Investigating the role of JARID2 in treatment resistance in glioblastoma

Muna Mubarak Hamed Al-Jabri

Submitted in accordance with the requirements for the degree of Doctor of Philosophy

The University of Leeds

Leeds Institute of Medical Research at St James's

Faculty of Medicine and Health

July 2023

The candidate confirms that the work submitted is her own and that appropriate credit has been given where reference has been made to the work of others.

Acknowledgements

I would like to give my sincere gratitude to the Almighty "Allah" for allowing me to complete this piece of work successfully. Then, I would like to express my gratitude to my three supervisors, Dr. Lucy F. Stead, Dr, Thomas A. Hughes and Dr. Laura C. Matthews, for the incredible learning experience I have had over the last five years. First and foremost, I want to express my sincere gratitude to Lucy, who has always been a very encouraging supervisor and has provided constructive criticism and advice to help me develop as an independent researcher. I would also like to thank Tom and Laura for their high-quality scientific guidance, their feedback and useful discussions throughout this work.

I would like to extend my thanks to everyone on Lucy's team for making the previous five years such a pleasure. A big thanks go to my friends Marilena Elpidorou and Rhiannon Barrow for being my lab mentor and perhaps the greatest instructors I could have had during this learning experience. In addition, a particular thank you go to Georgette tanner who helped me in the bioinformatics part of the work. Also, a special thanks goes to Martina Finetti who tried to complete few tasks for me during COVID-19 pandemic. I would also like to thank all the people on level 5 for their support.

A special thanks go to my family and more specifically my parents for their love, supports and continuous encouragement during this challenging PhD completion journey. Last but not least, I want to extend a heartfelt thank you to my partner Ali for his love, support, and ability to constantly make me smile, even during the most stressful times.

I would like to dedicate this thesis to my beloved brother, Fahad Al Jabri, who died recently. He was always believed in my ability to be successful in the academic area. You are gone but your belief in me has made this journey possible.

Abstract

Glioblastoma (GBM) is the deadliest primary brain tumour in adults with a median survival of 14-20 months from initial diagnosis. Despite aggressive treatment involving maximal surgical debulking followed by radiation therapy and chemotherapy, GBM remains incurable owing to a high rate of fatal recurrence. Understanding why unresected tumour cells survive treatment is necessary to better treat this disease. GBM tumours recur due to the presence of treatment-resistant cells, and this resistance phenotype is posited to be mediated by several epigenetic mechanisms including DNA methylation, histone modifications and chromatin remodelling. Work within the Glioma Genomics group in Leeds has specifically highlighted a potential role for histones, so this study focuses on understanding the role of histone modifications in GBM treatment resistance. More specifically, this group has recently proposed, for the first time, that Jumonji and AT- Rich Interacting Domain 2 (JARID2) plays a role in tumour recurrence via chromatin remodelling in GBM. JARID2 is an accessory protein of Polycomb Repressive Complex 2 (PRC2) which is the sole complex responsible for trimethylation on lysine 27 of histone H3 (H3K27me3). JARID2 promotes PRC2 recruitment to chromatin and regulates its enzymatic activity. PRC2 and JARID2 have a fundamental role in neurodevelopment but the role in GBM treatment resistance needs to be investigated. This can be achieved by generating genome-wide profiles of histone modifications associated with JARID2, and the binding of JARID2 itself, in paired primary (untreated) and recurrent (post-treatment) samples. Despite the tremendous work in generating global genome-wide profiling of histone modifications in various types of cancers, few genome-wide maps for H3K27me3 are available for GBM, therefore, current interest is placed on generating and comparing genome-wide mapping of histone modifications to locate and identify key epigenetic changes that are associated with GBM development and progression. Another histone mark (H3K4me3, which is a transcriptional activator) is known to work in concert with H3K27me3 during cell lineage determination in the brain, so it was also deemed necessary to prolife this mark. Thus, this study aimed to establish a workflow for generating and comparing the global distribution patterns of two histone modifications, along with binding patterns of the catalytic component of PRC2 (Enhancer of zeste homolog 2: EZH2) and JARID2, in paired primary and recurrent GBM samples. My hypothesis is that histone remodeling is driving the changes in the gene expression observed in GBM through treatment. I established

a workflow and then generated a genome- wide chromatin landscape for H3K27me3, H3K4me3 and EZH2 binding from matched fresh frozen pair primary and recurrent GBM samples of our in-house dataset. Then, I performed an integrative analysis on histone marks along with EZH2 by correlating their modifications with the changes in gene expression. The analysis was performed on a subset of genes that have been found to be dysregulated in GBM's patient following standard treatment due to the epigenetic remodeling of their promoters. The findings revealed that these genes are significantly found in the bivalent state, but the balance of histone marks is altered by therapy. Also, it revealed that histone modifications are driving gene expression in those genes more than the others This leads to the hypothesis that this bivalency is what causes the tumours to be able to adapt to treatment. I concluded that JARID2 genes facilitate tumour recurrence through transcriptional reprogramming in patients following standard therapy. Additionally, it implies that bivalent areas promote GBM tumorigenicity and are linked to chemo-resistance.

Table of contents

List of Figures

List of Tables

Chapter 1
Introduction

1.1. An overview of glioblastoma (GBM)

1.1.1. Clinical characteristics and Classification of GBM

Among all invasive cancers, malignant brain tumours account for a small proportion (approximately 2%), with gliomas comprising 80% of those (1-3). Glioblastoma (GBM), previously known as glioblastoma multiforme, is the most prevalent, aggressive and untreatable subtype of glioma in adults, accounting for more than 60% of all brain tumours. The term multiform denoted that GBM is heterogeneous in nature, consisting of a variety of cellular phenotypes and distinct mutational profiles (4-6). GBM is often diagnosed in older patients with a median age of 64 years at diagnosis. The incidence rate rises considerably with age, peaking between 75 and 84 years and drops after 85 years (7, 8). According to several international studies, the overall annual incidence rate of GBM ranges from 3.19 to 4.17 cases per 100,000 people and is rising in many countries (8). Additionally, GBM incidence has been found to be higher in males as compared to females (3.97 vs. 2.53 in the United States) (9).

The World Health Organization (WHO) has classified GBM as grade IV astrocytoma (6, 10). It is further clinically subdivided into primary (de novo) and secondary GBMs. Primary GBMs arise de novo, supposedly from glial cells or their progenitors, without any precursor lesions or prior symptoms. It accounts for about 90% of GBM cases. On the other hand, 10% of GBM cases are secondary GBM, which arises from pre-existing lower-grade astrocytoma (1, 2). The transcriptional and genomic characteristics of these two groups can be used to distinguish between them. The most prevalent mutations in primary GBM, for example, include p16 deletion, phosphatase and tensin homolog on chromosome ten (PTEN) mutation, EGFR amplification with loss of heterozygosity (LOH) on chromosome 10q, and oncogene amplification of the mouse double minute 2 (MDM2) gene. On the other hand, secondary GBM is associated with mutations in retinoblastoma (RB), tumour protein 53 (TP53) and LOH on 17p, 10q, and 19q (11, 12).

It has been demonstrated that these genetic alterations modulate the oncogenic pathways, leading to GBM invasion and progression. Disruption of the growth factor tyrosine kinase receptor pathway (EGFR), P53 pathway that includes MDM2 and TP53, retinoblastoma (RB) tumour suppressor pathway and Wnt signaling pathway have been found to be critical for GBM development and progression (13, 14). The detail of signaling pathway disruptions in GBM is provided below.

1.1.2. Signaling pathways disruption in GBM

The most common changes in primary GBM, accounting for 40% to 60% of cases, are EGFR amplification and overexpression. EGFR, the transmembrane receptor tyrosine kinase protein, is essential for controlling cellular growth, migration, differentiation, and tumour-induced neovascularization (15, 16). Aberrant EGFR contributes to aberrant activation of several downstream signaling pathways. For instance, activation of EGFR leads to the activation of the phosphatidylinositol-3-kinase (PI3K)/AKt and the mammalian target of rapamycin (mTOR) signaling pathways (PI3K-AKt-mTOR) which induce cancer proliferation, tumour development and therapy resistance (16, 17). Additionally, amplification and overexpression of EGFR result in an abnormal activity of son of sevenless 1 (SOS1) and growth factor receptor bound protein 2 (GRB2) which enhance cell proliferation, tumour transition, migration and development (14). Several small-molecule kinase inhibitors such as gefitinib, erlotinib, afatinib, dacomitinib and osimertinib have been tested against EGFR in the context of GBM, however, these inhibitors did not show any therapeutic efficacy in the clinical trials (15, 18). This is mainly due to the challenges in targeting EGFR which include the presence of mutations, GBM heterogeneity and ineffective blood-brain barrier penetration (18).

P53 pathway including MDM2 and TP53 is also one of the most commonly altered pathways in numerous types of cancer including GBM. It is mostly implicated in the prevention of the tumour by orchestrating a wide range of cellular responses, such as damaged cell apoptosis, maintaining genomic stability, inhibiting angiogenesis, and regulating cell metabolism and the tumour microenvironment (19, 20). Disruption of P53 signaling pathways including MDM2 and TP53 is occuring in 87% of GBM cases, leading to defects in the mechanisms governing cell cycle arrest, senescence, DNA repair and apoptosis (21). Deregulation of these mechanisms has been found to be associated with GBM progression. The key participant in the P53 signaling pathway is encoded by TP53, and is often altered in GBM (28-30% of cases) and other forms of malignancies (19). The second alteration of P53 mutation is MDM2 amplification. MDM2, an E3 ubiquitin ligase, ubiquitinates p53 in order to degrade it by the proteasome. MDM2 amplification has been found in different types of cancers and it is often associated with a poor prognosis (21, 22). Growth arrest, apoptosis, DNA repair, and other tumour suppressor mechanisms can all be lost as a result of p53 inactivation by MDM2 amplification (23). Several small molecule inhibitors have been developed to inhibit tumour progression in patients with P53 deregulation. Atorvastatin and rosuvastatin, the FDA-approved drugs for P53 mutation, have been shown to successfully inhibit the growth of

tumours in the cells that harbor P53 mutation. However, the clinical safety and efficacy need to be assessed (24). The MDM2 inhibitors nutlins, idasanutlin, navtemadlin, APG-115, BI-907828, CGM097, siremadlin, and milademetan have been extensively studied. Although several of these inhibitors have progressed to clinical development, their effectiveness has not yet been established (25). To determine the effectiveness of MDM2 inhibitors in the treatment of GBM and to pinpoint the patient population that would benefit most of this therapeutic strategy, more study is required.

retinoblastoma (RB) tumour suppressor pathway, which consists of retinoblastoma tumour suppressor (RB), cyclin-dependent kinases inhibitors and activators, and the E2F-family of transcription factors, is essential for controlling cell cycle progression and cell death (26). In GBM, the pRB pathway was disrupted in 78% of the primary GBM patients, and 7.6–11% of cases had RB gene deletions or mutations (27). The most common RB pathway alterations in GBM are Cyclin/CDK4-6 amplification (15%), CDKN2A/B deletions or inactivating mutations (40%), and RB1 gene deletions or inactivating mutations (40%) (12, 28, 29). Several CDK inhibitors have been developed to reactivate pRB, and only PD0332991 (Palbocicilib), an inhibitor of CDK4/6 has been shown to prevent the downstream suppression of pRb (30). However, for such highly selective drug, more trials are required to assess its treatment efficacy since tumour evolution may easily develop bypass mechanisms to get around such a single agent (31).

In addition, aberrant activation of WNT signaling has been shown to facilitate GBM development and invasion by maintaining stem cell characteristics (14). In general, Wnt pathway is involved in the regulation of key biological processes such as cellular proliferation, polarity, differentiation, motility, and stem cell activity, which are all essential for development, regeneration, and homeostasis (32). WNT is typically activated in GBM by genetic abnormalities, such as a loss in FAT1, a detrimental WNT signaling effector that is present in 20% of GBM patients (33). Due to their crucial involvement in carcinogenesis and cancer progression, as mentioned above, WNT pathway proteins have been recognized as a desirable and reliable cancer target (32). WNT pathway inhibitors have been developed over the past ten years, and some of these have even undergone clinical trials. But so far, there haven't been any FDA-approved medications that target WNT pathways (33).

Despite enormous advances in understanding the biology of cancer and in the improvements of cancer diagnosis and treatment, the prognosis for GBM has remained unchanged for 2 decades (34, 35). Patients with GBM have a very poor prognosis witha median survival of 14-20 months from initial diagnosis, making it a critical matter of public health (1, 36). TMZ resistance and tumour recurrence

are the main reason behind this poor prognosis: almost 100%of GBM tumours recur (3, 6, 37). Therefore, it is important to understand the reasons behind GBM recurrence and the mechanisms that GBM develops to resist treatment in order to design more effective therapeutic strategies (3). Extensive evidence has recently suggested that GBM heterogeneity mediates tumour recurrence and treatment resistance (38, 39). Therefore, understanding tumour heterogeneity is a crucial step in advancing personalized medicine and enhancing clinical outcomes (5, 40).

1.2. Challenges in GBM therapy

1.2.1. Current treatment regimen

Current standard management of GBM consists of maximal surgical debulking followed by radiation therapy (RT) with concomitant and adjuvant temozolomide (TMZ), a DNA alkylating chemotherapeutic agent (36, 40). Following Stupp et al. (2005)'s seminal study, this standard of care was initially implemented in 2005 (41). GBM patients receive treatment with the Stupp regimen after a safe surgical excision of the tumour. This regimen consists of radiation therapy administered as 60 Gy total in daily fractions of 2 Gy; concurrent daily TMZ administration (75 mg/m2/day) from the first to the last day of radiation therapy; and adjuvant TMZ administration (six cycles, 150-200 mg/m2/day for 5 days during each 28-day cycle) (41-43). Research has demonstrated that the addition of TMZ to radiation therapy increases the overall survival of patients by two months (42).

TMZ, commercially known as Temodar, is an orally administrated chemotherapeutic agent that is rapidly and completely absorbed after oral administration due to its biochemical characteristics (44). Additionally, TMZ penetrates the BBB since it is a tiny lipophilic molecule, making it the drug of choice for treating GBM patients (45). TMZ is known to cause cell cycle arrest at G2/M, which ultimately results in apoptosis. At physiological pH (pH > 7), TMZ gets activated through a non-enzymatic conversion into a short-lived active metabolite 5-(3-methyltriazen-1-yl)-imidazole-4-carboxamide (MTIC). MTIC is further hydrolysed to produce 5-amino-imidazole-4-carboxamide (AIC) and the highly reactive methyldiazonium cation. The latter preferentially adds a methyl group to guanine residues at the oxygen-6 (O6, 6%), nitrogen-7 (N7, >70%), and nitrogen-3 (N3, 9%) at adenine residues. This results in the production of the cytotoxic bases O6-methylguanine (O6-MeG), N7-methylguanine (N7-MeG), and N3-methyladenine (N3-MA), which have a positive, clinical effect. The alkylation of DNA, particularly at the O6 and N7 sites of guanine, is thought to be the cause of MTIC's cytotoxicity, which results in DNA double strand breaks and apoptosis (45, 46). Methylation

of guanine at O6 (O6-MeG), despite only present in a small percentage (7%), is cytotoxic, mutagenic, and essential for TMZ-induced cytotoxicity. It results in the insertion of thymine instead of cytosine residues during DNA replication leading to cell death (47).

Although TMZ is still the primary medication used to treat GBM today, numerous studies have shown that cells can develop resistance to TMZ. More than 50% of GBM patients who receive TMZ treatment do not respond to the treatment, leading to tumour progression and recurrence (48). The main TMZ resistance mechanisms are provided below in details.

1.2.2. TMZ resistance and tumour relapse

Several molecular mechanisms related to DNA damage repair have been shown to contribute to TMA resistance, including O^6-methylguanine-DNA methyltransferase (MGMT), mismatch repair pathway (MMR), and base excision repair (BER, the poly (ADP)-ribose polymerase (PARP) pathway) (48). To date, the main contributor to TMZ resistance is MGMT due to its role in preventing DNA damage from DNA alkylation, leading in a diminished TMZ cytotoxic effect (49). The MGMT gene, which is located on chromosome 10q26, encodes a DNA-repair protein that removes methyl groups from the O6 position of guanine, repairing TMZ-induced DNA damage and reducing the effectiveness of the medication (50).

A measure of intrinsic resistance to TMZ is the epigenetic status of MGMT, which includes promoter methylation, histone modifications, and miRNA modulation of transcription levels (51, 52). MGMT expression is governed by the CpG methylation state of the MGMT gene promoter region (53). For instance, promoter hypermethylation, which is associated with silencing of MGMT gene, reduces MGMT protein expression, preventing it from performing its function as a DNA damage protector and enhances the response to the chemotherapy (53, 54). In contrast, the unmethylated MGMT promoter is associated with the enhancement of MGMT protein production, which results in TMZ resistance (49). There is growing evidence from meta-analysis studies that MGMT status might be subject to alteration during tumour's treatment, development, or recurrence (55). After recurrence following TMZ treatment, it has been noticed that tumours with initial MGMT methylation had a lower methylation ratio, indicating that the decrease in MGMT promoter methylation is a mechanism for developing therapeutic resistance to TMZ (44).

Another predictive marker of response to TMZ is the identification of mutation in Isocitrate dehydrogenase (IDH) gene (56, 57). IDH enzymes are categorized into three isoforms namely IDH1,

IDH2 and IDH3. These enzymes are involved in several metabolic processes such as lipid metabolism, Krebs cycle and redox regulation. The enzyme IDH1 is located in the cytoplasm and peroxisomes, whereas the enzymes IDH1 and IDH2 are localized in the mitochondrial matrix (57, 58). As primary roles, IDH enzymes catalyze the oxidative decarboxylation of isocitrate to produce -ketoglutarate (KG) in the citric acid cycle (58). IDH mutations are a frequent event in all human malignancies including GBM. The prevalence of IDH mutations is considerably higher in secondary GBM, which accounts for 73% of clinical cases, than in the primary GBM (57, 58). Based on IDH status, GBMs can be classified as either GBM-IDH-wt (wild type) or GBM-IDH-mut (mutant) (59). In GBM, Patients with mutant IDH1 have a better disease outcome compared with the patients with wild type IDH1 (58). There has been a growing attempt to incorporate molecular tumour features into GBM diagnosis and treatment, however, due to the involvement of numerous molecular processes, TMZ resistance mechanisms are still not fully understood (44). In addition to these molecular biomarkers, TMZ resistance and tumour recurrence have been found to be strongly associated with intrinsic intra-tumour heterogeneity that GBM cells poses (10, 40).

1.2.3. GBM Heterogeneity and tumour recurrence

Ninety per cent of GBM tumours recur within 1-2 cm from the resected tumour edge because of the infiltrative nature of the GBM (1, 2). Infiltrative GBM cells spread into the surrounding brain parenchyma and migrate along the white matter tracts or the blood vessels (3, 4). These areas cannot be removed surgically due to the risk of neurological damage, making a complete surgical resection impossible and meaning that the unresected cells left behind serve as an origin of recurrence. The recurrent tumours exhibit tremendous cellular and molecular heterogeneity compared to those in the initial tumour therefore, they are usually not sensitive to the original treatment (5, 6).

From the histopathological point of view, GBM appears as an irregular heterogeneous tumour consisting of differentiated and undifferentiated cells characterized by self-renewal and proliferation (10, 38). In general, GBM displays remarkable heterogeneity within a single tumour (intra-tumour heterogeneity) and between tumours (inter-tumour heterogeneity) (37, 39, 60). Intra-tumour heterogeneity refers to the presence of subpopulations of cells with different phenotypes owing to variability in cellular transcriptomic, genetic and epigenetic features. These subpopulations harbour distinct molecular signatures conferring different levels of sensitivity to therapies (6, 60). Several reports have shown that intratumoral heterogeneity serves as a potential

hallmark that contributes to tumour progression and poor responsiveness to therapy (5, 61). Glioblastoma's heterogeneity can be explained by two main proposed models as in the case of other cancers: the stochastic (or clonal evolution) model and the hierarchical cancer stem model (35, 62, 63).

The clonal evolution model (**Figure 1.1a**) hypothesizes that during tumour progression, a normal cell undergoes a series of genetic mutations or epigenetic alterations, inherited or acquired, leading to the formation of cancer cells (6, 35, 38). These cells are then clonally proliferated and expanded in response to tumour microenvironment selective pressures such as acidosis, hypoxia, immune evasion and competition for spaces and resources and give rise to clonal outgrowths. Clones with growth advantages will expand to form tumour bulk that harbours multiple sub-clones of cells with stem-like properties and the clones with less fitness will probably become extinct. These sub-clones may undergo other waves of cumulative mutations, resulting in complex sub-clonal cells with different biological features and molecular profiles (1, 6, 36). On the other hand, the cancer stem model (**Figure 1.1b**) proposes that tumours are mainly initiated and maintained due to the presence of a subset of stem cell-like cells termed cancer stem cells (CSCs). These cells exhibit high plasticity, tumorigenic and indefinite self-renewal features. Only these CSCs can proliferate and give rise to new tumours which have different molecular profiles compared to the mother cells. CSCs divide asymmetrically to form bulk tumours constituting new CSC, progenitor cells and differentiated cancer cells, with the symmetric division responding to expanding the stem cell population (1, 38, 64). **Figure 1.1** illustrates the mechanisms of tumour initiation and maintenance based on the two proposed models. In both models, the resulting sub-clones possess the ability to resist treatment and, therefore, they survive and expand resulting in the re-occurrence of the disease. While the cancer stem cell model better explains inter-tumour heterogeneity, the clonal evolution model better describes the development of intra-tumour heterogeneity (40, 65).

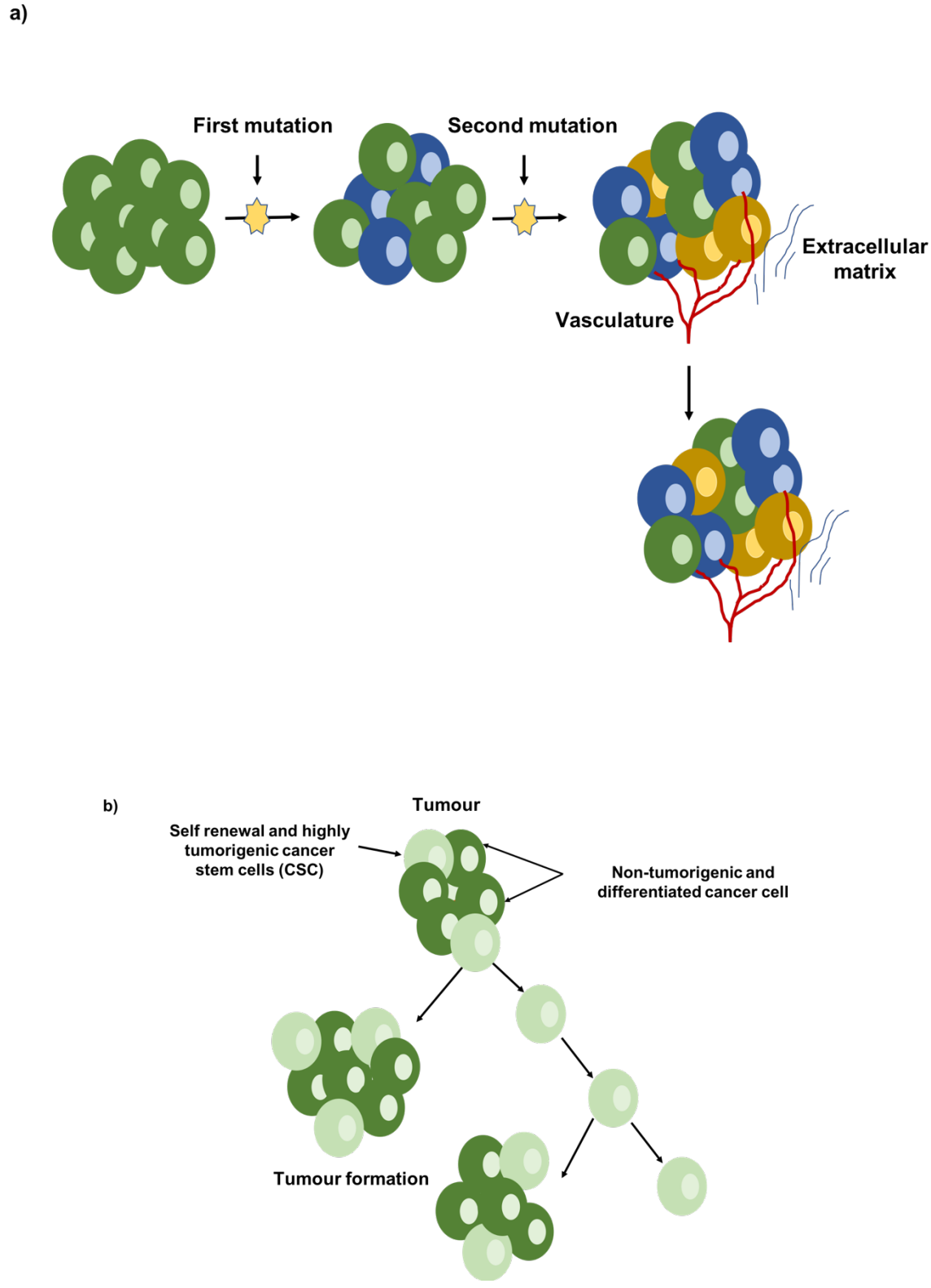


Figure 1-1: The proposed models for tumour initiation and progression. (A) The clonal evolution model proposes that normal cells (dark green) undergo a series of mutations, leading to the formation of cancer cells. These cells are then clonally proliferated and expanded and give rise to clonal outgrowths. Clones with growth advantages will expand to form tumour bulk that harbours

multiple sub-clones of cells with stem-like properties. **(B)** The cancer stem model hypothesizes that tumours are mainly initiated due to the presence of a subset of stem cell-like cells termed cancer stem cells (CSCs) (Light green). These cells exhibit high plasticity, tumorigenic and indefinite self-renewal features. Only these CSCs can proliferate and give rise to new tumours. CSCs divide asymmetrically to form the bulk of the tumour with new CSC, progenitor cells and differentiated cancer cells and symmetrically to expand the stem cells.

Thus far, numerous investigations have shown that tumour heterogeneity is the key component of tumour recurrence and progression and this includes complex genetic mutations, epigenetic abnormalities, growth rate, protein modification and apoptosis. It causes a significant challenge in designing new therapeutic agents, therefore understanding the molecular events that drive GBM resistance is essential in designing more effective therapeutic strategies (10, 66).

This work focuses on understanding the epigenetic mechanisms involved in GBM resistance and recurrence as recent work conducted by our group has shown that transcriptional changes occur dynamically after treatment in GBM. The study included gene expression analysis on 45 matched pairs of primary and recurrent GBM tumours and the results showed that genes that contain a Jumonji and AT-Rich Interacting Domain 2 (JARID2) binding site in their promoters are commonly and significantly dysregulated after standard treatment. JARID2 has a role in chromatin remodelling and epigenetic modifications, specifically via histone marks, as outlined below (67).

1.3. The role of epigenetic alterations in drug resistance and recurrence in GBM

1.3.1. Overview of Epigenetics

Research in the post-genomic era revealed the involvement of epigenetic alterations in tumorigenesis in almost all cancers, including GBM. These alterations can result in different profiles of gene expression and, thus, elicit phenotypic changes (68, 69). Nearly all cells of multicellular organisms share the same genetic information, meaning that they contain identical deoxyribonucleic acid (DNA). However, during cell development, each single cell differentiates into a distinct cellular phenotype (70, 71). Waddington used the term ‘epigenetics’ to describe this phenomenon, which is currently defined as long-term heritable changes in DNA conformation and chromatin structure that affect gene expression without affecting the underlying DNA sequence (69, 72, 73). The communication of epigenetic information is mediated by multiple mechanisms

including DNA methylation, covalent histone modification, non-coding RNA and chromatin remodelling. These mechanisms are controlled by sets of modifier enzymes termed writers, readers and erasers that respectively add, recognize and remove the epigenetic modifications on DNA and histone proteins (68, 74, 75). Numerous studies indicated that these modifications are critical to regulating chromatin structure and DNA accessibility, thereby regulating gene expression. It is believed that chromatin is the main barrier to transcription, therefore, understanding chromatin structure and how epigenetic alterations remodel its structure will be essential in targeting epigenetic mechanisms that are associated with tumour resistance in order to design appropriate therapeutic targets for different types of tumours including GBM (76-78).

1.3.2. Chromatin Structure and Function

Genomic DNA within the nucleus of eukaryotic cells is tightly compacted and condensed into arrays of nucleosomes called chromatin. Nucleosomes, the fundamental basic and recurring unit of chromatin, are composed of approximately 147 bp of DNA encircled around a histone octamer of two copies each of four core histone proteins: two H2A/H2B dimers and a central H3/H4 tetramer (79-81). The adjustment of the wrapping of DNA around this octamer comprises the physical foundation for the regulation of transcription of nucleosomal DNA (75, 82). In addition to these core histones, a linker histone protein H1 stabilizes and organizes the nucleosomes into a higher-order chromatin structure through its binding to the linker DNA, the DNA between adjacent nucleosomes (78, 83, 84). Nucleosome spacing defines the structure of the chromatin and can be widely divided into euchromatin, which is a less condensed form that corresponds to the transcriptionally active regions, and heterochromatin; a highly condensed form that corresponds to the inactive transcriptional regions (74, 78, 85).

Chromatin conformation is essential for the regulation of gene expression and chromosome function since it effectively controls all DNA processes such as transcription, DNA replication and DNA repair (77, 78, 86, 87). A growing body of evidence has shown that chromatin structure and DNA accessibility to transcriptional machinery are dynamically regulated by modifications to both DNA and core histone tails. These modifications involve two major mechanisms: DNA methylation and histone modifications (75, 84, 87). Details of these two mechanisms are provided below.

1.3.3. DNA methylation in GBM

DNA methylation is a silencing chemical modification that occurs de novo and is maintained during cell division by the enzymatic family of DNA methyltransferase (DNMTs) on the cytosine residues of the DNA molecule (40, 88). The enzyme covalently transfers a methyl group from S-adenosyl-methionine to the carbon-5 position of cytosine (C5), thus forming 5-methyl-cytosine (5mC). DNA methylation occurs most commonly at the CpG islands (C = cytosine, p = phosphate bond and G = guanine) of the promoter regions of genes, where a cytosine residue occurs next to guanine residue causing the chromatin to compact and making it inaccessible to the transcriptional machinery (69, 89, 90). It is involved in several evolutionary biology-related processes such as genomic imprinting, silencing retroviral elements, regulating the expression of germline-specific genes and X chromosome inactivation. Additionally, it plays a crucial role in the regulation of transcriptional potential and regulation of gene expression (69, 91). However, recent studies suggested that aberrant DNA methylation of gene promoters is a crucial process contributing to the progression and oncogenesis of multiple cancers (90, 92).

It has been reported that global loss of methylation, termed hypomethylation, in the CpG islands of the promoter regions of the genes induces genetic instability and facilitates transcriptional activation of oncogenes, leading to mutagenesis and tumour progression. Whereas, an increase in the level of methylation, termed hypermethylation, is associated with gene silencing of tumour suppressor genes, which is a hallmark of carcinogenesis (90, 93, 94). The correlation between DNA methylation and tumour progression has been extensively studied in multiple types of cancer, including GBM, using several methods such as DNA methylation arrays, methylation-specific PCR and bisulfite sequencing (40, 88, 95). In GBM, drug resistance was found to be modulated by promoter hypermethylation and the best known, and clinically utilised, example is the hypermethylation of O-6-methylguanine-DNA methyltransferase (MGMT), a DNA damage repair gene that protects cancer cells from chemotherapeutic alkylating agents. Silencing of this gene was found to be significantly associated with longer survival of GBM patients, therefore, such change appears as a promising target for GBM treatment (94, 96). Hypomethylation of gene promoters was also reported in GBM and one particular important example is a promoter hypermethylation of SOX2 ((Sex Determining Region Y)-box 2) that is considered to be a stem cell-related transcription factor (TF). This family of transcription factors is found to be associated with glioma progression (40, 88).

DNA methylation is one of the best-addressed epigenetic modifications in cancer and in particular in GBM (69, 77, 92), but less widely characterised is the role of histone modifications in the regulation of gene expression and chromatin state in cancer. Recent research has been increasingly focused on the role of post-translational modifications (PTMs) of histones in cancer and, specifically, in GBM for gaining deeper insight into the complex interplay of different epigenetic modifications in cellular processes. These studies improve our knowledge of how these modifications alter chromatin structure at all stages of cancer development from initiation to progression (97, 98). Despite the expanding body of GBM research, little is known about how the epigenome promotes the progression of GBM and the exact epigenetic mechanisms of therapy resistance in GBM need to be elucidated. There is an increased interest in generating genome-wide histone modification maps for gliomas as there are very limited data on this field (10, 89, 99).

1.3.4. Epigenetic modifications: role of histone modifications in chromatin machinery

Histones are highly conserved, alkaline proteins with a positively charged N-terminal tail constituting 20-35 residues and protruding from the globular domain of the histone (74, 80, 81). These tails are subjected to a remarkable number of covalent post-translational modifications including methylation, acetylation, ubiquitination and phosphorylation. These modifications are catalyzed by different modification enzymes that cooperate to regulate the chromatin state. The “histone code hypothesis” suggests that DNA transcription is regulated by various patterns of histone modifications which may have an impact on nucleosome stability and hence, the dynamic state of the chromatin (74, 77, 85, 100).

Histone modifications involve covalent addition and removal of various molecules on the N-terminal tail of the histone. Most of them are dynamic and enzymatically reversible in nature and can be returned back to their original state during normal physiology and by epigenetic therapy (101, 102). The most extensively modified histone is H3, followed by H4. It has been reported that covalent modifications on the residues of histone proteins, and more specifically methylation and acetylation, have either a direct or indirect effect on the chromatin structure, thereby, leading to alterations in gene transcription (102, 103). These transcriptional alterations in gene expression are found to be associated with the development and progression of various types of cancers including GBM (81, 101, 104). In this study, I am particularly interested in understanding the role of histone methylation in gene regulation. This is because histone methylation functions remain largely unknown due to its complexities (68, 104). An example of this complexity is that histone methylation

does not cause any change in histone charge but rather creates a docking site for chromatin-related proteins that contain a specific methyl binding domain (68, 75). Whereas, histone acetylation causes a change in histone charge that affects the interaction between the negatively charged DNA and histone, causing a partial unwinding of the DNA from the nucleosome (75). Another layer of complexity is that histone methylation can be associated with either active transcription (methylation at H3K4 and H3K36) or repressed transcription (methylation at H3K27 and H3K9) depending on the methylation sites or the extent of its methylation, whereas histone acetylation of lysine residues is usually associated with active transcription (76, 101). Therefore, it is still unclear how histone methylation regulates gene expression to promote cancer cell initiation and survival (101, 102). Histone methylation will be addressed in this study so an overview of it is given below in more detail.

1.3.4.1 Histone Methylation in GBM

Histone methylation takes place primarily on the N-terminal tail of both lysine and arginine residues and is catalysed by histone methyltransferases (HMTs) (92, 104). HMTs catalyze the covalent transfer of methyl groups from S-adenosylmethionine (SAM) to the specific lysine and arginine residues of histone proteins. Lysine residues are methylated by lysine methyltransferases (KMTs) and can be mono-, di- and tri-methylated, while, arginine residues are methylated by arginine methyltransferases (PRMT) and can be only mono- or di-methylated (101, 102, 104). Variations in the number of methyl groups added and the type of the modified residues can elicit different chromatin statuses and, thus, different transcription patterns. For instance, trimethylation of lysine 27 (H3K27me3) and lysine 9 (H3K9me3) of histone 3 is associated with transcriptionally repressed heterochromatin, whereas, di- and trimethylation of lysine 4 of histone 3 (H3K4me3) is associated with transcriptionally active euchromatin (87, 102, 105). In this study, I will focus mainly on the most well-characterized H3 methylation, namely H3K27me3 and H3K4me3.

It has been demonstrated that these two methylations are mainly enriched in the promoter region of the gene (61, 63). Notably, some gene promoters are bivalent i.e. contain both therepressive mark H3K27me3 and the active mark H3K4me3. Bivalent regions were initially hypothesized to be prevalent in embryonic stem cells (ESCs), but recent studies confirmed the existence of bivalent regions in terminally differentiated cell types as well as glioma stem cells (GSCs) (64, 65). These regions serve as an epigenetic control mechanismenabling rapid regulation of genes associated with cell differentiation and lineage determination during embryogenesis (48, 64). Bivalent promoters were initially thought to keep genes in a poised state but enable them for rapid activation

upon appropriate developmental cues and/or environmental stimuli, while maintaining a transcriptionally repressed state (36, 47). However, recent subsequent studies have proposed a unifying model which showed that bivalent regions do not poise genes for rapid activation but protect gene promoters from de novo DNA methylation, while maintaining a reversibly repressed state. Further, H3K4me3 at bivalent promoters is instructive for rapid activation of transcription and that activation of bivalent genes is neither greater nor more rapid than that of other transcriptionally repressed genes without H3K4me3 (106, 107) .

In general, H3K27me3 is mediated by histone methyltransferase Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit (EZH2), a subunit of Polycomb Repressive Complex 2 (PRC2), which mediates maintenance and self-renewal of stem cells by trimethylation of H3K27 and suppresses stem cell differentiation. Whereas, H3K4me3 is catalyzed by Trithorax protein complexes (101, 102, 108).

PRC2 is composed of three core subunits: suppressor of zeste 12 (SUZ12), embryonic ectoderm development 1 to 4 (EED1-4), and the catalytic subunit enhancer of zeste 1 or 2 (EZH1/2). Besides these core components, PRC2 interacts with other cofactors that regulate its enzymatic activity and modulate its binding to chromatin. These include retinoblastoma binding proteins 4 and 7 (RBBP4/7), adipocyte enhancer-binding protein (AEBP2) and Jumonji AT-rich interaction domain (JARID2) (**Figure 1.2**) (109, 110).

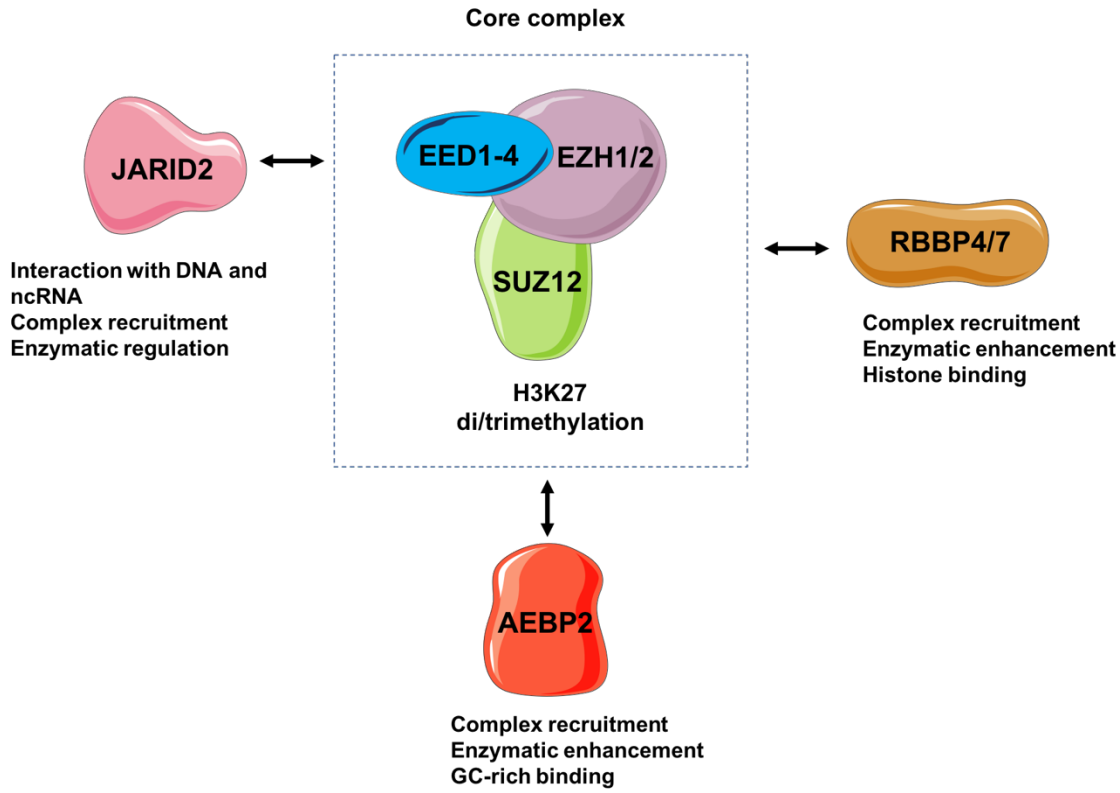


Figure 1-2: The PRC2 complex in mammals with core components and cofactor proteins. PRC2 is composed of three core subunits: suppressor of zeste 12 (SUZ12), embryonic ectoderm development 1 to 4 (EED1-4), and the catalytic subunit enhancer of zeste 1 or 2 (EZH1/2). In addition, PRC2 interacts with other cofactors components such as retinoblastoma binding protein 4 and 7 (Rbbp4/7), adipocyte enhancer-binding protein (Aebp2) and jumonji AT-rich interaction domain (JARID2). These components regulate PRC2 enzymatic activity and modulate its binding to chromatin (110).

PRC2 plays a central role during development and in cell differentiation (67). Several studies revealed that regulation of histone modifications by PRC2 is a key factor of tumour cell plasticity, which is necessary for glioblastoma cells to survive and adapt to their microenvironment (69, 71). Disruption of PRC2 activity, through overexpression of its enzymatic subunit EZH2, leads to poor prognosis in GBM patients, further highlighting the relevance of this histone modification in glioma biology. EZH2 and PRC2 were found to suppress many genes involved in cell-cycle regulation, cell differentiation and proliferation and self-renewal. It has been found that overexpression of EZH2 in glioma cells leads to an increase in glioma cell self-renewal, proliferation and migration (69, 103, 111). In addition to EZH2, it has been recently shown that a Jumonji and AT-Rich Interacting Domain 2 (JARID2), an accessory protein of PRC2, has the capability to bind DNA and thus dock PRC2 to specific sites where it can deposit its associated histone marks. Additionally, numerous studies reported the strong relationship between PRC2 and JARID2 in regulating the catalytic activity of PRC2 which is mediated by EZH2-mediated methylation (110, 112). JARID2 either activates or inhibits the catalytic activity of PRC2 (110).

This was further emphasized by previous work in our group which highlighted JARID2 as a potential master regulator of transcriptional changes through treatment but furthermore showed that these changes were taking place at genes that are commonly found to be bivalent in both normal brain and glioma tissue (67). The work included transcriptional analysis of 217 pairs of GBM samples that are specifically wild-type for isocitrate dehydrogenases (IDH^{wt}) and recurred locally following standard treatment (i.e. radiation and Temozolomide). As described in **Section 1.2.2**, patients with IDH1 mutations have a better prognosis compared to those with IDH^{wt} (71). Recent studies reported that wild-type IDH enzymes play a critical role in promoting GBM growth and recurrence (**See section 1.2.2. for more detail**) (72). To characterize the changes in transcriptional profiles in these two groups through treatment transcriptional analysis was performed. Two sources of paired GBM samples were used for this purpose, the Discovery cohort which consists of 168 longitudinally paired

samples from 84 patients and the Validation cohort which consists of 46 paired samples from 23 patients. RNAseq data for the discovery cohort was processed in house, whereas, for the validation cohort, RNAseq data was processed via a distinct pipeline within the Glioma Longitudinal AnalysSiS (GLASS) consortium. Differential expression analysis was performed using Deseq2 and the results revealed that genesthat were differentially expressed between matched primary and recurrent GBM samples were enriched for terms associated with neurodevelopment and cell lineage determination. In order to examine if particular regulators were implicated in these dysregulated genes in primary versus recurrent GBMs, gene set enrichment analysis (GSEA) was performed per patients using a novel, comprehensive gene sets for DNA-binding factors. A gene was assigned to a DNA-binding factor's gene set if its promoter (transcription start site from gencodev27 $\pm$1kbp, or $\pm$2 or 5kbp where specifically stated) included a binding site for a DNA-binding factor in at least two separate ChIPseq experiments. GSEA showed that genes that contain a Jumonji and AT-Rich Interacting Domain 2 (JARID2) binding site in their promoters (JBSgenes) were commonly and significantly dysregulated during standard treatment. JBS genes are subsets of genes with JARID2 binding sites in their promoters. Gene set enrichment analysis revealed the association of these genes with key signaling pathways that are critical for GBM development such as signaling pathways regulating pluripotency of stem cells, neuroactive ligand-receptor interaction, Wnt signaling, Adenosine 3',5'-cyclic monophosphate (cAMP) pathway, mitogen-activated protein kinase (MAPK) pathway and pathways in cancer such as PI3K-AKt-mTOR pathway P53 signaling pathways, vascular endothelial growth factor (VEGF) signaling pathway and Neurogenic locus notch homolog protein (NOTCH) signaling pathways. Genetic alterations in these pathways have been found to be associated with GBM proliferation, invasion, proliferation, self-renewal and cell survival.

Further analysis involved the examination of the stability of inclusion within the leading edge to see whether the same JBSgenes were causing the enrichment across patients. Of the 5234 JBSgenes in the Discovery cohort, 443 were found in the leading edge (LE) of at least 50% of patients (LE50) and 81 in more than 70% of patients (LE70). The Validation cohort showed similar results, with 444 genes found in LE50 and 87 genes in LE70. Further analysis was performed on these sets to examine the directionality of dysregulation, and the results showed that the direction of the fold change in expression of the LE70 genes from primary to recurrence (Log2FC) was consistent within patients, but varied between individuals. The LE70 genes were upregulated from primary to recurrence in 60% of patients (referred to as Up responders) while the same genes were downregulated in the other 40% of patients (referred to as Down responders). With the same proportion of Up and Down responders, this finding was recapitulated in the Validation cohort. These findings suggested that

JBSgenes drive patient enrichment; for example, the identical genes are downregulated in D response patients and increased in U response patients regardless of response subtype. To investigate therapy-driven changes in gene expression across these two responders, transcriptional reprogramming from primary to recurrent was performed using principle component (PC1) analysis on Log2FC profiles. The analysis showed that genes with the highest principle component 1 (PC1) were enriched for JBSgenes across these two responder types. This finding further confirmed that through treatment, up and down responder patients undergo transcriptional reprogramming in opposing directions, driven by a specific set of genes with JBSgenes enrichment being the most significant. These findings demonstrated that JBSgenes facilitate GBM recurrence through treatment by indirect transcriptional reprogramming of surviving cells and in opposing directions. The role of JARID2 in chromatin remodeling and epigenetic modifications, specifically via histone marks, is outlined below.

1.4 The role of JARID2 in relation to PRC2 in GBM progression

At the molecular level, JARID2 is the best-characterised cofactor of PRC2. It is a member of the jumonji family of proteins and contains a DNA binding domain known as the AT-rich interaction domain (ARID); a zinc finger domain; a jumonji N (JmjN) and jumonji C (JmjC) domain (110, 113). Multiple studies have shown that JARID2 is required for the complete recruitment of PRC2 to its target genes and that late or incomplete recruitment of PRC2 to chromatin, along with lower enzymatic activity, is observed in the absence of JARID2 (114). Despite the fact that JARID2 has a DNA-binding domain, its DNA-binding affinity is low and requires stimulation from other factors (115).

A model has been proposed of increased stimulation of the JARID2-PRC2 interaction *in vitro* via the additional interactions with specific long non-coding RNAs (lncRNAs) that also maximize PRC2 recruitment to DNA, resulting in increased H3K27me3. However, the role of JARID2 in PRC2 recruitment through lncRNA in specific contexts, such as during epigenetic reprogramming in GBM, remains unclear (113, 114).

The latest ENSEMBL (release 96) human gene annotation revealed the presence of three different JARID2 isoforms as shown in **Table 1.1**. Isoform one is considered to be the canonical gene product.

JARID2 isoforms	Predicted protein size (kDa)	Length (amino acid)
Isoform 1	140 kDa	1,246 aa
Isoform 2	120 kDa	1.074 aa
Isoform 3	105 kDa	960 aa

Table 1-1: Table 1: JARID2 isoforms from ENSEMBL release 96 (April 2019).
Table includes JARID2 isoforms based on ENSEMBL release, their protein sizes in kDa and the corresponding length in amino acids.

A recent study has identified the presence of a novel form of JARID2 that exists predominantly in lineage-committed human cells with a molecular weight of approximately 75 kDa. This form, denoted ΔN-JARID2, results from the cleavage of the N-terminal region from full-length (i.e. 140kDa) JARID2 leaving a stable C-terminal region. It was shown that ΔN-JARID2 is important for cell differentiation as its formation results in the dissociation of PRC2 and subsequent activation of previously repressed genes (112).

Our preliminary data suggest that gene expression changes associated with JARID2 occur during GBM recurrence. To investigate this further, we need to characterise and compare the binding site profiles of JARID2 (preferably both cleaved and un-cleaved forms) and EZH2 (as the catalytic subunit of PRC2) and the prevalence of H3K27me3 and H3K4me3 in paired primary and recurrent GBM samples. Genome-wide mapping has emerged as a new opportunity to decipher the histone code and enhance our understanding of how these modifications work together to regulate gene expression and how they contribute to diseases, and more specifically to the development of cancers (75, 116). For example, can the active mark (H3K4me3) and the repressive mark (H3K27me3) occur on the same nucleosome and what will be the effects of these integrations on the transcriptional machinery. Such information cannot be obtained by single-gene studies; therefore, genome-wide mapping is a powerful indicator for the identification and characterization of the combinatorial patterns of histone modification for each cell type at each gene locus (117).

1.5 Genome-wide mapping of histone marks and other regulatory domains in GBM

Genome-wide mapping of histone modifications, referred to as chromatin state maps, provides precise, descriptive data about the regulatory roles of histone modifications on gene expression, which is more informative than RNA expression profiling (117, 118). In the past few years, chromatin immunoprecipitation followed by sequencing (ChIP-seq) has emerged as a powerful tool for mapping and identifying global genome-wide patterns of these modifications (117, 119). Considerable amounts of genome-wide data have been generated via ChIP-seq for diverse sets of human cancers including prostate, breast, lung, colon and melanoma (97). Despite the presence of large numbers of histone modifications that seem to be involved in the regulation of gene expression, H3K27me3 and H3K4me3 have been successfully profiled on a large scale in different types of cancers. For instance, genome-wide analysis of the H3K27me3 profile in prostate cancer showed extensive enrichments of this mark in promoter regions in advanced disease in comparison to the normal tissues, suggesting the correlation of H3K27me3 with prostate cancer progression and aggressiveness (120). In another study, genome-wide profiling of 8 histone marks namely H3K4me3, H3K4me1, H3K27me3, H3K9me3, H3K36me3, H3K27ac, H3K9ac and H3K79me2 in 5 major breast cancer subtypes revealed the presence of subtype-specific chromatin state signatures for these major breast cancer subtypes (121). Despite the tremendous and exciting work in generating global genome-wide profiling of histone modifications in various types of cancers, few genome-wide profiling datasets for H3K27me3 are available for GBM, therefore, current interest is placed on generating and comparing genome-wide mapping of histone modifications to locate and identify key epigenetic changes that are associated with GBM development and progression (96, 116). This study aims to generate and compare the epigenomic profiles of H3K27me3 and H3K4me3 along with the binding site profiles of JARID2 (preferably both cleaved and un-cleaved forms) and EZH2 in matched pairs of primary and recurrent GBM samples using epigenomic mapping approaches. An overview of these, including workflow and subsequent sequencing data analysis, is provided below.

1.6 Epigenomic mapping technologies

1.6.1 Sequencing technologies

Studies on genomes, epigenomics, and transcriptomics have been made possible only because of the remarkable developments in high-throughput sequencing technologies over the past 20 years (122, 123). Traditionally, DNA sequencing information was elucidated using a low throughput

technique called Sanger sequencing. This technique was first introduced by Frederick Sanger and his colleagues in the 1970s (122, 124).

A further significant development in DNA sequencing happened around 1977, when Sanger's chain-termination technique was developed to involve the use of chemical analogues of the deoxyribonucleotides (dNTPs) known as dideoxyribonucleotides (ddNTPs). This chemical lacks the 3' hydroxyl group that is required for the elongation of the DNA strand, and therefore, the lack inhibits its binding with the 5' phosphate of the next dNTP. In this technique, Sanger et al. mixed radiolabelled ddNTPs into a DNA elongation reaction and performed four parallel polymerase reactions consisting of each single ddNTPs which is then ranning on polyacrylamide Gel. Because there will be a radioactive band in the appropriate lane at that location of the gel, autoradiography may be used to determine the nucleotide sequence of the original template (122, 124). In the following years, Sanger sequencing underwent a number of improvements which involved the replacement of radiolabelled molecules with fluorophores and the incorporation of capillary-based electrophoresis for better detection. These improvements aided to the development of high throughput sequencing (HTS) technologies or sometimes called next generation sequencing (NGS). These technologies can sequence multiple DNA molecules simultaneously, enabling hundreds of millions of DNA molecules to be sequenced at a time (122). The fundamental concepts behind NGS are similar to Sanger sequencing. Nucleotides that have been dye-labelled are added to the growing DNA strand, and the colour of the dye is used to identify each base (124). NGS works by fragmenting the genome (i.e. DNA or RNA sample) into smaller pieces which are treated by enzymes to synthesize the complementary DNA strands. The latter is then subjected to DNA sequencing. A typical NGS workflow involves DNA fragmentation, library preparation, clonal amplification of DNA libraries, sequencing and data analysis (125). The first commercial NGS machine was the 454 machine, introduced by Life Science in 2004. Other platforms emerged later such as Illumina/Solexa, ABI SOLiD, HiSeq X, NextSeq 500, NovaSeq and Ion Torrent (126).

One of the areas on which high throughput sequencing technologies has had high impacts on it is genome-wide mapping of chromatin accessibility and histone (123, 127). Chromatin immunoprecipitation followed by sequencing of the enriched DNA (ChIP-seq) is the first of these technologies that made the identification and characterisation of transcription factor binding sites and genome-wide histone marks possible in a large region of the genome with high resolution (117, 127). Here, I will discuss the most commonly used HTS technologies in more detail.

1.6.2 Chromatin immunoprecipitation coupled with sequencing (ChIP-seq)

ChIP-seq has been the method of choice for identifying and mapping histone modifications on a genome-wide scale with higher genomic coverage and spatial resolution since 2007 (128, 129). The basic and fundamental procedure of this technique is mainly depending on the isolation and enrichment of target proteins by immunoprecipitation which is then purified and added to the universal adapter for PCR amplification, followed by sequencing using HTS (130). In short, ChIP-seq starts with crosslinking of proteins and their bound DNA using formaldehyde followed by shearing the crosslinked chromatin into small fragments (~200-600bp) using sonication. DNA fragments associated with the protein of interest are pulled down/immunoprecipitated using protein-specific antibodies. These pulled-down fragments are subjected to reversal crosslinking and the resulting DNA fragments are sequenced using HTS (128, 131) (**Figure 1.3**).

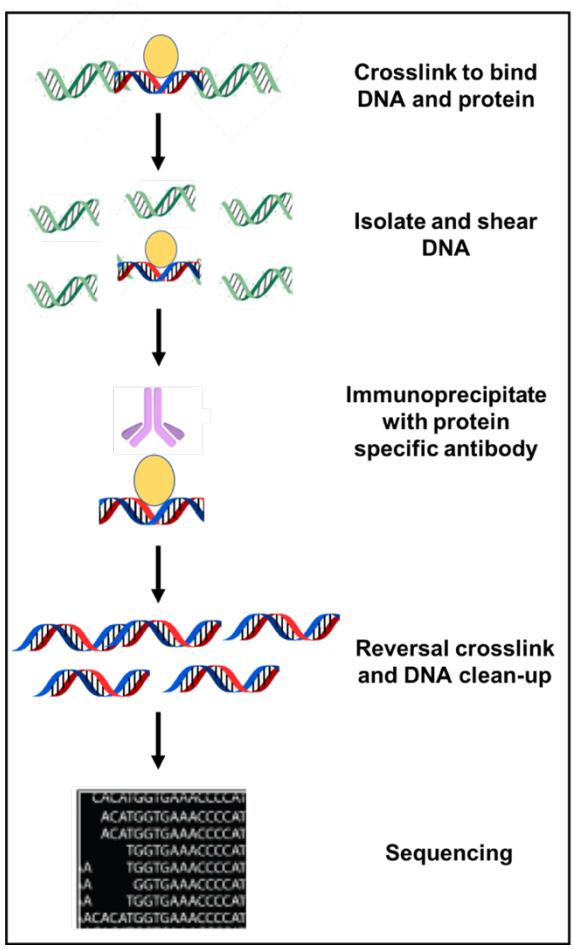


Figure 1-3: Schematic representation of the ChIP-seq workflow. Chip-seq starts with crosslinking of proteins and their bound DNA using formaldehyde followed by shearing the crosslinked

chromatin into small fragments (~200-600bp) using sonication. DNA fragments associated with the protein of interest are pulled-down/ immunoprecipitated using protein-specific antibodies. These pulled-down fragments are subjected to reversal crosslinking and the resulting DNA fragments are sequenced using HTS.

Despite the fact that ChIP-seq has become a gold standard for mapping histone modifications across the genome with higher resolution and less noise (132), it is still fundamentally limited. The primary reported limitation is that it requires abundant starting materials in the range of 1-20 million cells per immunoprecipitation to assess transcription factors or histone proteins. Additionally, high cost of sequencing and reagents are still limiting factors for most researchers, though the situation has improved and the cost is decreasing with the development of new generation sequencing technologies. Moreover, issues related to experimental design in terms of control sample, the sequencing depth and the quality of the antibodies poses substantial challenges in applying ChIP-seq (131-133). The ENCyclopedia Of DNA Elements (ENCODE) consortium published a set of technical design guidelines and considerations for ChIP-seq experiments in order to reduce bias and background noise and aid the consistent generation of high quality genome-wide data (134). The detail of these guidelines is provided below.

1.6.2.1 ENCODE guidelines for proper ChIP-seq experimental design

The analysis of each ChIP-seq experiment requires a suitable control data set since DNA breakage during sonication is not constant and generates uneven fragmentation of the genome. It has been noted that some regions of open chromatin tend to be fragmented more easily than closed regions. In addition, platform-specific sequencing efficiency biases may lead to non-uniformity. In order to assess the relevance of a peak in the ChIP-seq profile, it should be compared to the same location in a matched control sample (135, 136). There are two reported types of control samples that are frequently utilized: input DNA (a portion of the DNA sample that has been cross-linked and fragmented under the same conditions as the immunoprecipitated sample) and a “mock” IP DNA (DNA sample that has been immunoprecipitated with a control antibody that reacts with an irrelevant, non-nuclear antigen such as immunoglobulin, G (IgG). These controls have been tested for different types of artifacts and the results show that there is no consensus on which type is more convenient and suitable for downstream analysis. Input DNA has been used widely in nearly all ChIP-seq analyses so far (134, 135).

With regards to antibody quality, a successful ChIP experiment and the value of ChIP-seq data critically depend on the quality of the antibodies used. The selection of antibodies that can target the protein of interest with high sensitivity and specificity was addressed extensively in ENCODE guidelines. Antibodies with higher sensitivity and specificity will give a high enrichment level in comparison with the background which facilitates the accurate detection of binding events (134, 137). There are several commercially available antibodies, some of which are marked as ChIP grade, however, their quality is highly variable and can even differ across batches of a certain antibody (137). Multiple commercial antibodies have been tested as part of ENCODE project, and the validation results revealed that 20–35% of these antibodies were unsatisfactory (135). Therefore, ENCODE developed specific guidelines to assess the specificity of the antibodies used in ChIP applications.

The suggested primary mode of assessment for histone modifications involves a standard immunoblot analysis on protein lysates from whole cell extract (WCE) (138). To pass the immunoblot test, 50% of the signal should be observed in a single band and ideally, this band should correspond to the expected size of the target protein. In the absence of a band at the expected size or in the presence of multiple non-specific bands, further tests should be performed. Therefore, ENCODE listed secondary modes of characterization to assess the sensitivity and specificity of the tested antibodies. This involves knockdown or knockout of the target protein using either small interfering ribonucleic acid (siRNA) or short hairpin RNA (shRNA) followed by immunoblot analysis, or immunoprecipitation of an epitope-tagged version of the protein (IP) followed by Western or mass spectrometry (MS) (**See chapter 2, section 2.2.5 for the detail of these techniques**) (139). At least one successful characterization is required to ensure the specificity of the antibody. For the knockdown approach, the antibody passes the test if a reduction in signal of > 50% is observed in comparison to the control sample and no reduction is observed in the control knockdown sample (i.e. scrambled siRNA). If the antibody fails to pass these parameters, immunoprecipitation of an epitope-tagged version of the protein followed by Western analysis or mass spectrometry (MS) can be used. This is considered a powerful approach for identifying physiologically relevant protein-protein interactions. This approach involves the detection and comparison of an overexpressed or exogenous epitope-tagged version of protein with the endogenous version (138). The antibody passes this test if a comparable and clear band is observed for the overexpressed (larger band) or exogenous epitope-tagged protein (band seen with antibodies against the protein tag) at the expected size of the target protein. Despite the fact that these approaches provide confidence

concerning an antibody's acceptance, the validation is laborious, time-consuming and expensive (134, 137).

In addition to ChIP experimental design guidelines, ENCODE documented parameters that should be considered when evaluating and analysing ChIP-seq data (134). Computational analysis pipeline steps, which are needed for the comprehensive characterization of epigenetic states in the ChIP-Seq profiling dataset, were ascertained and the details and potential tools for each step are provided in the following section.

1.6.2.2 Computational pipeline for ChIP-seq data analysis

ChIP-seq generates a massive amount of data that requires computational analysis to identify the epigenetic landscapes and key chromatin signatures in an accurate manner (135). A generic computational ChIP-seq pipeline includes raw data quality assessment, trimming low-quality reads, sequence alignment, peak detection and data visualization, most often using either the UCSC genome browser or integrative genomic viewer (140). Once the data are visualized, different downstream analyses can be performed either by analysing the resulting peaks, such as peak annotation in relation to the transcription start site (TSS), or characterizing and annotating the chromatin states (141).

In general, raw sequencing data generated from the NGS platform contains short DNA sequences with quality scores (140). A ChIP-seq pipeline starts with assessing the quality of raw sequencing data, and the most commonly used tool is FastQC (142). This tool provides a comprehensive overview of the main quality metrics that help the user to spot any issue with the data such as the quality score of the sequence, GC content, sequence duplication level, overrepresented sequence and the percentage of adapter content. The second step of the pipeline is filtering and trimming low-quality reads and adapter sequences and there are several well-established tools for data trimming such as Trimmomatic, cutadapt and trim galore (143). Trimmed reads are then mapped to an appropriate reference genome and there are many mapping tools that have been developed. This includes Burrows-Wheeler Aligner (BWA) (144), bowtie (145), Efficient Large Scale Alignment of Nucleotide Database (ELAND) (146), MAQ (147) and SOAP2 (148). The selection of the appropriate tool is mainly depending on the speed requirement, the sequencing platform and the hardware resources (140, 149). Mapped reads are then subjected to several quality checks in order to assess and evaluate ChIP-seq data as suggested by ENCODE (134). This includes:

- global ChIP enrichment in terms of the fraction of all mapped reads that are located in identified peak regions (FRiP);
- the fraction of nonredundant mapped reads (NRF) which is defined as the proportion of reads that uniquely map to specific locations in the genome to all reads that are uniquely mappable;

$$\text{NRf} = \frac{\text{Number of unique start positions of uniquely mappable reads}}{Number\ of\ \text{uniquely mappable reads}}$$

- the normalized strand coefficient (NSC), ratio of the background cross-correlation to the fragment-length cross-correlation peak;
- the relative strand correlation (RSC), the proportion between the fragment-length peak and the read-length peak.

These are the main quality metrices that were suggested by ENCODE to inspect the quality of ChIP-seq data (150, 151). A measure of library complexity in terms of PCR bottlenecking coefficients 1 and 2 (PBC1 and PBC2) was also suggested. These measure the skewness in the distribution of read counts per location is towards 1 read per location (151). PBC1 is defined as:

$$\text{PBC 1} = N_1/N_{distinct}$$

Where N1 = number of genomic regions where a single read map precisely and uniquely and $N_{distinct}$ = the number of genomic regions where at least one unique mapping reads maps.

Whereas, PBC2 is defined as the number of genomic regions where single read maps uniquely to the number of genomic regions where two reads map uniquely (PBC 1 = N1/N2)

According to ENCODE guidelines and practices, for good quality data (134):

- FRiP should be < 3;
- NRF should be > 0.9;
- NSC should be > 1;
- RSC should be > 1;
- PBC1 should be > 0.9;
- PBC2 should be >3.

A step-by-step ChIP-seq analysis pipeline is provided in **Chapter 2, section 2.2.8.1**.

The primary objective of ChIP-seq analysis is to identify the functional components of the human genome and explain the binding features of the target proteins accurately. Several peak calling tools and algorithms were developed and published to fulfil this purpose such as model-based analysis of ChIP-seq (MACS), SPP, PeakSeq and CisGenome (140, 152). These tools were optimized and tested by ENCODE on publicly available data on 12 histone modifications and the results suggested that are no major differences among these tools and that the performance of each peak calling tool mainly depends on the peak calling parameters that are chosen by the submitter. These parameters include peak position, signal value in terms of fold enrichment and the p-value. Peak calling tools, in general, determines areas that are enriched as a result of protein binding by using the coverage properties of ChIP and Input samples (153).

MACS2 is one of the most commonly used peak callers (154). It performs several steps ranging from filtering duplicated reads and building the peak model to peak detection and statistical assessment to evaluate if the location of enrichment is likely to be a true binding site. MACS2 workflow started by removing the duplicated (i.e. redundancy) reads at the same locations (i.e. reads with the same coordination and strands) as these reads may interfere with the true ChIP signal. MACS2 offers a variety of options for handling these reads and by default, only one read is kept at each location. Then MACS2 builds the model by scanning the whole dataset of the ChIP sample and simulating the distance between the paired forward and reverse strand peaks. Finding enriched regions, those with high confidence fold enrichment (M-fold) than the background, in the genome is done by sliding a window across the genome. Next, MACS2 randomly uses 1000 regions with fold enrichment between 10 and 30 to build the model between the positive and negative strands peaks and estimates the fragment length ‘d’ (**Figure 1.4**).

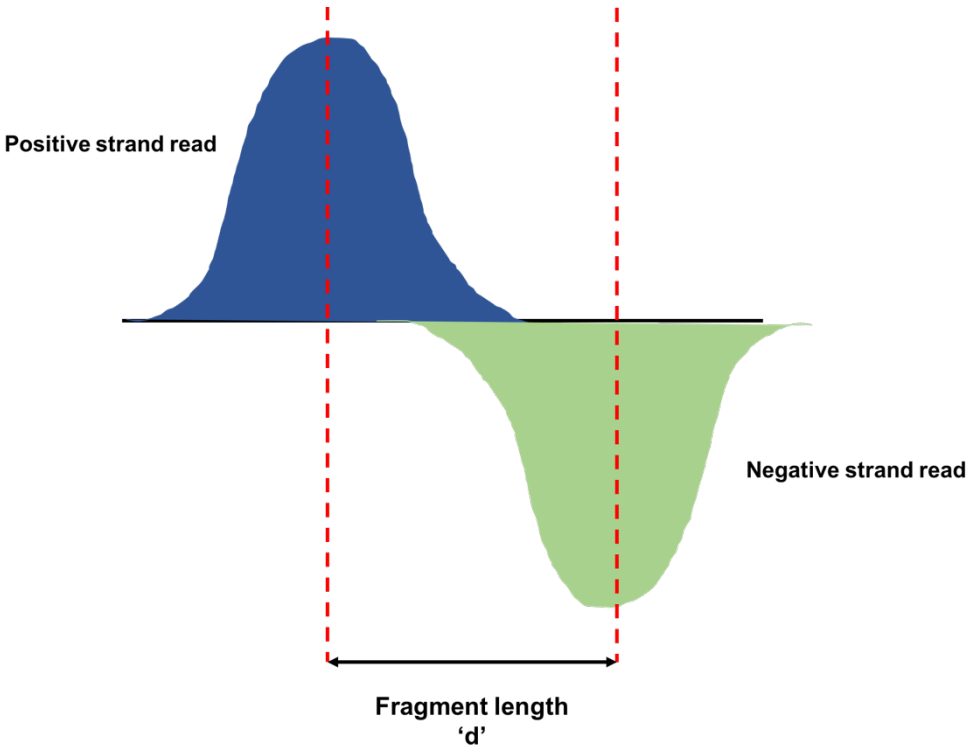


Figure 1-4: Schematic diagram showing how MACS2 build a peak model using the estimated DNA fragment length 'd'.

After that, MACS2 extends the readings in the 3' direction by *d/2* during the real peak detection phase. Then, in the presence of the control sample, MACS2 linearly scale the control and ChIP sample to the same read number. Next, potential peaks are chosen by scanning the genome once again with a window size that is twice the fragment length (i.e. *2d*). To account for regional biases in reading background levels, MACS2 computes a p-value for each peak using a dynamic (λ)Poisson distribution. If a control sample is available, it is used to calculate the local background. Finally, p-values are calculated using the Benjamini-Hochberg correction (155). MACS2 workflow is summarized in **Figure 1.5**.

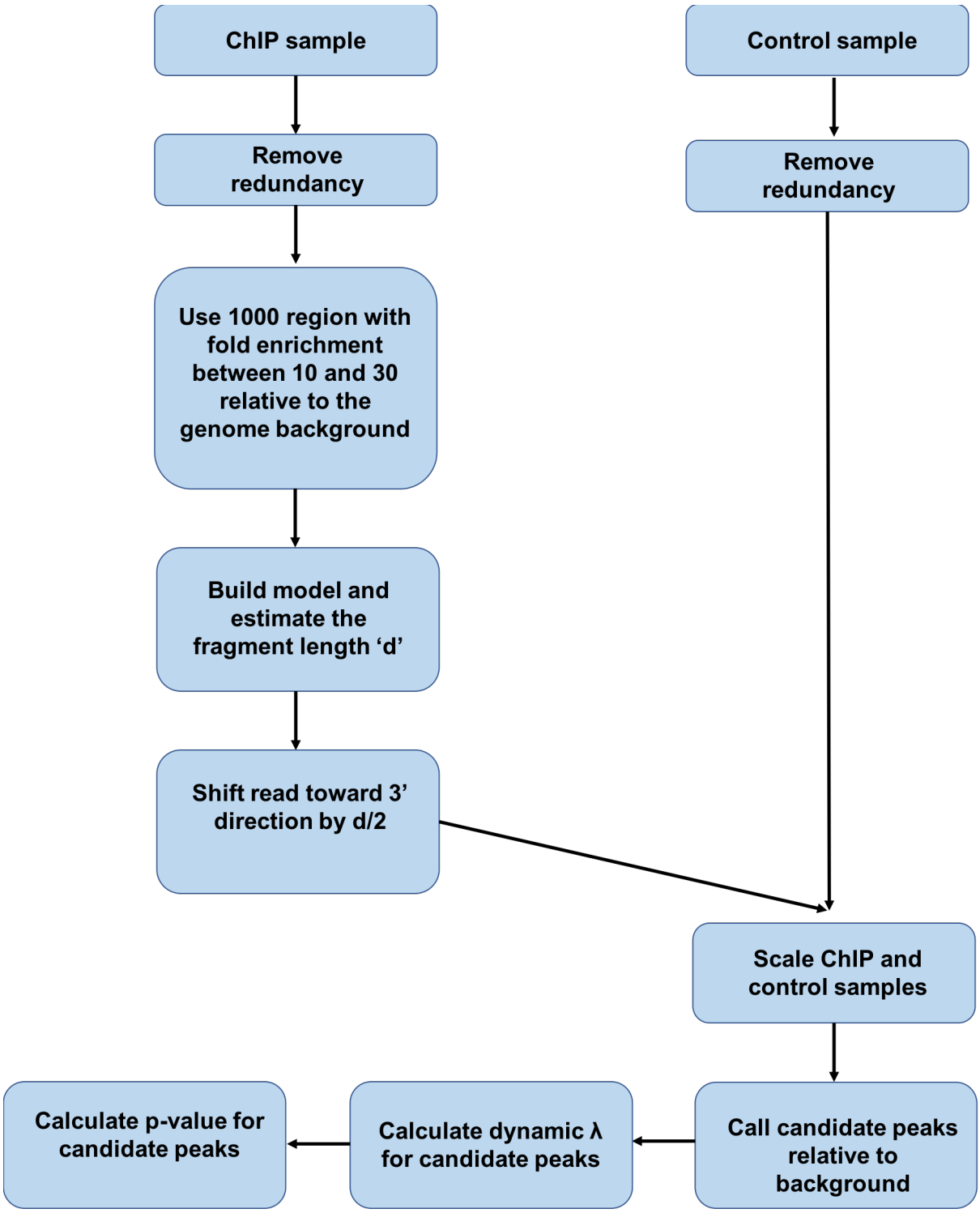


Figure 1-5: General workflow of peak calling with MACS2.MACS2 performs multiple steps ranging from removing redundancy reads and building the peak model to peak detection and statistical assessment to determine if the location of enrichment is likely to be a true binding site.

In the presence of replicated samples, ENCODE suggested the use of an irreproducibility discovery rate (IDR) to assess peak consistency between replicates (134). The IDR framework provides very

reliable thresholds based on reproducibility and unifies a method for measuring the reproducibility of findings from duplicate experiments. The IDR technique generates a curve that quantitatively examines when the results are no longer consistent across duplicates, in contrast to the typical scalar metrics of reproducibility. The IDR approach, to put it simply, compares two sorted ChIP-seq peaks. These sorted lists should offer identifications over the complete range of high confidence/enrichment (signal) and low confidence/enrichment (noise), i.e. they shouldn't be pre-thresholded. The IDR approach then fits the bivariate rank distributions over the repetitions to distinguish between signal and noise based on a specified level of rank consistency and identification repeatability, or the IDR threshold (156). According to ENCODE guidelines, the replicates are considered concordant if the consistency ratio is less than 2 (134).

Despite the presence of different peak calling tools and algorithms, these tools summarize only one ChIP-seq dataset (one experiment with a single antibody) at a time in which a single mark can be studied in isolation through the identification of narrow or broader domains. Therefore, a segmentation approach called ChromHMM was developed to conceptually integrate and combine information of multiple marks across multiple datasets (157).

1.6.2.2.1 Integrative Epigenomic data analysis with integration with ChromHMM

ChromHMM enables the researcher to characterize and annotate the chromatin states across multiple cell types using epigenomic information (157). It is based on a multivariate Hidden Markov Model (HMM) which is a probabilistic model that specifically models multiple 'observed' events based on invisible 'hidden' states. HMM uses the probabilistic nature of a multi-state model to identify each segment based on the combinatorial presence and absence of multiple marks and the spatial constraints of how these mark combinations occur relative to one another across the genome (157, 158). There are actually several tools that implement HMM in the recognition of chromatin states, aside from ChromHMM, including Segway and EpiCSeg (157, 159, 160). However, ChromHMM is the most widely used tool for chromatin state identification and annotation and more specifically in ENCODE and RoadMap projects (157). ChromHMM divides the genome into intervals of 200 nucleotides by default, which is corresponding to the resolution of a nucleosome and spacer area, however, the interval size can be changed by a user-specified parameter. Then, it evaluates whether each mark is present or absent for each genomic interval based on the significance of the observed read count compared to a Poisson background distribution.

ChromHMM learns a chromatin state model and an annotation of state occurrences across the genome using the generated presence-absence calls (159).

ChromHMM differs from the remaining tools in a number of ways. Unlike other tools which model the signal levels of each mark independently, ChromHMM focuses its modelling power on combinations of epigenomic marks by employing binary presence/absence input features. This has made it possible for ChromHMM to identify chromatin states such as a state associated with Zinc finger genes and putative bivalent promoter states that are frequently missed by other methods when applied to the same datasets. Additionally, ChromHMM can be used for large-scale applications because of its reliable and effective implementation, which includes multi-core parallelization. This is proved by learning models based on a dozen markers across more than 100 cell and tissue types while using the whole genome for training. Furthermore, ChromHMM is easy to use and install and can work with aligned reads to generate chromatin state enrichment analysis (157). Despite these characteristics, ChromHMM possesses some limitations. First, there is a significant loss of information when the read count is converted into a binary value, as it is impossible to differentiate between different degrees of activity. Second, choosing a threshold is crucial for the final segmentation, however, there is no clear way of deciding which threshold to use. Third, the model assumes that the presence of one mark is independent of the other mark which is inconsistent with other observations (160).

1.6.3 Cleavage under targets & release using nuclease (CUT&RUN)

To alleviate some of the reported limitations of ChIP-seq (**See section 1.6.1**) and to enhance the profiling of protein-DNA interactions accurately, a new method was developed by Skene and Hanikof. Cleavage under targets and release using nuclease (CUT&RUN) is a feasible replacement for ChIP-seq that relies on a target-specific primary antibody and micrococcal nuclease (MNase) to isolate the binding sites of DNA-protein complexes. An overview of the CUT&RUN technique, the general workflow along with the analysis pipeline are presented in detail in **Chapter 6**.

As the purpose of this study is to understand JARID2-associated epigenetic remodelling in GBM in recurrent versus primary tumours, I decide to use the ChIP-seq, as a traditional approach, along with CUT&RUN to facilitate genome-wide profiling of histone marks H3K27me3 and H3K4me3 along with EZH2 and JARID2 as a regulatory element in matched pairs of primary and recurrent GBM tumours.

2. Hypothesis

Mechanisms associated with JARID2 facilitate chromatin remodelling in GBM cells, enabling them to adapt to treatment. Characterising and comparing the binding profiles of JARID2 and EZH2 (the catalytic subunit of PRC2) and the histone marks H3K27me3 and H3K4me3 in matched pairs of primary and recurrent GBM samples will give insights into the role of chromatin remodelling in conferring treatment resistance in GBM.

3. Aims and objectives

The main aim of this project was to develop a workflow that would enable the generation of genome-wide chromatin state map of H3K27me3, H3K4me3, JARID2 and EZH2 binding for paired primary and locally recurrent IDH wildtype GBM brain tumours from patients that had received standard treatment.

To address this aim I had to achieve several objectives, the results of which each constitutes a results chapter in this thesis

Objective 1: Validate antibodies against JARID2, as this has not been commonly used in chromatin profiling

Objective 2: Develop a computational pipeline for analysing chromatin profiling data in paired samples

Objective 3: Apply the optimised workflow to a pair of patient samples to begin gaining insight into treatment resistance in GBM

Objective 4: Ascertain whether CUT&RUN could be applied to patient GBM samples, as a potential superior approach to ChIPseq

Chapter 2
Materials and Methods

2.1 Materials and Reagents

2.1.1 Reagents

2.1.1.1 Radio-immunoprecipitation assay (RIPA) lysis buffer

Tris-HCl (pH 8.0)	50mM	Sigma-Aldrich
NaCl	150mM	Sigma-Aldrich
NP-40	1% [v/v]	Sigma-Aldrich
SDS	0.1% [v/v]	Melford laboratories Ltd
Sodium Deoxycholate	0.1% [v/v]	BDH
proteinase inhibitor cocktail		Sigma-Aldrich (P8340)

2.1.1.2 IP incubation buffer

NaCl	25mM	Sigma-Aldrich
Tris-HCl (pH 8.0)	20mM	Sigma-Aldrich
EDTA	2mM	Ambion
Glycerol	10% [v/v]	Sigma-Aldrich
Ethanol	10% [v/v]	Sigma-Aldrich
Proteinase inhibitor cocktail	1x	Sigma-Aldrich (P8340)

2.1.1.3 Wash buffer for IP

NaCl	150mM	Sigma-Aldrich
Tris-HCl (pH 8.0)	50mM	Sigma-Aldrich
EDTA	0.5mM	Ambion
NP-40	0.1% [v/v]	Sigma-Aldrich
Proteinase inhibitor cocktail	1x	Sigma-Aldrich (P8340)

2.1.1.4 Tris Acetate-EDTA (TAE) buffer (50x)

Tris HCl (pH 8.3)	2M	Sigma-Aldrich
Glacial acetic acid	0.9M	Sigma-Aldrich
EDTA	0.5mM	Sigma-Aldrich

2.2 Methods

2.2.1 Ethical considerations and patient samples

All tumour samples that were available for this research project in the form of fresh frozen tissues were used in accordance with ethical approval acquired from NHS NRES Committee South Central - Oxford A (REC 13/SC/0509).

2.2.2 Antibodies selection

A key component, required for all of my project aims, is to first validate JARID2 antibodies for ChIP and CUT&RUN applications. This is critical to ensure that the antibody-antigen interaction is specific and suitable for its intended applications. Four JARID2 antibodies were selected that target different regions of the JARID2 protein, since this would provide potential for analysing different protein species and may increase the chances of the epitope being accessible in the context of ChIP and CUT&RUN. The selection was based on the available reviews and whether it has been tested for the planned applications (i.e. Chip-seq and CUT&RUN assays). JARID2 antibodies that have been used in this study are listed in detail in **Table 2.1**.

As per the ENCODE guidelines, the specificity of these antibodies was confirmed by siRNA knockdown of JARID2 in M059K, HEK293T and GBM63 cell lines, and subsequent assessment of western blots to determine whether the antibody in question indicated a similar reduction at the protein level. Antibodies used in ChIP-seq assay (i.e. H3K27me3, H3K4me3 and EZH2 were provided by Active Motif (**Table 2.2**).

For CUT&RUN, H3K4me3 and IgG antibodies were provided in the CUT&RUN kit, whereas, the H3K27me3 antibody was selected based on other publications which reported the success of this antibody in different ChIP-seq and CUT&Tag experiments. **Table 2.2** lists all the primary and secondary antibodies used in this study.

JARID2 targeted regions	Species raised in	Polyclonal/ monoclonal isotype	WB dilution (1/X)	IP dilution	Supplier	Catalogue number
1-100 aa of the N-terminal regions of isoform 1 (140kDa)	Rabbit	Polyclonal, IgG	1:1000	N/A	Novus Biology	NB100-2214
Around Asp 1114 of isoform 1 (140kDa) and isoform 2 (120kDa)	Rabbit	Monoclonal, IgG	1:1000	N/A	Cell signaling Technology	D6M9X
100-200aa of isoform 1 (140kDa)	Rabbit	Monoclonal, IgG	1:1000	5ug per IP	Abcam	Ab 192252
1130- 1230 aa of isoform 1 (140kDa)	Mouse	Polyclonal, IgG	1:1000	N/A	Abcam	Ab 93288

Table 2-1: List of JARID2 antibodies used in this study.
Table includes the JARID2 antibodies, their targeted regions, the species that the they raised in, the type of isotype, western blot and IP dilutions, the supplier and the catalogue number.

Antigens	Species raised in	Polyclonal/ monoclonal isotype	WB dilution (1/X)	IP dilution	Applications	Supplier	Catalogue number
H3K27me3	Rabbit	Polyclonal	N/A	N/A	CUT&RUN	Diagenode	C15410195
H3K27me3	Rabbit	Polyclonal	N/A	N/A	ChIP-seq	Active Motif	39155
H3K4me3	Rabbit	Monoclonal	N/A	N/A	CUT&RUN	Cell signaling Technology	86652
H3K4me3	Rabbit	Polyclonal	N/A	N/A	ChIP-seq	Active Motif	39159
EZH2	Rabbit	Polyclonal	N/A	N/A	ChIP-seq	Active Motif	39901
Flag	Mouse	Monoclonal	1:1500	5ug per IP	WB/IP	Millipore	F1804
ß-actin	Mouse	Monoclonal	1:10000	N/A	WB	Sigma-Aldrich	A1975
GAPDH	Rabbit	Monoclonal	1:10000	N/A	WB	Cell signaling Technology	21118s
HRP-linked antibody	Rabbit	Monoclonal/ polyclonal	1:2000	N/A	WB	Cell signaling Technology	7074s
HRP-linked antibody	Mouse	Monoclonal/ polyclonal	1:2000	N/A	WB	Cell signaling Technology	7076P2

Table 2-2: List of all other primary and secondary antibodies used in this study.
Table includes the primary and secondary antibodies that were used in this study, the species that the they raised in, the type of isotype, western blot dilutions, IP dilutions, applications, the supplier and the catalogue number.

2.2.3 Cell cultures

2.2.3.1 Cell lines cultures

Routine cell culture was carried out using standard aseptic techniques in a tissue culture room. M059K and HEK293T cells were available and purchased originally from the American Type Culture Collection (ATCC). M059K cells were routinely cultured in T-75 flasks in Dulbecco's Modified Eagle Medium/Nutrient Mixture F-12 (DMEM/F12, ThermoFisher Scientific, Cat No: 11320033) supplemented with 10% Fetal bovine serum (FBS, ThermoFisher Scientific, Cat No: 10270106), 0.5 mM non-essential amino acid (NEAA, ThermoFisher Scientific, Cat No: 11140050) and 1 mM sodium pyruvate (ThermoFisher Scientific, Cat No: 11360070). Whereas, HEK293T cells were routinely grown in Dulbecco's Modified Eagle Medium (DMEM, Sigma-Aldrich, Cat No: D6429) containing 10% FBS and 1% Penicillin/Streptomycin (ThermoFisher Scientific, Cat No: P0781). Cell cultures were incubated at 37^{o}C in 5% CO_2.

The cells were passaged when they were over 70% confluency. During cell passaging, the media was removed and the cells were washed twice with 1x Dulbecco's phosphate-buffered saline (PBS, Sigma-Aldrich, Cat No: D8537). Cells were trypsinized with 1ml of 1x pre-warmed Trypsin/EDTA (Sigma-Aldrich, Cat No: 59418C) and incubated in a 5% CO_2 incubator at 37^{o}C for 5 min. To inactivate the trypsin, 9 ml of fresh pre-warmed media was added to the flask. The cells were centrifuged at 200 x g for 5 minutes at room temperature to collect the pellet, and then they were resuspended in 10 ml of new medium. According to the supplier's recommendations, both cells were sub-cultured in a split ratio of 1:8 every two to three days.

2.2.3.2 Primary cell cultures

2.2.3.2.1 Poly-L-Ornithine and laminin coating protocol for primary cell cultures

To promote neural cell growth and enhancement of attachment of GBM primary cells plastic ware was coated with materials for poly-L-Ornithine and laminin. Specifically, Flasks were coated first by the addition of a 10ml working solution of poly-L-ornithine (10µg/ml, Sigma-Aldrich, Cat No: P3655) and incubated at RT for an hour. Flasks were rinsed once with TC grade water (Sigma-Aldrich, Cat No: W3500) and 10ml working solution of laminin (2µg/ml, Sigma-Aldrich, Cat No: L2020) was added and the flasks were kept in room temperature overnight and stored at -20^{o}C.

2.2.3.2.2 Cell cultures

Prior to culturing primary cells, coated flasks were thawed and the laminin solution was removed. Then the flasks were rinsed once with PBS. GBM63 cells were maintained in Neurobasal-A serum free-media (ThermoFisher Scientific, Cat No: 10888022) supplemented with N2 supplement (Invitrogen, Cat No: 17502048) and B-27 (Invitrogen, Cat No: 17504044) at a final working concentration of 0.5x. These two supplements support the proliferation and survival of neural cells in the culture. In addition, 20ng/ml each of working concentration of human recombinant epidermal growth factor (EGF, R&D systems Cat No: 236-EG-200) and basic fibroblast growth factor (bFGF, Peprotech Cat No: 100-18B-1000) were added to the media and incubated under the same conditions as described in **Section 2.2.3.1**. The cells were cultured for up to 6 – 12 weeks to achieve sufficient confluency (70-80% confluency) the cells were passaged to a new culture to avoid over confluency as described above (**see Section 2.2.3.1**).

2.2.4 Molecular biology technique

2.2.4.1 Transfection of cells with lipofectamine RNAiMAX and siRNA to knockdown gene expression

An initial attempt to validate JARID2 antibodies is to knock down JARID2 gene using a small interfering RNA (siRNA) assay. ON-TARGETplus JARID2 siRNA SMARTpool and On-TARGETplus non-targeting pool (Dharmacon-horizon Discovery, Cat No: L-009244-00-0005 and Cat No: D-001810-10-05, respectively) were used for the transfection experiment. The SMARTpool is a mixture of 4 siRNA provided as a single reagent and it targets 4 exons. The exact localization of siRNA targets at the human genomic DNA level (**Table 2.3**) were checked using the NCBI Basic Local Alignment Search Tool (BLAST) and the Ensemble genome browser. Upon arrival, siRNAs were resuspended in 250µl of RNase-free 1x siRNA buffer (Dharmacon-horizon Discovery, Cat No: B-002000-UB-100) for a final concentration of 20µM.

siRNA target	Sequence
JARID2	GAAGAACGGGUGGUACGUA GCUCAGGACUUACGGAAAC GACAAAGGCGUCCUCAAUG

	AAUGAAGCGUCGCCAUAUA
Non-target	UGGGUUUACAUGUCGACUA UGGUUUACAUGUUGUGUGA UGGUUUACAUGUUUUCUGA UGGUUUACAUGUUUUCCUA

Table 2-3: Sequences of JARID2 and non-target siRNA used for knock-down
Table includes the siRNA target and the related siRNA target sequences.

M059K cell line was transfected using the following transfection protocol. On the day of transfection, the cells were seeded in a 6-well culture plate at a density of 2 x 10^5 cells per well. Simultaneously, transfection complexes were prepared by adding 15µl of lipofectamine RNAiMAX (ThermoFisher Scientific, Cat No: 13778075) to 250µl Opti-MEM reduced serum media (Gibco, Cat No: 11564506) in one Eppendorf. Whilst, 15µl the 20µM stock siRNA was added to 250µl of Opti-MEM in another Eppendorf for a final concentration of 150 nM. The 2 solutions were then mixed thoroughly and incubated at room temperature for 20 min. Finally, 500µl of the transfection mixture was added to each well-containing 1.5ml of the diluted cell suspension and cultured for up to 72 hrs.

The experiment was repeated as described above by this time HEK293T and GBM63 cells were transfected with 50 nM for 72 hrs (This task was completed by Marilena Elpidorou, a postdoc in the Stead group). The workflow of the transfection is summarized in **Figure 2.1**.

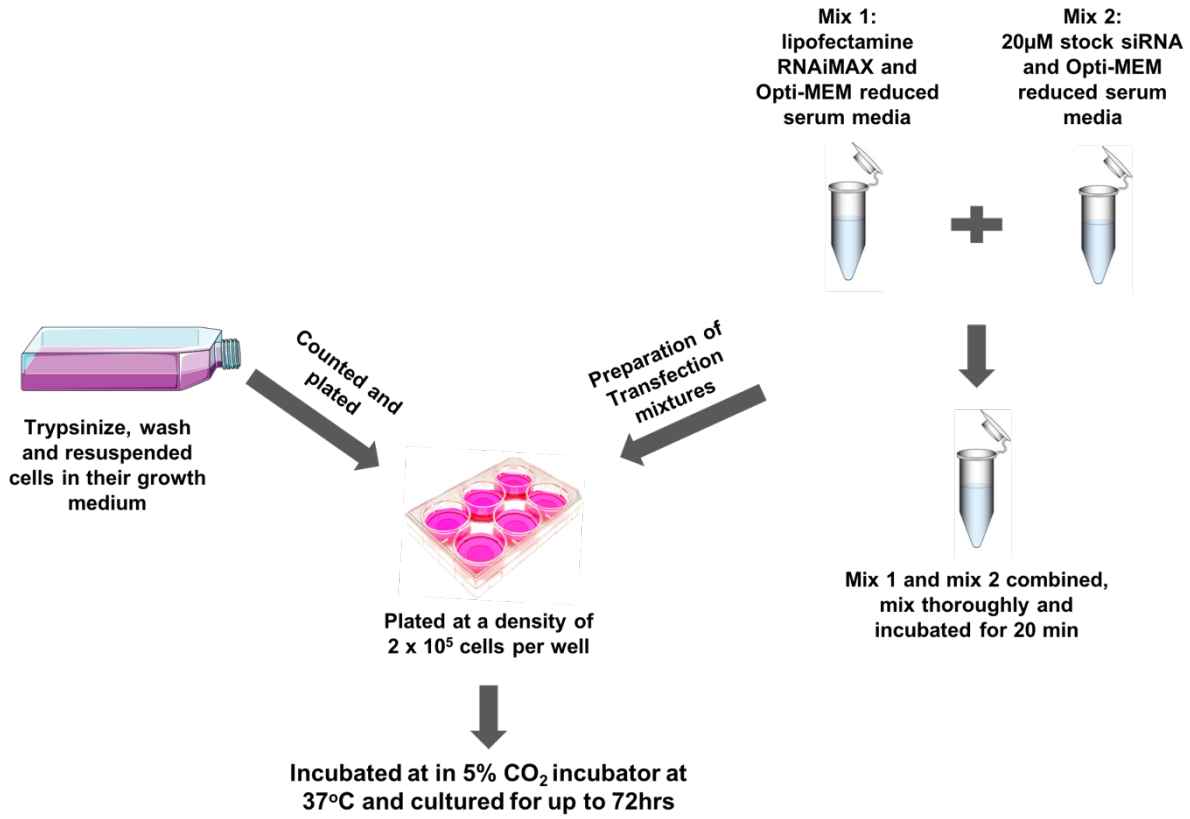


Figure 2-1: Schematic representation of siRNA reverse transfection procedure for 6-well plate. On the day of transfection, M059K cells were counted and plated in 6-well culture plates at a density of 2 x 10^5 cells per well after being suspended in a full growth medium. Simultaneously, transfection complexes were made by adding 10µl and 15µl of lipofectamine RNAiMAX to 250µl of Opti-MEM reduced serum media in one Eppendorf and 20 µM of stock siRNA (final concentrations of 100nM and 150nM per well) to 250µl of Opti-MEM in another Eppendorf. The contents of both Eppendorfs were then carefully combined and incubated for 20 minutes at room temperature. Finally, 1.5ml of the diluted cell suspension and 500µl of the transfection mixture were added to each well, and the cells were then grown for up to 72 hrs.

2.2.4.2 Total RNA extraction, purification and quantification

Following the transfection of siRNAs into M059K cells, the cells were lysed directly using the RNeasy Plus Mini kit (Qiagen, Cat No: 74134) following the manufacturer's protocol at the time points of 24, 48 and 72 hrs. In short, cells were trypsinized as described in **Section 2.2.3.1** followed by resuspension of pellets in RLT plus buffer to lyse the cells. the lysates were transferred into a QIAshredder spin column placed in a 2ml collection tube, centrifuged for 2 min at maximum speed (i.e. > 8000 x g) and the homogenized lysates were transferred to a gDNA eliminator spin column

placed in a 2ml collection tube. It was then centrifuged for 30s at > 8000 x g and the flow-through was placed into a new 2ml collection tube. Next, 600µl of 70% of ethanol (Sigma-Aldrich) was added to the flow-through, mixed well by pipetting and transfer the mix into a new RNeasy spin column placed in a 2ml collection tube. The mixture was centrifuged for 15s at > 8000 x g and the flow-through was discarded. Then, 700µl of RW1 buffer was added to the RNeasy spin column, centrifuged for 15s at > 8000 x g to wash the spin column membrane and the flow-through was discarded. After that, 500µl of RPE buffer was added to the RNeasy spin column and centrifuged for 15s at > 8000 x g. This step was repeated again and then the RNeasy spin column was placed into a new 1.5 ml collection tube followed by the addition of 25µl of RNase-free water to elute the DNA and centrifuged for 1 min at > 8000 x g. The eluted RNA was then quantified and the quality was assessed using Nanodrop-1000 Instrument (ThermoFisher Scientific).

2.2.4.3 Preparation of cDNA for the quantitative polymerase chain reaction (qPCR)

Quantitative polymerase chain reaction (qPCR) was performed after initially reverse transcribing RNA (up to 400µg) into complementary DNA (cDNA) using high capacity RNA-to-cDNA kit (Applied biosystem, Cat No: 4387406), following the manufacturer's protocols. The resultant cDNA was diluted with 55µl of RNase-free water (ThermoFisher Scientific, Cat No: AM9906). 3µl of the diluted cDNA was aliquoted in the corresponding wells of MicroAmp® Optical 96-Well Reaction Plates (Applied Biosystems). Then, 12µl of master mix (**see Table 2.4 for the list of master mix reagents**) for JARID2 probe (Hs01004467_m1, ThermoFisher Scientific, Cat No: 4331182) or GAPDH (Hs99999905_m1, ThermoFisher Scientific, Cat No: 4331182) probe was added for each well in 3 technical triplicates reactions in a total volume of 15µl. The QuantStudio 5 Real-Time PCR equipment was used to run the plate after it had been sealed with MicroAmp Optical Adhesive film (Applied Biosystems). The relative expression levels of the JARID2 gene were calculated with the 2^ΔΔCt method using GAPDH as endogenous control and negative siRNA as a control sample.

Component	Volume (µl) per sample	Supplier/Catalogue number
TaqMan master mix	7.5 µl	ThermoFisher Scientific/ Cat No: 4444557
TaqMan probe	0.75 µl	ThermoFisher Scientific, Cat No: 4331182
RNase free water	3.75 µl	ThermoFisher Scientific, Cat No: AM9906
Total	12 µl	

Table 2-4: TaqMan reaction mix for JARID2 and β-actin probe
Table include the component of the TaqMan reaction mix and the volume (µl) of each component for each sample.

2.2.4.4 Plasmids

JARID2 plasmid (pCR8, Addgene, Cat No: 114443) was received as transformed bacteria in a stab culture format. The stabs culture was scraped with a sterile inoculating needle followed by an immediate streaking of the bacteria onto an LB agar plate containing 100ug/ml spectinomycin. The plate was incubated overnight at 37^{o}C to allow for bacterial growth. Then, a single colony was selected and transferred to a new 5 ml of LB broth using a pipette tip and was left in the tube for a few seconds and then incubated in an orbital shaker (Excella E25, Eppendorf, Hamburg, Germany) at 37^{O}C at 250rpm overnight, to ensure adequate aeration for the culture. Then, DNA was extracted and purified from the bacterial culture using a mini Prep kit (Qiagen, Cat No: 27104) following the manufacturer's protocol as described below.

2.2.4.5 Plasmid DNA purification – Plasmid DNA mini preps

DNA was extracted from bacterial cultures by centrifuging 1 ml of the bacterial culture at 6800 x g for 3 min at room temperature. The pellet was resuspended in 250µl of resuspension buffer (P1) followed by the addition of 250µl of alkaline lysis buffer (P2). The reaction was mixed by inverting the tube 4-6 times until the solution become clear and then 350µl of neutralizing buffer (N3) was added. The mixture was mixed immediately by inverting the tube 4-6 times and then centrifuged for 10 min at 17,900 x g. 800µl of the supernatant was added to a QIAprep spin column and centrifuged for 60s at 17,900 x g to bind DNA to the column. The QIAprep spin column was first washed with 0.5ml of buffer PB and centrifuged for 60 s. The second wash was performed with 0.75ml of PE buffer and centrifuged again for 60 s. The spin column was placed in a clean 1.5 ml microcentrifuge tube and the DNA was eluted from the spin column by adding 30µl of EB buffer and left to stand for 1 min followed by a centrifugation for 1 min at 17,900 x g. Prior to cloning, the DNA concentration was measured using a Nanodrop-1000 Instrument (Thermo Fisher Scientific).

2.2.4.6 Gateway cloning

Gateway cloning technology enables to effectively and rapidly transfer DNA fragments between cloning vectors for gene functional analysis and protein expression. It is based on particular

recombination sites known as *att*B site in *E.coli and att*P site in bacteriophage lambda (λ). The cloning reaction is mediated by two proprietary enzyme mixtures termed LR clonase and BP clonase, which produce two distinct forms of recombination reactions known as LR reaction and BP reaction, respectively and they are summarized in **Figure 2.2** (161).

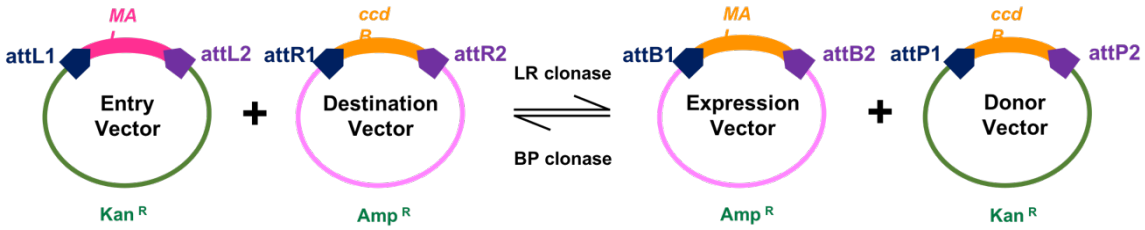


Figure 2-2: Schematic diagram of gateway cloning technology. Gateway cloning technology is an effective and reversible technique for transferring DNA fragments across plasmids. The site-specific recombination between the "att" sites is used in this procedure. The LR reaction, as described above, recognizes the attL and attR sites between an entry vector and a destination vector, respectively, and creates the desired expression vector. An entry clone of interest is produced by the reversible BP reaction, which takes place between the attB and attP sites of an expression vector and a donor vector, respectively.

In the present study, LR Cloning Gateway (Thermo Fisher Scientific, Cat No: 11791020) was used following the manufacturer's protocol. In brief, a reaction mix was prepared by adding 150ng of Gateway Entry (pENTR) vector, 150ng of Destination vector (pDEST) and 4µl TE buffer (PH 8.0). We used JARID2 (pCR8) as the entry vector and the GW306 N-term expressing the 3x flag tag (a kind gift from Professor Colin A. Johnson) as a destination vector. The LR Clonase II enzyme mix was thawed on ice for about 2 min and vortexed twice for 2 s before usage. Then, 1µl of the enzyme mix was added to the reaction mix, vortexed twice and incubated at 25^{o}C for 1 hr. After incubation, 1µl (2µg) of proteinase K solution (Invitrogen, Cat No: 10665795) was added to the above reaction mix and incubated at 37^{O}C for 10 min to stop the reaction. The resulted expression plasmids (i.e.3xFlag-tagged JARID2) were transformed into *E. coli* DH5-Alpha Competent cells (New England Biolabs, Cat No: C2987I) (**see Section 2.2.4.7**) and the DNA was extracted using mini preps from bacterial cultures (**see Section 2.2.4.5**). Sequence verification of the plasmid was performed using Sanger sequencing with the primers listed in **Table 2.5** (**See Section 2.2.4.9**).

2.2.4.7 Transformation of competent bacteria and culture preparation

An aliquot of *E. coli* DH5-Alpha competent cells was gently thawed and kept on ice. 2µl of the resulting plasmid DNA was added to 25µl of the competent cells, pipetted gently up and down and incubated on ice for 30 min. During the incubation time, a water bath was prepared at 42^{O}C and a SOC medium (New England Biolabs, Cat No: B9020S) was put in it. After the 30 min incubation, a heat shock was performed on the transformed cells at 42^{O}C for 45s and transferred immediately on ice for 2 min. Then, 200µl of pre-warmed SOC media was added to each transformation mix and incubated in an orbital shaker for 1 hr at 37^{O}C and at 250rpm (Excella E25, Eppendorf, Hamburg, Germany). Next, 50µl of the transformed cells were plated on an LB agar plate with the appropriate antibiotic that selects plasmids which contain an antibiotic resistance gene. The plate was incubated upside down to avoid condensation of agar gel for 16 hrs at 37^{O}C. The bacterial cultures were then prepared as described in **Section 2.2.4.4**.

Primer design was carried out using web-based software found at https://www.sigmaaldrich.com/webapp/wcs/stores/servlet/LogonForm?storeId=11001. It enables the researchers to design intronic primers for exonic PCR amplification. The following criteria were considered when designing the primers: an optimum annealing temperature of 50-80^{o}C, a minimum of 20pb generating PCR products of 200-600bp in length and a guanine-cytosine content (GC) between 30-70% to ensure stable binding between the template and the primer. In addition, the primers were designed with the absence of secondary structure which is defined as the base pairing interactions within a single nucleic acid polymer or two polymers as the presence of them can result in poor or no yield of PCR product. Primer sequences were checked using the BLAST tool (http://blast.ncbi.nlm.nih.gov/Blast) to ensure that they are uniquely and specifically bound to the gene of interest. Primer sequences used for this purpose are listed in **Table 2.5**.

The primers were provided by Sigma-Aldrich and they were diluted using dH2O to a stock solution with a final concentration of 100µM which is then diluted further to reach a working concentration of 2µM.

Gene name	Primer sequence	Tm^O
JARID2 CMV forward	CGCAAATGGGCGGTAGGCGTG	76.9
JARID2 internal forward	GCCTAAGACAGAAGATTTTCTTA	57.5
JARID2 internal forward	GCAAACAGGTGCTATCCCTC	63.5
JARID2 BSPEI side	GCGAGGAATATCATGAGCATGT	65.0
JARID2 EcoNI side	GCCCGAGTGCAAGCTCAACGAT	73.5
JARID2 NotI side	GCCATTCTCCATGGAGAAGTTA	64.0
JARID2 hGH poly (A) signal side	TTAGGACCAGGATCAGAACG	60.9
JARID2 SV40 promoter	GTGAAGAAGGAAGTGCCGGA	66.5
JARID2 NeoR/KanR side	GGGAGCAGGCTTCAGCTAAC	65.2
JARID2 internal forward	ATGGAGAAGGAGATCCTGGA	62.6
JARID2 internal forward	TTCCATACATTGACTACTTA	49.4

Table 2-5: List of primers used for cloning verification.
Table includes gene name, nucleotide sequence of each primer and the primer melting temperature (Tm^O)

2.2.4.8 Sequence verification of 3xFlag-tagged JARID2 plasmid via Sanger DNA sequencing

Verification of the resulted expression plasmid (3xFlag-tagged JARID2) was achieved using BigDyeÒ Terminator v3.1 Sequencing Kit (ThermoFisher Scientific, Cat No: 4337455) after purification. The sequencing reaction was carried out in a 96-well sequencing plate in a total volume of 10µl and consisted of 1µl Big Dye® 39Terminator v3.1, 1.5µl Big Dye® Sequencing Buffer (5x), 1µl of either the 5' or the 3' primer at 2µM, 5.5µl of dH_2O and 1µl of purified plasmid. The plate was placed in a Veriti Thermal Cycler (Thermo Fisher Scientific) and the mixtures were subjected to the following incubations: denaturation for 1 min at 96°C, followed by 45 cycles of 96°C for 10s, 50°C for 5s, 60°C for 4 min, and then hold at 4°C until ethanol precipitation. The precipitation was achieved by adding 5µl of 125mM EDTA (pH8.0) and 60µl of 100% ethanol to each mixture followed by centrifugation for 30 min at 3100 x g and 22°C. The plate was then subjected to an inverted spin for 15s at 18 x g followed by the addition of freshly prepared ethanol (70%). Next, the contents were centrifuged again at 800 x g for 15 min at 4°C and subjected to an inverted spin for 15s at 18 x g. The precipitated pellets were then left to air dry, face up and protected from light for 15 min at room temperature. The resulted dry pellets were resuspended in10µl of highly deionized formamide "Hi-Di" (Thermo Fisher Scientific, Cat No: 4311320) to hydrate each DNA sample. The plate was put on the ABI 3130xl Genetic Analyzer using standard protocols and a POP7 polymer (Applied Biosystems). Sequencing data was visualized and analysed using 4Peaks (Mek&Tosj.com).

2.2.4.9 Transient transfection for gene over-expression

Prior to transfection, HEK293T cells were sub-cultured in a 6-well plate and allowed to reach 60-70% confluency as described in **Section 2.2.3.1**. On the day of transfection, transfection complexes were prepared by mixing 250µl of Opti-MEM reduced serum media (Gibco, Cat No: 11564506) with 6µl of lipofectamine 2000 (ThermoFisher Scientific, Cat No: 11668030) in a fresh Eppendorf tube. The mixture was mixed gently by flicking the tube and then incubated for 5 min at room temperature. Then, 1µg of plasmid DNA was added to the complex, vortexed, spun down and incubated for 20 min to facilitate the encapsulation of the plasmid in the lipid bilayer of Lipofectamine 2000. Before the transfection, the normal culturing media (i.e. DMEM) was replaced with Opti-MEM followed by the addition of the transfection complexes to each well. The cells were mixed gently with the complexes by rocking the plate back and forth and incubated for 5 hrs. After that, the media was changed back to DMEM to avoid loss of transfection activity. Lysates were then prepared after 48 hrs-96 hrs for protein analysis as described in **Section 2.2.5.1**.

2.2.5 Methods of protein analysis

2.2.5.1 Preparation of cell extracts and determination of protein concentrations

Prior to protein analysis, cells were collected and subjected to cell lysis. Cells in each well of 6-well plates were washed twice with 500 µl of ice-cold PBS and incubated with 350 µl of ice-cold PBS for 5 min on ice. The cells were then detached using cell scrapers and transferred into 1.5 ml tubes. Next, cells were centrifuged at 14000 x g for 5 min and the supernatant was removed followed by the resuspension of the pellet in 80 µl of radio-immunoprecipitation assay lysis buffer (RIPA lysis buffer, **see Section 2.1.1.1 for the buffer recipe**) containing proteinase inhibitor cocktail (Sigma-Aldrich, Cat No: P8340). The suspension was incubated on ice for 30 min followed by centrifugation at 14000 x g for 10 at 4^{o} C. Finally, the supernatant was removed carefully and transferred into a new 1.5 ml tube. The protein in each sample was determined and quantified using the Pierce ^{Tm} BCA protein assay kit (ThermoFisher Scientific, Cat No: 23225) as per manufacturer's instructions. The concentration of the protein was measured on spectrophotometer at a wavelength of 595nm and compared to a range of diluted albumin (BSA) standards. The BSA standards ranges from 0 to 2000 µg/ml.

2.2.5.2 SDS-PAGE and western blotting

For immunoblotting, equal amount of protein (20µg) per sample was mixed with equal volume of 2x laemmli sample loading buffer, heated for 5 min at 95^{o}C on a heat block and were kept on ice until needed. Samples were electrophoresed in 4-15% mini-PROTEAN TGX precast gels (Bio-Rad, Cat No: 4561083). Precision Plus Protein^{TM} All Blue Protein Standards (Bio-Rad, Cat No: 1610373) was used as a protein standard (i.e. a ladder). The gel was run in 1x pre-mixed running buffer (Tris/Glycine/SDS, Bio-Rad Cat No: 1610732) at 120V for approximately 1.5 h. Proteins were transferred onto Trans-Blot Turbo mini nitrocellulose transfer (0.2µm, Bio-Rad, Cat No: 1704158) using Trans-Blot Turbo transfer system (Bio-Rad, 17001915). The transfer was performed for 7 min at 25V.

For the purpose of optimization, the experiment was repeated using different protocol. After heating, samples were electrophoresed 4-12% NuPAGE Bis-Tris gel (ThermoFisher Scientific, Cat No: NP0326BOX) along with Plus Protein^{TM} All Blue Protein Standards. The gel was placed in a transfer tank filled with 1x NuPAGE MES-SDS running buffer (ThermoFisher Scientific, Cat No: NP0002) and ice. It was run for 1.5 h at 120V. Prior to protein transfer, polyvinylidene difluoride (PVDF, 0.2µm, ThermoFisher Scientific, Cat No: LC2002) was activated by ethanol. Proteins were then transferred onto PVDF using NuPAGE transfer buffer (ThermoFisher Scientific, Cat No: NP00061) for an hour at 30V.

2.2.5.3 Antibody binding and visualization of the targeted protein

After protein transfer, the membrane was blocked in 20 ml blocking buffer (5% Marvel milk solution in PBS solution containing 0.1% tween 20 (ThermoFisher Scientific, Cat No: 85115) and incubated for 1 h and at room temperature. For Flag antibody, the membrane was blocked in 3% bovine serum albumin solution (BSA, Sigma Aldrich, Cat No: A4503) in 1x PBST for 30 min at room temperature. After incubation, the membrane was probed with primary anti-JARID2 antibody diluted in 5% Milk solution at a concentration of 1:1000 and incubated overnight on the shaker in the cold room (4^{o}C). For Flag antibody, the membrane was probed with primary anti-flag antibody diluted in 3% BSA at a concentration of 1:1500. Next, the membrane was washed three times with 1x PBST buffer at intervals of 10 min and incubated with the secondary anti-rabbit IgG-HRP-linked antibody diluted in 5% Milk for JARID2 antibodies or 3% BSA for Flag antibody at a concentration of 1:5000 for 1h at room temperature. After that, the membrane was washed again as described above and then SuperSignal West Femto substrate solution (ThermoFisher scientific, Cat No: 34095) was added as

recommended by the manufacturer to develop the membrane. The image of the membrane was obtained using ChemiDoc™ MP Imaging System with a UV transilluminator. To immunoblot the membrane with different primary antibody, it was stripped with 5ml of Restore™ Plus stripping buffer (ThermoFisher Scientific) for 10 min, blocked with 20 ml blocking buffer and then washed three times. The same membrane was later re-probed with primary GAPDH antibody or beta-actin as loading controls at a concentration of 1:10000. The membrane was processed as described above. The analysis was performed on Image Lab (v. 6.0.1) software (Bio-Rad) and the intensity of all bands was quantified and compared to GAPDH or beta-actin depending on which loading control used in the experiment.

2.2.5.4 Co-immunoprecipitation (Co-IP) using protein A agarose

Co-immunoprecipitation is a traditional assay to study protein-protein interactions by overexpressing a whole cell lysate for a specific protein and "pull down" any other protein that are associated with it (162).

Prior to co-immunoprecipitation (Co-IP), all buffers such lysis buffer, incubation buffer, and wash buffer were prepared and kept on ice (**See Sections 2.1.2.2, 2.1.2.5 and 2.1.2.6 for buffer recipe**). The overall experiment was divided into 5 steps. The first step consisted of lysates preparation as described in **Section 2.2.5.1**. In the second step, the collected lysates were homogenized with 20G needle and the suspension was gently rocked on an orbital shaker (Excella E25, Eppendorf, Hamburg, Germany) for 15 min at 4^{o}C to lyse the cells. The suspension was then centrifuged in pre-cooled centrifuge at 14,000x g for 15 min and the supernatant was transferred immediately to a fresh clean Eppendorf tube. The protein concentration was determined as described in **Section 2.2.5.1** and 500µg of protein per pull down was transferred to a new tube and filled up to a volume of 1000µl with the incubation buffer. In the third step, protein A agarose slurry was prepared. 200µl of protein A/G agarose beads (Roche diagnostics) were washed twice with PBS with and centrifuged for 1 min at 1000x g. In the second wash, 50% of the bead slurry was restored with PBS. In the fourth step, the lysates were pre-cleared with the bead slurry to minimizes non-specific binding of protein to the agarose. The pre-clearing was achieved by the addition of 80µl of the bead slurry to the lysates and incubated for 30 min at 4^{o}C on an orbital shaker. The lysates were then centrifuged at 1000x g at 4^{o}C for 1 min to remove the protein A agarose and the supernatant was transferred to a fresh Eppendorf tube. Next, the supernatant was coupled with target protein-specific antibodies

(1µg of antibody for each 500µg of protein) for 2 hrs at 4°C on an orbital shaker. The immunocomplex was then captured by the addition of 80µl of protein A agarose slurry and the mixture was incubated on an orbital shaker overnight at 4°C. After incubation, the mixture was centrifuged for 1 min at 1000x g and the supernatant was discarded to collect the pre-coupled agarose beads. Next, the samples were washed quickly three times with 500µl of ice-cold wash buffer. In the fifth step, the pre-coupled agarose beads were resuspended in 20µl of 2x SDS, mixed by inverting and incubated for 30 min at room temperature to dissociates and elutes the immunocomplex from the agarose. Finally, the mixture was spun for 2 min at 14,000x g and the supernatant was transferred to a new Eppendorf tube. The sample was processed for western blotting as described in **section 2.2.5.2 and section 2.2.5.3** to identify potential protein-protein interaction. The Co-IP protocol was summarized in **Figure 2.3**.

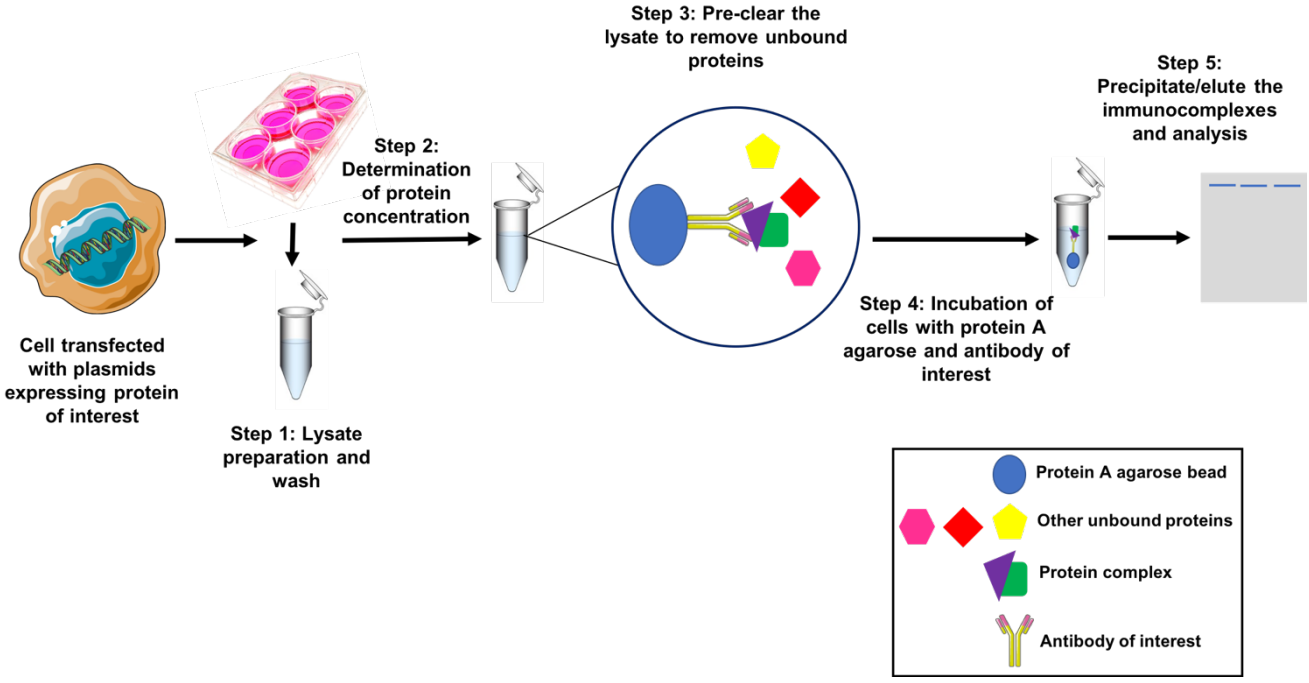


Figure 2-3: Schematic diagram of co-immunoprecipitation procedure and principles. Lysates from transfected cells were prepared and the protein concentration was determined. 500µg of protein was used for each IP. The immunocomplex was captured by the addition of protein A agarose slurry. Agarose beads which contain the protein of interest was dissociated and eluted from the pre-coupled agarose beads and the supernatant was processed via western blot analysis.

2.2.6 Chromatin Immunoprecipitation sequencing (Chip-seq)

2.2.6.1 Tissue sectioning

Formalin-fixed, paraffin-embedded patient samples were acquired from the Brain Tumour Northwest Research Tissue Bank at Royal Hospital Preston. They were used in this study following project specific favourable ethical opinion from the National Research Ethics Service Committee South Central - Oxford A (Research Ethics Committee code 13/SC/0509). The samples were specifically selected as paired primary and locally recurrent IDH wildtype GBM brain tumours from a patient that had received standard treatment

The block was sectioned according to instructions given by Active Motif company which performed 2 Chip-seq experiments on this sample. The tissue block was kept into the block holder of the microtome which is a cutting tool that is used to produce thin slices of the tissues known as sections. For the purpose of this work, 20 sections of 10µm thick was provided for each Chip reaction. Total of 5 1.5ml tubes each with 20 curls were prepared for H3K4me3, H3K27me3, EZH2, JARID2 and input reactions.

2.2.6.2 Library preparation and sequencing

Tissue sections were sent to Active Motif (Carlsbad, CA) for ChIP-Seq analysis. Active Motif performed the chromatin preparation, ChIP protocols, library preparation, and library sequencing. Cells were dampened with 0.125 M glycine and fixed with 1% formaldehyde for 15 min. The fixed cells were mixed with lysis buffer and then agitated in a Dounce homogenizer. With the use of Active Motif's EpiShear probe sonicator (cat# 53051), the resulting lysates were sonicated, and the DNA was sheared to an average length of 300–500bp.

To prepare the input sample, fractions of chromatin were treated with RNase and proteinase K. the mixture were heated to break down crosslinks. This was followed by SPRI beads clean-up (Beckman Coulter), and Clariostar quantification (BMG Labtech). With the use of protein G agarose beads, an aliquot of chromatin (50 ug) was precleared (Invitrogen). 4 ug of an anti-JARID2 (Novus biology), anti-H3K4me3, anti-H3K27me3 and anti-EZH2 was used to identify genomic DNA areas of interest. Complexes were cleaned, then treated with RNase and proteinase K after being eluted from the beads using SDS buffer. Crosslinks were broken down overnight at 65 ^{o}C, and ChIP DNA was then extracted using phenol-chloroform, followed by ethanol precipitation.

SYBR Green Supermix was used in triplicate for quantitative PCR (qPCR) experiments on particular genomic regions (Bio-Rad). By running qPCR for each primer pair using input DNA, the signals were adjusted for primer efficacy.

Using the ChIP and input DNAs, Illumina sequencing libraries were prepared using the conventional enzymatic procedures of end-polishing, dA-addition, and adaptor ligation. A robotic system (Apollo 342, Wafergen Biosystems/Takara) was used to carry out the steps. The resultant DNA libraries were measured and sequenced using Illumina's NextSeq 500 following a final PCR amplification step (75nt reads, single end).

2.2.7 Cleavage under targets and release using nuclease (Cut&Run)

2.2.7.1 Disaggregation of tissues into a single cell suspension

Prior to performing CUT&RUN, cells were dissociated from 2 fresh frozen patient tumours namely, NB17/39 and NB169/12. All steps were processed on ice unless otherwise stated. Tissues were placed onto clean and pre-cooled 6cm petri dish followed by the addition of 400µl accutase (Sigma-Aldrich, Cat No: A6964) drop by drop on one side and some Neurobasal-A serum free-media (NB) on the other side of the dish. Using 2 sterile scalpels, the tissue was chopped quickly until it became creamy consistency, transferred into 50ml sterile centrifuge tube using glass pasteur pipette and resuspended directly with 40ml cold PBS. The solution was then spun at 200 x g for 5 min at 4^{o}C and the supernatant was discarded. Finally, the pellet was resuspended in fresh NB medium and kept it ready for CUT&RUN assay.

2.2.7.2 CUT&RUN workflow

Tissues and cell cultures were harvested as described in **Sections 2.2.3.1 and 2.2.7.1** respectively before use. The cells were counted using haemocytometer and trypan blue to get 1 million cells for each antibody/MNase reaction. An additional 1 million cells were obtained for input sample (i.e. no antibody). Cells were pelleted by centrifugation for 3 min at 600 x g at room temperature.

CUT&RUN was carried out following the same principles and instructions as per manufacturer's protocols (Cell Signaling Technology, Cat No: 86652). The first step consisted of cell preparation and binding of primary antibody. During this step, the cell pellet was first resuspended in 1 ml of 1x CUT&RUN wash buffer (Cell Signaling Technology, Cat No: 31415), centrifuged for 3 min at 600 x g followed by the removal of the supernatant. The pellet was washed a second time. Next, 100µl of 1x wash buffer was added to each reaction followed by the addition of 10µl of activated bead

suspension. Concanavalin A magnetic beads (Cell Signaling Technology, Cat No: 82307) were activated before adding to the cell suspension by adding 10µl of beads slurry per reaction with 100µl concanavalin A bead activation buffer (Cell Signaling Technology, Cat No: 91275) and the tube was mixed gently and placed on a magnetic stand until the solution become clear. The supernatant was discarded and the beads was washed again by the activation buffer as described above. 10µl of the activation buffer was added to the pellet and resuspended gently. Cell suspension with the activated concanavalin A bead slurry was rotated for 5 min at room temperature. After that, the suspension was briefly centrifuged at 100 x g to remove cell-bead suspension from the cap of the tube and the tube was placed on the magnetic stand until the solution became clear. The supernatant was discarded and 100µl of antibody binding buffer (Cell Signaling Technology, Cat No: 15338) was added to the cell pellet (i.e. cell-bead suspension). Then, 100µl was aliquoted separately for each antibody/MNase reaction followed by the addition of 2µg of the antibody and the mixture was mixed gently. The mixture was then rotated at 4^{o}C for 2 hrs. during the incubation period, the digitonin buffer (Cell Signaling Technology, Cat No: 16359) and pAG-MNase pre-mix solution were prepared for the second step which is the binding of pAG-MNase enzyme (Cell Signaling Technology, Cat No: 57183).

After incubation, the sample was centrifuged briefly at 100 x g to remove cell-bead suspension from the cap of the tube and the tube was placed on the magnetic stand until the solution turns clear. Next, the supernatant was discarded and the cell-beads suspension was resuspended with 1ml of digitonin buffer. The tube was placed again on the magnetic stand until the solution became clear followed by the removal of the supernatant. Then, 50µl of pAG-MNase pre-mix solution was added and mixed gently by pipetting up and down. After that, the sample was rotated at 4^{o}C for 2 hrs. After rotation, the pAG-MNase suspension was briefly centrifuged at 100 x g to remove cell-bead suspension from the cap of the tube and the tube was placed on the magnetic stand until the solution turns clear. At this stage, the antibody should be bound to its epitopes and the pAG-MNase should bound to the constant region of the antibody. The supernatant was discarded and 1ml of digitonin buffer was added to resuspend the beads. The tube was placed on the magnetic stand again until the solution turns clear and the supernatant was discarded followed by the addition of 150µl of digitonin buffer. Next, the tube was placed on ice for 5 min to cool before digestion and then the MNase was activated by the addition of 3µl cold Calcium Chloride (Cell Signaling Technology, Cat No: 55676). The sample was incubated for 30 min at 4^{o}C. After 30 min, the enzyme cleaved the DNA underlying the antibody and the reaction was stopped by the addition of 150µl of stop buffer (Cell Signaling Technology, Cat No: 48105) followed by an incubation for 10 min at 37^{o}C.

This buffer will allow the pAG-MNase to cut the DNA fragment to diffuse into the supernatant. At this stage, the third step of CUT&RUN workflow is started in which the DNA is digested and diffused. Finally, the sample was centrifuged 4°C for 2 min at 16,000 x g, placed on the magnetic stand until the solution turns clear and the supernatant, containing the pAG-MNase-antibody targeted DNA complex was transferred to a new 1.5 ml microcentrifuge tube. This is the enriched chromatin samples and can be stored at -20°C for up to 1 week. The CUT&RUN workflow was summarized in **Figure 2.4**.

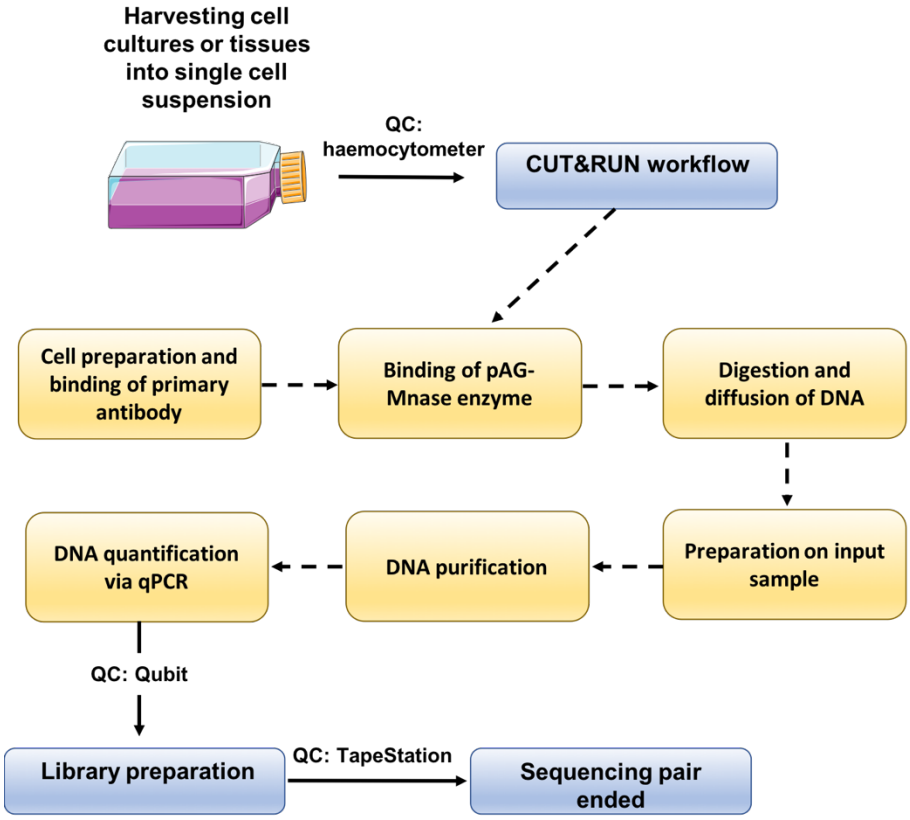


Figure 2-4: A schematic representation of CUT&RUN workflow and sequencing. Prior to performing CUT&RUN, cell culture or tissue were harvested to single cell suspension and counted via haemocytometer and trypan blue to get 1 million cells for each antibody/MNase reaction and input sample. The first step in CUT&RUN workflow was the preparation of the sample by washing the cell suspension twice with 1x CUT&RUN wash buffer, coated with Concanavalin A magnetic beads and immobilized them from on-target chromatin. Then, an antibody to the target of interest was added to the suspension and incubated for 2hrs at 4°C. Next, the enzyme pAG-MNase was added and upon activation of it with $Ca2^{+}$ ion, the enzyme cleaves the desired chromatin fragment. After that, the fragmented DNA/chromatin was diffused out of the cell and purified using DNA spin

column. Then, the input sample was prepared and purified. Finally, the purified DNA was quantified via qPCR and only 5ng of the purified DNA was processed for library preparation and sequencing.

2.2.7.3 Optimization of sonication conditions for input samples

Chromatin fragmentation should be tested and optimized before starting any experiment that required DNA shearing. For a successful CUT&RUN experiment, DNA fragments of 100-600 bp in length is recommended. In this work, the sonication conditions of the input sample were optimized using 4 conditions: 15, 20, 25 and 30 cycles of 30 sec ON/ 30 sec OFF with the Bioruptor UCD-200 sonicator (Digenode). DNA fragment sizes of the sonicated samples were determined using gel electrophoresis (**see Section 2.2.7.5**).

2.2.7.4 Preparation of the input sample

Input sample prepared from step 1 of CUT&RUN workflow (**See section 2.2.7.2**) was processed as follows. 200µl of DNA extraction buffer was added to the harvested cells and the sample was incubated at 55^{o}C with shaking. After 1 hr, the sample was placed on ice immediately. Due to the fact that only fragmented DNA (i.e., 10kb) can be purified using DNA purification spin columns, the cells were lysed and the chromatin was fragmented to a size of 100-600 bp by sonication (**See Section 2.2.7.3**). The sonication conditions were optimized as recommended by the manufacturer and the optimal sonication condition was selected for the remaining work. (**See Section 2.2.7.3 for sonication optimizations**). The sonicated sample was centrifuged at 4^{o}C for 10 min at 18,500 x g and the supernatant was transferred to a new 1.5 ml microcentrifuge tube prior to DNA purification.

2.2.7.5 Agarose gel electrophoresis

15µl of the sonicated sample was loaded in an agarose gel for DNA fragments evaluation. Agarose gel was prepared by dissolving 1g of agarose powder (Biolin, London, UK, Cat No: BIO-41025) in 100 ml of 1x TAE buffer. 6µl of SYBR DNA gel stain (Sigma-Aldrich, Cat No: 59430) was added to the melted agarose solution and poured into the gel cassette. Sample was loaded on the gel immersed in 1x TAE buffer along with a hyperLadder 100 bp Plus (Bioline, Cat No: BIO-33056). The gel was run at 80V for 1-1.5 hrs and the gel was observed on a ChemiDoc XRS gel imaging system (Bio-Rad).

2.2.7.6 DNA purification using spin columns

DNA from enriched and input samples (**resulted from Section 2.2.7.2 & 2.2.7.3, respectively**) was purified using DNA purification buffers and spin columns (ChIP, CUT&RUN) assay kit (Cell Signaling Technology, Cat No: 14209S) as described by the manufacturer's instructions. Prior to DNA purification, 24 ml of ethanol (96-100%) was added to the DNA wash buffer before use. Then, 1.5 ml of DNA binding buffer was added twice to each sample (i.e. enriched chromatin or input sample), vortexed briefly followed by the transfer of 600µl of the mix to a DNA spin column in the collection tube. The sample was centrifuged at 18,500 x g in a microcentrifuge for 30s. 750µl of DNA wash buffer was added to the samples and they were centrifugated twice at 18,500 x g for 30s. Next, sample eDNA was eluted in 50ul following a centrifugation at 18,500 x g for 30s.

2.2.7.7 DNA quantification by quantitative polymerase chain reaction

Initial quantification of DNA was carried out via quantitative polymerase chain reaction (qPCR) in 20µl containing 10µl of SYBR-Green mix (Sigma-Aldrich, Cat No: S4438), 2µl of Human RPL30 exon 3 primer (Cell signaling Technology, Cat No: 7014), Human RPL30 intron 2 primer (Cell signaling Technology, Cat No: 7015) and SAT2 primer (Novus Biology, Cat No: NBP1-71655 at a final concentration of 5µM , 0.2µl of ROX dye (ThermoFisher Scientific, Cat No: 12223012), 5µl of nfH_2O and 2µl of purified DNA. Each sample was tested in triplicates. A three-step cycle programme with different melting temperature were applied and repeated 40 times. The cycling steps were listed in **Table 2.6**.

Step	Temperature	time	Number of cycles
Initial denaturation	95°C	3 min	1 cycle
Denaturation	95°C	15 s	40 cycles
Annealing and extension	60°C	60 s	

Table 2-6: PCR reaction conditions program for CUT&RUN DNA quantification
Table includes PCR step, temperature (°C), time and number of PCR cycles

The analysis was performed manually by calculating the IP efficiency using the percent input method as follows:

Percent Input = 100% x 2(C[T] 100%Input Sample – C[T] IP Sample)

Signals from each immunoprecipitation are quantified as a percentage of the total amount of chromatin used as an input. The enrichment of each sample was compared to that of IgG, which was utilized as a control sample.

2.2.8 Library preparation and sequencing

Prior to library preparation, the purified DNA resulted from DNA purification step (**See Section 2.2.7.6**) was read in Qubit fluorometer using the Qubit HS assay reading. Library preparation was performed using NEBNext Ultra Library Prep kit for Illumina (BioLabs, Cat No: E7370L), with slight adaptations. These changes were recommended from the CUT&RUN protocol. Briefly, 5ng of purified DNA was added to 55.5µl nuclease free water (nfH_2O) followed by the addition of 9.5µl of NEBNext End Prep master mix. The reaction mixture was placed on the Veriti thermal cycle (Thermo Fisher Scientific) with heated lid set to ≥ 75^oC and run using end repair program as follows: 30 min at 20^oC, 30 min at 50^oC and hold at 4^oC. The temperature was reduced from 65^oC to 50^oC to avoid denaturing small DNA fragment. Prior to adapter ligation, NEBNext adaptor was diluted to 0.6uM (25-fold dilution) for sample with less than 5ng and 10-fold dilution for those with 5ng. Then, the adaptor ligation mix was prepared and 16µl was added to each sample followed by the addition of 2.5µl of the diluted NEBNext adaptor. The mixture was subjected to an incubation for 15 min at 20^oC in the thermal cycle with heated lid set to ≥ 40^oC. Then, 3µl of USER enzyme was added to each sample, and the samples were then placed in a thermal cycle for 15 min at 37^oC with a heated lid set to 40^oC.

Adaptor-ligated DNA was cleaned up and purified using AMPure XP beads (Beckman Coulter, Cat No: a63381) with a final concentration of 1.1X instead of 0.9 to increase capture of small fragments. Following purification, the library was amplified in a 12 cycle PCR reaction using NEBNext oligos kit, and purified again with AMPure XP beads. The size and quality of the fragmented DNA was evaluated using high sensitivity HSD1000 TapeStation tapes (Agilent Technologies), according to manufacturer's instructions. Library with adapter dimer was re-cleaned up again at 0.8x twice to remove the adaptor dimers and re-assessed again on HSD1000 TapeStation tapes. Then, the purified

library was washed twice with 80% ethanol, eluted into 26µl of elution buffer (EB, Qiagen, Cat No: 19086) and only 22µl of the elution was recovered. The eluted sample was re-evaluated on HSD1000 TapeStation tapes to ensure that the library is free of contamination. The suitable library along with a negative control were read in Qubit 1X dsDNA HS kit assay following the manufacturer's protocol. Sample was pooled at 5x the concentration and sequenced on the Illumina NextSeq 500 platform (Illumina, San Diego, CA) using a 75bp paired end protocol.

2.2.9 Optimization of computational pipeline

2.2.9.1 ChIP-seq pipeline design and implementation

A broad variety of computational tools have been designed to analyse ChIP-seq data and comprehend the genome-wide mapping of protein-DNA interactions from sequencing read. A general pipeline was developed and implemented on Linux platform (i.e. ARC3). The pipeline functions to converts a set of multiplexed Fastq files into either a bed format that can be used as input for various programs such as ChromHMM or Bedtools, or a bigWig file that can be used to visualize dense and continuous data as a graph in the genome browser tool. **Figure 2.5** outlines the proposed ChIP-seq processing pipeline from quality assessment of raw sequencing reads to promoter status calling step and the list of tools used to develop the pipeline is presented in Appendix A.

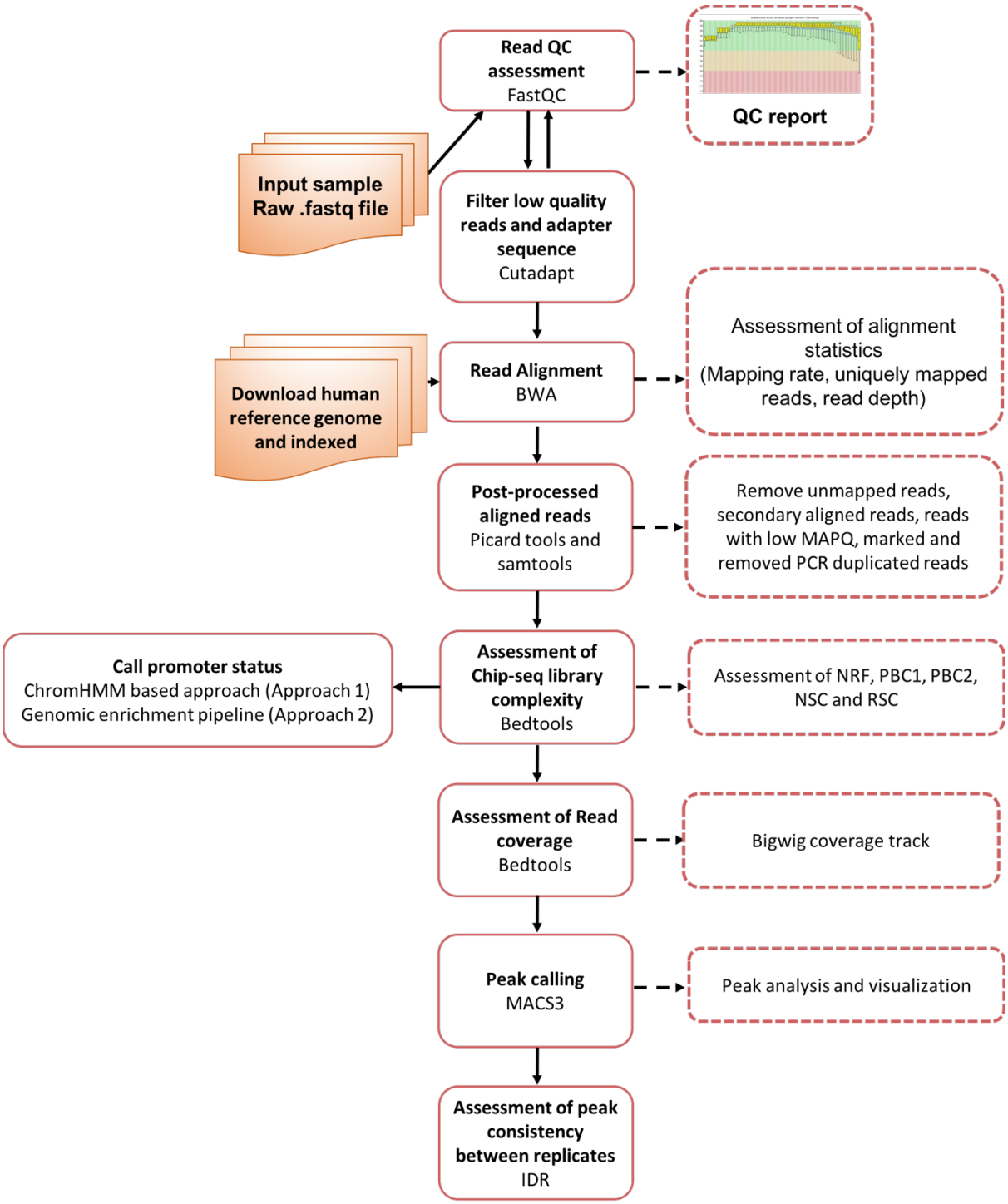


Figure 2-5: A schematic representation of the proposed ChIP-seq pipeline.

The pipeline was first developed using an external dataset. This dataset, published in Liau et al., was obtained from the NCBI Gene Expression Omnibus via accession number GSE74557. Fastq files (single-end reads) were downloaded for two cell lines (GSC8 and GSC8per), which each underwent ChIP-seq to detect the location of both H3K27me3 and H3K4me3 marks, compared to input DNA

controls (Appendix B). GCS8 cells are a patient-derived GBM cell line cultured in serum-free (stem-cell permissive) conditions; GSC8per cells are GSC8 cells that persist following prolonged (>8 week) treatment with dasatinib. The pipeline was implemented and optimized further using an in-house ChIP-seq dataset. DNA from a fresh frozen pair of samples in our lab underwent ChIPseq (performed externally at Active Motif, Inc) to assess EZH2, H3K27me3, H3K4me3 in both samples, and JARID2 in the recurrent sample.

Raw sequencing files created by the *Illumina* NextSeq 500 platform were obtained in "FASTQ" format. The sequencing data usually consists of millions of short reads with the length of approximately 36-75 nucleotides. The first step in the proposed pipeline was to examine the quality of the obtained data using FastQC (v0.11.9). FastQC is a quality analysis tool that is designed to perform a set of quality checks on the raw sequence data and spot any potential issues or biases in a QC report. The obtained reads were run on FastQC and it provided a QC report for each sample. I focused on assessing per sequence quality score, sequencing depth and length, sequence duplication levels and adapter content of each sample as these metrices considered the key quality metrices for ChIP-seq quality assessment. Once the sample passed the quality check, the adapter sequences and more specifically Illumina adapter sequences along with low quality ends of phred < 10 were trimmed using Cutadapt (v3.6). By default, a minimum overlap length of 3 was allowed to reduces the number of falsely trimmed bases and reads that are shorter than 20 bases are discarded to avoid having empty reads (i.e. reads that have a length of zero) in the final output. The quality of the trimmed reads is checked again and only samples that passed the quality check proceed to mapping.

Prior to mapping, the human reference genome (Release 39, GRCh38.p13) was downloaded from gencode https://www.gencodegenes.org/human/releases.html in the FASTA format which contains all the nucleotide sequence of the GRCh38.p13 version on all regions, including reference chromosomes, scaffolds, assembly patches and haplotypes. The reference genome was indexed using bwa (v0.7.17) and the trimmed reads were aligned to it with default parameters to generate sequence alignment map (SAM) file. Samtools (v1.11) was used to convert SAM file was to binary alignment map (BAM) file and then sorted and indexed. The alignment statistics were inspected and only the samples with a mapping rate > 80% was post-processed.

Samples with high mapping rate were post-processed by removing unmapped reads, mate unmapped reads, reads that align to more than one place (i.e. secondary alignment) and reads with

low mapping quality score (i.e. MAPQ < 25) using samtools. In addition, PCR duplicates were marked and removed using picard tool (v2.21.2) with default parameters. A report that summarized the main quality metrices in terms of percentage of duplicate reads, estimated library size and the number of unmapped reads was generated. Library complexity, defined as an estimation of the number of the distinct molecules in the given library, was inspected using Picard tools and Bedtools (v2.30.0). This included the assessment of the prevalence of uniquely mapped reads, which is known as the non-redundant fraction (NRF), PCR bottlenecking coefficient 1 and 2 (PBC1 and PBC2), normalised strand cross-correlation coefficient (NSC) and relative strand cross-correlation coefficient (RSC) (**See Chapter 1, Section 1.6.1.2 for detailed definitions of these terms**). The genomic coverage across the sample was assessed by generating the coverage track (Bigwig) and this was done by converting the resulting BAM file to the Bigwig file using the 'bamCoverage' function of the deeptools (v3.5.1) with default parameters.

To identify the enriched peaks in each sample, MACS3 (v2.2.7.1) was used. Post-processed experiment files, along with sample-matched input control files, were used for the peak calling with default settings, except that I set the '--broad' option to call broad peaks for H3K27me3, JARID2 and EZH2. P-value thresholds were set at 0.01 and 0.1 for narrow and broad peaks respectively. Bed files are the output format from MACS, giving the genomic location of each peak and its intensity. The reproducibility at the level of peak calling between replicates was measured using irreproducible discovery rate (IDR) with a q-value of 0.1 and IDR-threshold of 0.05.

2.2.9.2 Development of a bespoke approach to call the promoter status along with the integration of expression data

For genome-wide profiling of the histone and chromatin-binding proteins, I developed two approaches that enabled me to identify the co-occurrence of chromatin marks across a pre-defined promoter region. Approach 1 was developed via adoption of ChromHMM (**See Section 2.2.8.2.1**) to output promoter calling status so I could characterize the chromatin states in those specific regions in our datasets. ChromHMM tool is used widely to annotate and characterize the chromatin states across multiple experiments. Approach 2 defines and calls the chromatin states by scoring the enrichment of signal in defined regions compared to the background across the same sized windows across all genomic regions (**See Section 2.2.8.2.2**).

For the purpose of this work, I used the definition of the promoter which is +/- 1kb around the transcription start site (TSS). I created the promoter file from the gencode annotation (Version 27, GRCh38.p10) file which includes detailed gene annotation on the reference chromosomes, scaffolds, assembly patches and alternate loci (i.e. haplotypes). A simple script was developed in an integrated development environment (IDE) for Java known as NetBeans (v8.2) and the annotation file in a GTF format along with the chromosome size file in a text format were used as inputs. The latter was generated from the index file of the human reference genome (Release 27, GRCh38.p10) using samtools and it contains the chromosome name and it corresponding chromosome size in a tab delimited file. The script took these two files as inputs, searched for the gene transcript and extracted the chromosome number, chromosome start position whether it is located in the positive strand or negative strand, the gene id and the transcript id. The script then used the start position and expanded the region around it by 1000bp either side. It merged all the transcript ids that have the same gene id and start position (TSS) in one line. The final output was presented in a tab delimited file and contains the chromosome number, the promoter region as start and end position, the gene id and the transcript ids.

2.2.9.2.1 Chromatin state discovery and development of promoter calling approach using ChromHMM

Before using ChromHMM (v1.23), the post-processed bam file was converted to a bed file using the *Bamtobed* function of the Bedtools. The bed file was then binarized with the commands "BinarizeBed -b 200" of the ChromHMM. Two files are required to binarize the data, the chromosome size (**See Section 2.2.8.2**) and the cell mark file table. The latter is a tab delimited file in which each row contains the name of the bed file and the corresponding control bed file for each cell type and mark. **Figure 2.6** shows the selected options of cell mark file table for handling multiple cell types for ChromHMM.

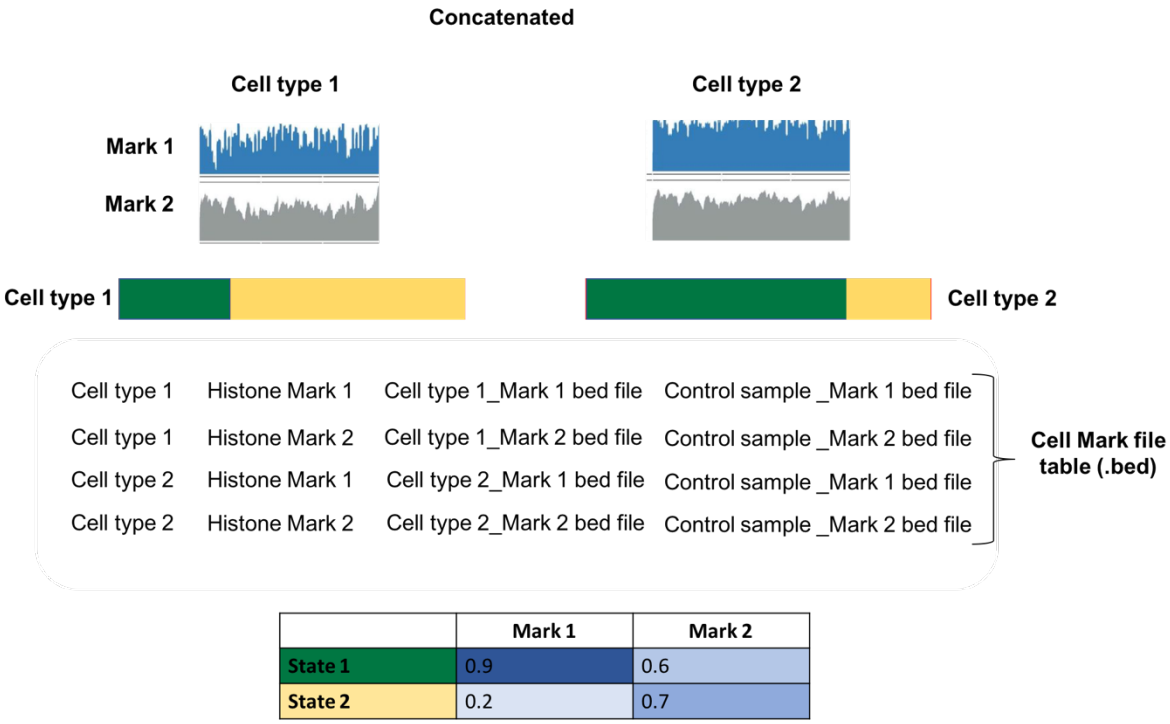
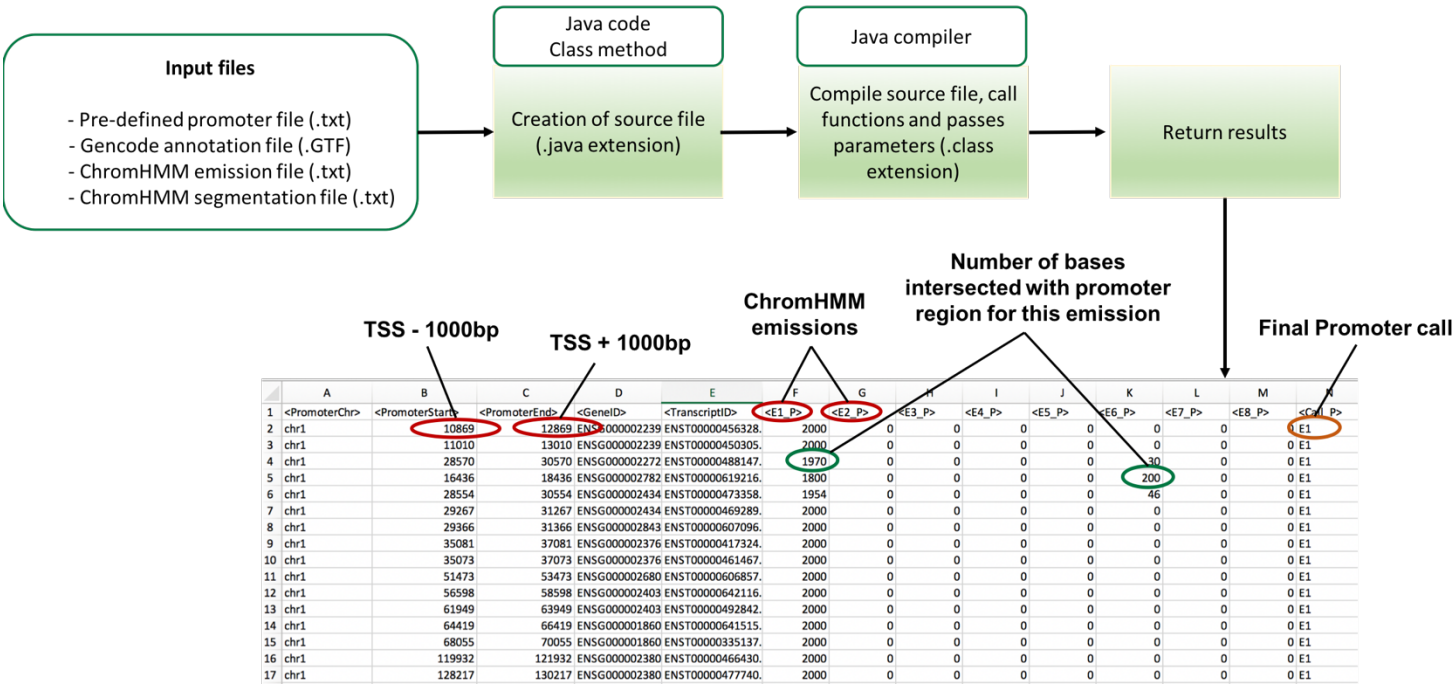


Figure 2-6: A diagram showed the selected options of cell mark file table for handling multiple cell types for ChromHMM. Cell mark file table is designed by effectively concatenating multiple cell types to adjust the binarization threshold locally during the binarization step of ChromHMM. This results in one common and shared model.

Next, a newly trained model was developed using the *learnModel* function of ChromHMM with default parameters but, depending on the analysed data, the number of states to be modeled can be decided by the user. In our case, models were trained to segment the genome into 4 chromatin states for the external dataset (in which two histone modifications were experimentally assessed) and 8 chromatin states for the in-house dataset (in which two histone modifications and a transcription factor were experimentally assessed). The states were then labelled, based on the probability of observing the histone marks in each state, as active (H3K4me3 only), repressive (H3K27me3 only) or bivalent (both marks present). In general, the resulting files (more specifically the emission file (.txt) and the segments file (.bed)) were used for downstream analysis.

To call the promoter status in our regions of interest, a java program was developed in NetBeans. The program took the segment and emission files resulting from ChromHMM, and the pre-defined promoter file and gene expression file as inputs. The program was designed first to search for an intersection between each segment and the promoter region. Then, the number of intersected

bases was calculated and reported in the corresponding emission column (**Figure 2.7**). The resulting tab delimited file was merged with the gene expression file based on the TSS id. These data were then analysed to get the distribution of the chromatin states across the promoter regions.



	A	B	C	D	E	F	G	H	I	J	K	L	M	N
1	<PromoterChr>	<PromoterStart>	<PromoterEnd>	<GeneID>	<TranscriptID>	<E1_P>	<E2_P>	<E3_P>	<E4_P>	<E5_P>	<E6_P>	<E7_P>	<E8_P>	<Call_P>
2	chr1	10869	12869	ENSG000002239	ENST00000456328.	2000	0	0	0	0	0	0	0	E1
3	chr1	11010	13010	ENSG000002239	ENST00000450305.	2000	0	0	0	0	0	0	0	E1
4	chr1	28570	30570	ENSG000002272	ENST00000488147.	1970	0	0	0	0	30	0	0	E1
5	chr1	16436	18436	ENSG000002782	ENST00000619216.	1800	0	0	0	0	200	0	0	E1
6	chr1	28554	30554	ENSG000002434	ENST00000473358.	1954	0	0	0	0	46	0	0	E1
7	chr1	29267	31267	ENSG000002434	ENST00000469289.	2000	0	0	0	0	0	0	0	E1
8	chr1	29366	31366	ENSG000002843	ENST00000607096.	2000	0	0	0	0	0	0	0	E1
9	chr1	35081	37081	ENSG000002376	ENST00000417324.	2000	0	0	0	0	0	0	0	E1
10	chr1	35073	37073	ENSG000002376	ENST00000461467.	2000	0	0	0	0	0	0	0	E1
11	chr1	51473	53473	ENSG000002680	ENST00000606857.	2000	0	0	0	0	0	0	0	E1
12	chr1	56598	58598	ENSG000002403	ENST00000642116.	2000	0	0	0	0	0	0	0	E1
13	chr1	61949	63949	ENSG000002403	ENST00000492842.	2000	0	0	0	0	0	0	0	E1
14	chr1	64419	66419	ENSG000001860	ENST00000641515.	2000	0	0	0	0	0	0	0	E1
15	chr1	68055	70055	ENSG000001860	ENST00000335137.	2000	0	0	0	0	0	0	0	E1
16	chr1	119932	121932	ENSG000002380	ENST00000466430.	2000	0	0	0	0	0	0	0	E1
17	chr1	128217	130217	ENSG000002380	ENST00000477740.	2000	0	0	0	0	0	0	0	E1

Figure 2-7: A diagram showing the steps of generating the java program in NetBeans to call the promoter status using the ChromHMM based approach. A java program was developed in NetBeans to generate the promoter status calling file. The programme took the pre-defined promoter file, gencode annotation file, ChromHMM emission file and segment file as inputs and processed them to generate the promoter calling file. The program started by intersecting each segment with the promoter region, assign each intersection with its corresponding emission and then calculate the sum of intersected bases for each emission. The final call for each promoter was assigned based on the emission that harbour the highest number of intersected bases.

2.2.9.2.2 Development of an alternative pipeline to characterize enriched genomic region

This approach was developed into a programme called GBMProm in partnership with AD Bioinformatics, in parallel to me coding the ChromHMM-based approach as above. The main idea behind the approach is to identify the enrichment of reads binding in a given genomic region and score the number of reads in that region in comparison to a suitable background (See Appendix C

for the full R code of GBPprom approach). The use of the background is to ensure that the changes in read depth are not related to the genomic copy number and GC content.

Prior to generating the read count, reads that align to notoriously difficult-to-accurately-mapregions of the genome (i.e. blacklisted regions) were filtered out. The average read count pergenomic region/window (λ) of fixed size (w, which equates to 2-kb for my definition of a promoter i.e. 1kb either side of a TSS) was calculated (**See Appendix C.1**) by counting the number of aligned readsfor the ChIP experiment (nr_{ChIP}) and control/background (nr_{input}), which is non-immunoprecipitated input DNA, and inputting that to the following equation:

$$\lambda_{\text{ChIP}} = \frac{W \times nr_{ChIP}}{genomeSize} \qquad \lambda_{\text{input}} = \frac{W \times nr_{input}}{genomeSize}$$

Where λ_{ChIP} and λ_{input} are the average read count per genomic region/window for ChIP experiment and input (i.e. control) sample respectively, W is 2-kb sliding window, nr_{ChIP} is the number of aligned reads for ChIP experiment, nr_{input} is the number of aligned reads for control/background and the genomeSize is the size of fixed intervals of the ChIPand control samples.

Reads mapping in promoter regions were also counted for both the ChIP experiment and theinput control. The probability of read enrichment for each promoter regions was counted and the scaling value (ε) was calculated as follows:

$$\varepsilon_{input} = \frac{nr\_promoter_{input}}{\lambda_{input}}$$

Where ε is the scaling value (i.e. the deviation in the promoter read count from what is expected in the input experiment), nr_{input} is the number of aligned reads for control/background in the promoter region, W is 2-kb sliding window and λ_{input} is the average read count per input (i.e. control) sample for the promoter region.

Then, based on the assumption that random read alignment would follow a Poisson distribution, A, with parameter $\lambda_{(ChIPpromoter)}$, the significance of the signal per promoter in each window was evaluated using a nominal p-value which is defined as P ($A \leq a$) in which:

$$\text{If } (\varepsilon_{input} > 1)\{ A \sim Poisson(\varepsilon_{input}\ x\ \lambda_{ChIP})\}$$
$$\text{Else } \{ A \sim Poisson(\lambda_{ChIP})\}$$

In statistics, Poisson distribution defines as a discrete frequency distribution that provides the likelihood of the number of independent events occurring within a certain time. Once all promoter *p*-values have been calculated, multiple testing correction was performed using the Benjamini-Hochberg (BH) procedure with a defaultFalse Discovery Rate (FDR) value < 0.05. BH is a procedure used in testing multiple hypothesis by controlling FDR, which is defined as the proportion of false positive test results to the total number of positive test results. FDR was used here to elucidate the significance of promoter signal in which promoters that had a corrected P value < 1 x 10^{-5} after correction are considered significant. The above steps were performed using calculate promoter signal script (**See appendix C.2**).

The FDR-adjusted p-values were used to score promoters as enriched or not and this was initially set to a default of 1 x 10^{-5} because this was selected within the publication I based this approach on (122- 125). The program gives a score of 0 if the adjusted p-value for each of the histone mark is higher than the selected threshold or 1 if the adjusted p-value is lower than the threshold atthe promoter region. By combining scores, all possible chromatin states were defined in a certain order (**See appendix C.3 for the full R script of scoring the enrichment of each promoter region**). **Figure 2-8** shows a workflow diagram of this alternative pipeline to characterizeenriched genomic region. To simplify the analysis of the chromatin states, I characterized these combined binary annotations to label promoters (i.e. active, repressive, etc).

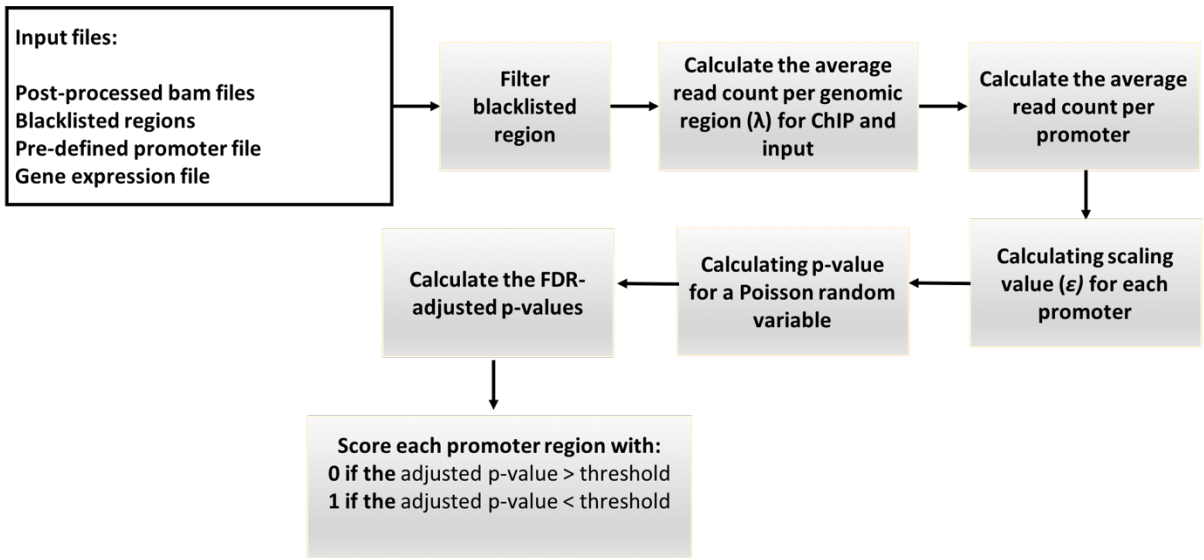


Figure 2-8: A workflow diagram of the orthogonal pipeline to call enriched genomic region.

A comparison between both promoter calling approaches were performed. For this purpose, I selected few promoters randomly and I assessed the reported promoter call from both approaches by visualizing the enrichment of each mark and see which approach gave the right call based on the enrichment. Also, I integrated the RNA-seq data to correlates the enrichment with gene expression at the selected promoters. Once the best approach was selected, I tried to analyse the external dataset and see if the selected approach along with the optimized p-value is suitable for the analysis and generates results similar to those published in the paper. I assessed the chromatin state transitions between GSC8 and GSC8per and the changes in gene expression.

2.2.9.3 CUT&RUN analysis pipeline

CUT&RUN raw sequencing data was analysed using the pipeline that was described in **Section 2.2.8.1** with slight modifications. Once the sample was post-processed, the quality of the data in terms of fragment size distribution, adapter content percentage, library size and read duplication rate was assessed. This was different in ChIP-seq data processing pipeline in which library complexity was computed as described in **Section 2.2.8.1**. The distribution of the fragment size of each sample was assessed using the *bamPEFragmentSize* function of Bedtools with default parameters. The remaining metrices were obtained from picard tool.

With regards peak calling, I applied MACS3 as described in **Section 2.2.8.1**. In addition, I applied SEACR (v1.3) as recommended by CUT&RUN guidelines with default parameters and the output was

compared with MACS3 outputs. Based on the results, I optimized the parameters of the selected tool to get better results. The pipeline was then processed exactly as described above.

Chapter 3
Experimental optimization and validation of JARID2 antibodies

3.1 Introduction

The primary aim of this study is to compare the epigenetic profiles of JARID2 and EZH2 binding and prevalence of H3K27me3 and H3K4me3 in patient–derived GBM cell lines derived from the primary and matched recurrence from the same patient. This can be achieved by performing ChIP-seq on these cell lines, however, the success of ChIP-seq experiments mainly relies on antibodies that can recognize the target proteins correctly. Therefore, it is necessary to validate all antibodies used in the ChIP-seq experiment to generate high-quality data. Antibodies against EZH2, H3k27me3 and H3k4me3 were validated in our group previously, so as per objective 1, I focused on validating JARID2 antibodies as this validation is central to my project. In this chapter, I detailed the approaches that I used to assess the specificity and the sensitivity of the selected JARID2 antibodies according to the ENCODE and modENCODE consortia and demonstrate that one antibody should work for our intended applications.

3.1.1 Antibodies

For many decades, antibodies have been crucial in the development of protein detection. They are among the most frequently used reagents in biomedical research, and in diagnostic and therapeutic applications (138, 163). Protein-based biochemical assays such as western blot and immunoprecipitation, cell-based assays such as flow cytometry and immunohistochemistry and proteomic assays use antibodies. Antibodies, also known as immunoglobins, naturally exist as a protective Y-shaped glycoprotein produced by B-lymphocytes in response to a foreign antigen. The antibody structure is composed of four polypeptide chains compromising, two identical heavy chains (H) and two identical light chains (L) linked to each other by disulphide bonds (**Figure 3.1**) (164, 165). Each chain of the heavy and light chains consists of constant (C_H and C_L) and variable domains (V_H and V_L).

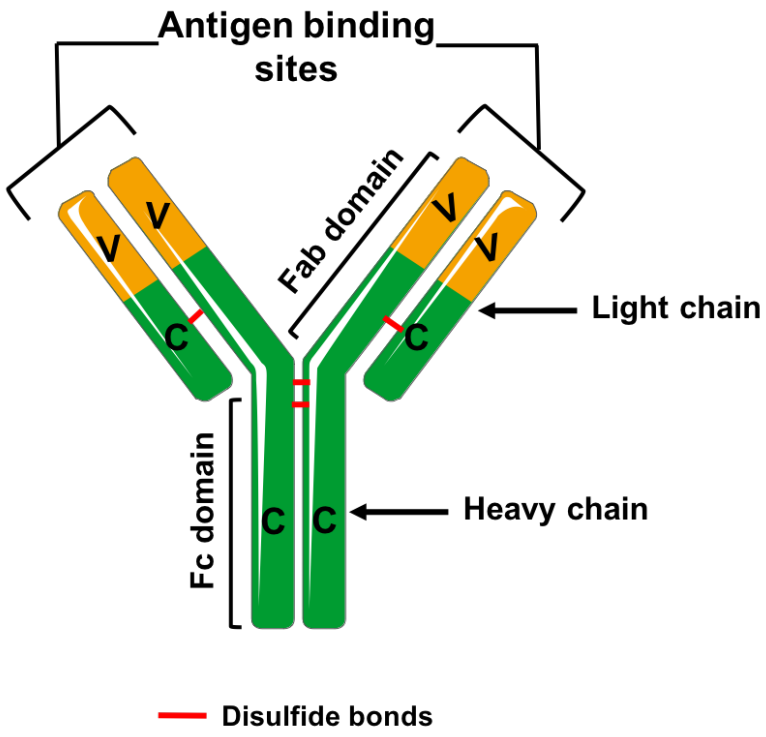


Figure 3-1: Schematic representation of antibody structure. An antibody consists of four polypeptide chains: two identical heavy chains and two identical light chains connected by disulphide bonds (Red lines). Each light and heavy chain contains an amino-terminal variable (V) region and constant (C) region.

This structure, in simplistic terms, facilitates antibody molecules to carry out two main dual functions in various regions of its structure. The first function is the recognition of the antigens which is achieved by antigen-binding fragment (Fab) and the second function is the removal of the antigen which is promoted by the interaction of the crystallizable fragment (Fc) with phagocytes or components of the complement pathway (164, 166, 167). Five major isotypes of antibody molecules can be found in the serum - IgM, IgG, IgA, IgD and IgE but the most frequently used isotype is IgG (164, 168). They are differing in the constant domain of the heavy chain they contain and their function (167). Two different types of antibodies have been made available to researchers to fulfil different research needs: monoclonal antibodies (mAbs) and polyclonal antibodies (pAbs). The details of these two types along with their production strategy are described below.

3.1.2 Commercial production of research antibodies

Production of antibodies has emerged as an important element across many research disciplines. The antibodies are mainly produced by safe immunization of a purified antigen into host species, commonly mice or rabbits. This results in high expression levels of antigen-specific antibodies in the serum that can subsequently be extracted from the host animals (168-170). The generation of mAbs are first established by Kohler and Milstein in 1975 using a hybridoma technique and they awarded Nobel prize for this (171). In the case of mAbs, immortal myeloma cells are fused with antibody-secreting spleen cells from immunized animals to produce monoclonal hybridoma cell lines, which express a particular and unique antibody in the cell culture (170, 172, 173). In the case of pAbs, host animals are injected with antigen or antigen/adjuvant combinations in order to induce efficient antibody responses. Serum must typically be collected in order to monitor the response and to get the antibody. PAbs are capable of recognizing several different epitopes on the antigen, whereas, mAbs bind only to one single epitope on the antigen. The production of pAbs production is inexpensive and relatively quick in comparison to the generation of mAbs which is considered more expensive and time consuming (169, 172-174). Also, the generation of mAbs from hybridoma cell lines introduces genetic variability during hybridoma formation and this results in a frequent lacking of antibody specificity. The main reported issues of pAbs are batch-to-batch variations, high background and cross reactivity (175).

Antibodies used as experimental reagents, typically, are classified as either primary antibodies that bind directly to specific antigens of interest or secondary antibodies that bind to the target-bound primary antibodies (170, 176)(**Figure 3.2**). Secondary antibodies are usually conjugated with fluorophores, such as rhodamine or fluorescein isothiocyanate (FITC), or enzymes, such as horseradish peroxidase (HRP) or alkaline phosphatase (AP), or biotin to enhances the detection and visualization of unconjugated primary antibodies bound to antigens (176-178). The selection of conjugate is dependent upon the desired application. These applications include western blotting, flow cytometry, ELISA, immunopurification, immunohistochemistry (176, 179). Despite the presence of various types of antibodies, their production requires careful designing, planning and implementation (169).

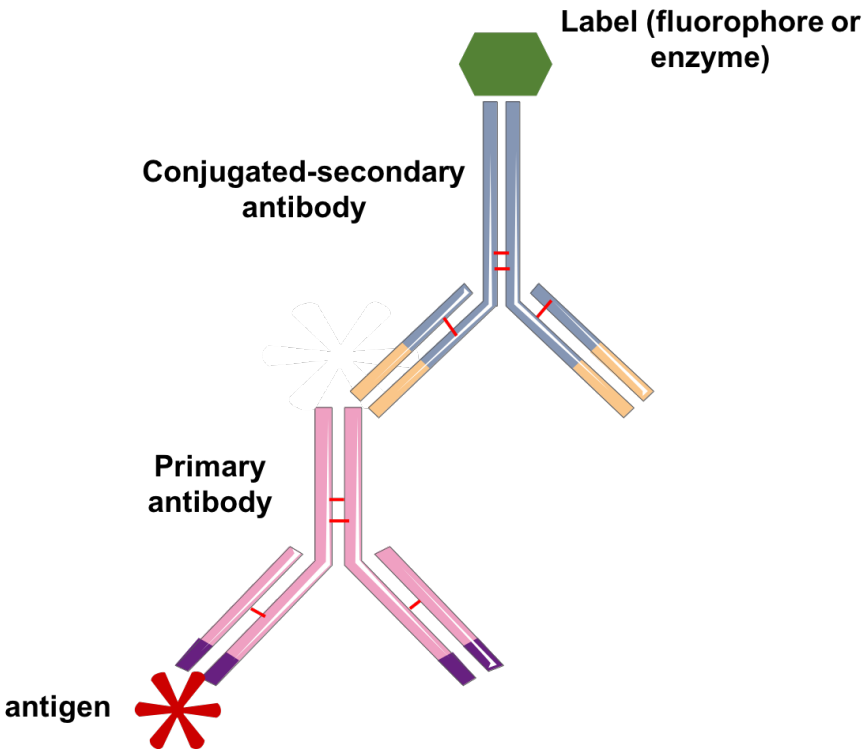


Figure 3-2: Schematic representation of primary and secondary antibodies. Primary antibodies are antibodies that bind directly to an antigen, whereas, secondary antibodies are antibodies that bind to the target-bound primary antibodies. The secondary antibodies are usually conjugated or labelled with either fluorophore or enzymes such as horseradish peroxidase (HRP), alkaline phosphatase (AP), rhodamine, fluorescein isothiocyanate (FITC), or biotin to facilitate the detection of the signal.

Commercial-scale production of antibodies is increasing each year to meet the market demands and much effort has therefore been made to generate high quality and stable antibodies (180, 181). However, the quality of the commercial research antibodies has been repeatedly put into question lately (175, 182). Major challenges that the scientists face are the lack of highly specific antibodies and inadequately validated antibodies (182, 183). Nonspecific antibodies often lead to inaccurate and irreproducible findings, therefore, confirming antibody specificity is critical to achieve accurate and consistent data (180, 182). Antibodies must be validated for each intended application and the validation criteria should at least include target specificity, especially in the application in which it is going to be used, and reproducibility (138, 139).

3.1.3 Antibody Validation

Antibody validation can be defined as the procedure of assessing the selectivity, specificity and the reproducibility of the selected antibody (138, 139). Despite the fact that antibodies play a prominent role in the reproducibility of research data, there are no universal guidelines that define how antibodies should be validated but there are some accepted standardized methods that can be used to determine its validity (138, 183). Given this limitation, the main objective (objective 1) of this work is to set up a method to validate the specificity of the selected antibodies and make it available as antibodies validation mechanisms for any future work.

I applied different methods as recommended by ENCODE consortium to assess the specificity of the selected antibodies and this involves a primary and a secondary method (**See Chapter 1, Section 1.6.1.1**). First, I applied the knockdown approach to knockdown the target protein using small interfering ribonucleic acid (siRNA) followed by western blot analysis (**See Chapter 2, Section 2.2.4.1 for detail of siRNA assay**). As described in **Chapter 2, Section 2.2.4.1**, siRNA knockdown is a biological mechanism where small synthetic interfering RNA mediates gene silencing by targeting and degrading mRNA transcript (184, 185). This assay is routinely used in the scientific laboratories to evaluate the specificity of a new antibody, because if mRNA is degraded, it means no more protein can be made and so cells with the gene knocked down by siRNA should have a reduced band on the Western Blot. However, even if the antibody is specific, siRNA knockdown is not always completely effective and it is usually transient, thus depending on the reduction level of mRNA and the stability of the target protein, there could not be a noticeable reduction in a band on the Western Blot (186-188). This can lead the researchers into attempting another effective method of validation which is overexpressing the protein in question and inspecting for a denser band (183).

I next used an overexpression method in which an epitope tagged version of JARID2 was overexpressed, immunoprecipitated with tag antibody and immunoblotted using protein specific antibody. This method enables us to confirm the specificity of the selected antibodies (**See Chapter 2, Section 2.2.4.9**). It has been widely used as a complementary assay to verify the antibody specificity. It is one of the most commonly applied method to study and identify protein-protein interaction (134). This interaction is detected only if the antibodies against the protein of interest is specific to its target. The main advantage of using epitope tagged protein is to test the specificity of antibodies against the protein of interest in IPs pulled down with different antibodies (i.e. flag) (138, 189, 190).

The selection of JARID2 antibodies was based on the available reviews and publications and whether they were tested and used for Chip-seq application. The list of all selected JARID2 antibodies are provided in the **Table 2.1** (**See Chapter 2, Section 2.2.2**).

In this chapter, I present the approaches used to validate JARID2 antibodies and the results obtained from each approach. Also, I described how I troubleshooted the problems with the obtained results.

3.2 Results

3.2.1 JARID2 siRNA knockdown effectively induces the mRNA degradation of target transcripts

JARID2 exists as three separate transcript isoforms based on the latest ENSEMBL (release 96) human gene annotation (**Figure 3.3**). Genomic alignments of each individual sequence of the SMARTpool siRNA reagents (**See Chapter 2, Section 2.2.4.1 for a detail descriptions of these reagents**) with the human reference genome indicated that these transcripts can each be targeted by at least two out of four smart pool siRNAs.

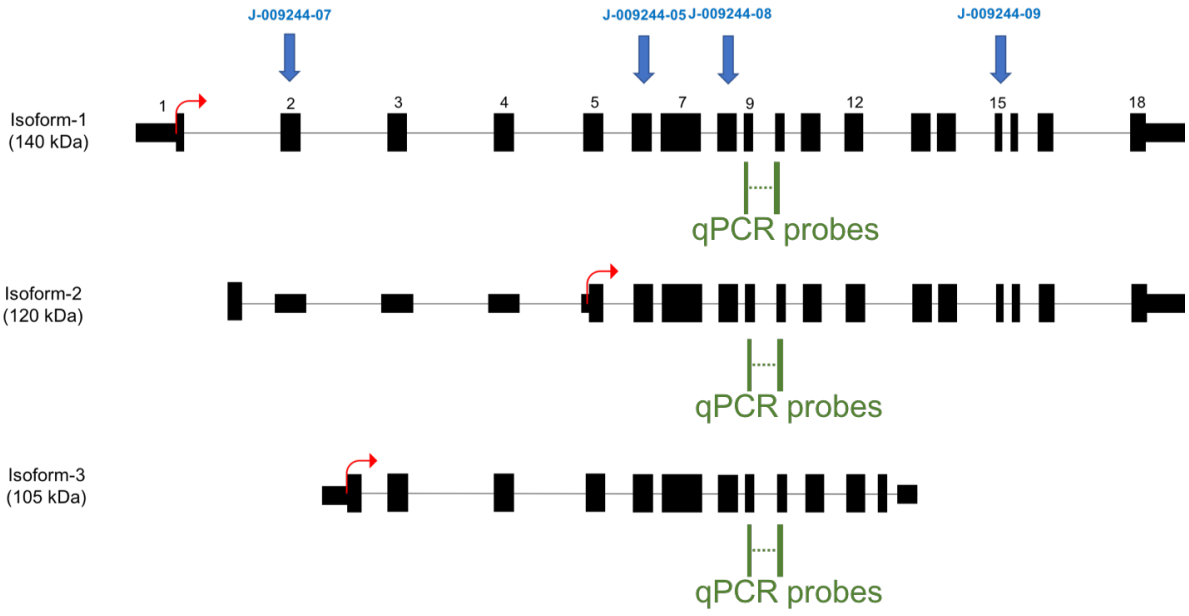
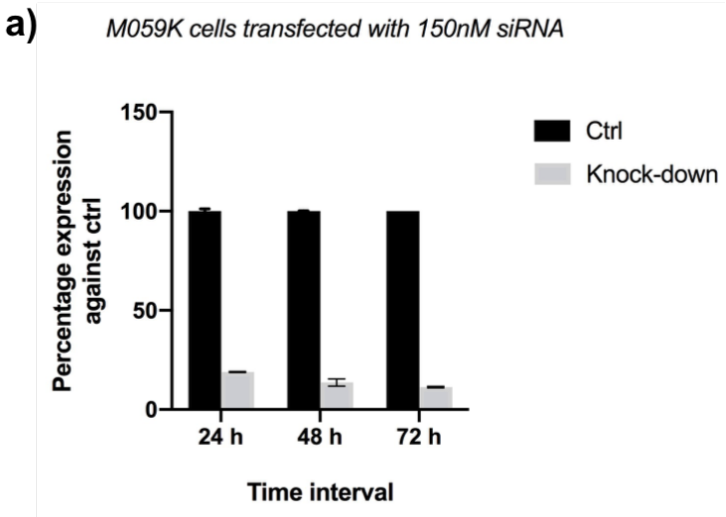


Figure 3-3: Schematic representation of JARID2 isoforms showing siRNA binding sites and qPCR TaqMan probes. JARID2 exists as three different isoforms, each encoding a protein of differing molecular weight (140, 120 and 106kDa). Genomic alignment of each individual sequence of the SMARTpool siRNAs used in this study (blue arrows) indicated that all three isoforms are adequately targeted. qPCR probes are represented by green lines. Translation start sites are represented by

curved arrows. Schematic Adapted From: Al-Raawi, D. et al., 2019. A novel form of JARID2 is required for differentiation in lineage-committed cells. The EMBO journal, 38(3), p.e98449.

Validation of the selected JARID2 antibodies was first performed using siRNA knockdown of JARID2 in the M059K GBM cell line. M059K cells were transfected with either pooled siRNA targeting JARID2 or a pool of non-targeting siRNA which was used as a negative control for 24 hr, 48 hr and 72 hr. Knockdown efficiency of JARID2 in the mRNA level was examined and quantified using qPCR. qPCR analysis showed a successful knockdown of JARID2 up to 72 h. (**Figure 3.4 a**). As shown below, the expression level of JARID2 in cells that were transfected with JARID2 siRNA was lower than the cells that were transfected with non-targeting siRNA by more than 70%.

The experiment was repeated using HEK293T cells (a human embryonic kidney cell line known to be highly transfectable) and the GBM63 patient-derived GBM cell line. These cells were transfected with siRNAs at final concentrations of 50nM for only 72h. A similar result was obtained in which a successful knockdown of JARID2 was observed at the 72hr time point (**Figure 3.4 b**). The expression of JARID2 was reduced by > 75% in the cells that were transfected with JARID2 siRNA, compared with control. In the first experiment, the reduction was observed at 24 h and this suggested that Lipofectamine® RNAiMAX, as the siRNA transfection reagent, is suitable for siRNA transfection assay. It is also indicating that this reagent is compatible with the selected cell lines.



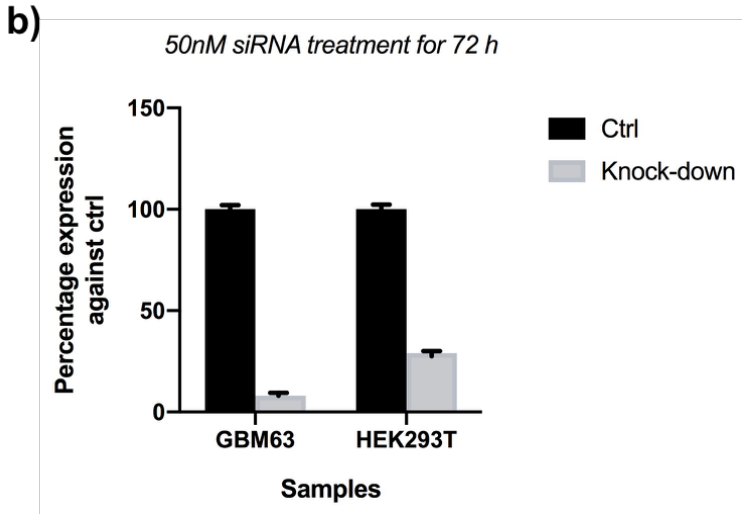


Figure 3-4: siRNA knockdown efficiency of JARID2 in M059K, HEK293T and GBM63 cells. Total cellular RNA was extracted from MO59K cells transfected with 150nM of JARID2 siRNA or non-targeting control siRNA for 24, 48 or 72 h and JARID2 transcript quantified using TaqMan qRT-PCR **(A)**. Total cellular RNA was extracted from GBM63 and HEK293T cells transfected with 50nM of JARID2 siRNA or non-targeting control siRNA for 72 h and JARID2 transcript quantified using TaqMan qRT-PCR **(B)**. Graphs depict relative gene expression normalized to a house keeping gene, GAPDH. Bars represent mean ± S.D of 3 technical replicates.

3.2.2 JARID2 siRNA knockdown had no observable effect on the JARID2 protein level

JARID2 exists as three different isoforms, each encoding a protein of different molecular weight (140, 120 and 106kDa). In addition, a cleaved product of full-length JARID2 denoted ΔN-JARID2 (~80 kDa) was recently identified (**Figure 3.5**). For the purpose of this study, four different JARID2 antibodies (**See table 2.1 in Chapter 2, section 2.2.2 for the list of antibodies**) that targets different regions of JARID2 protein were used in this study.

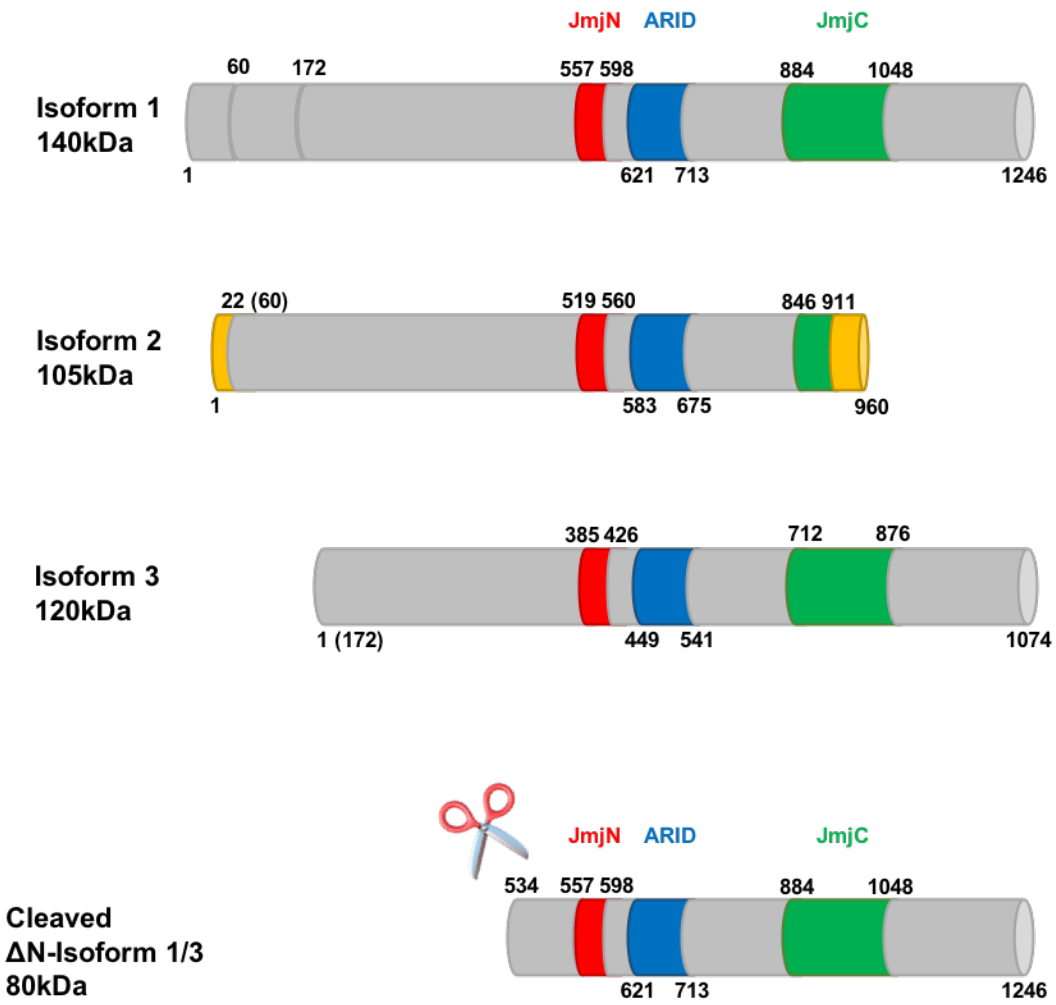
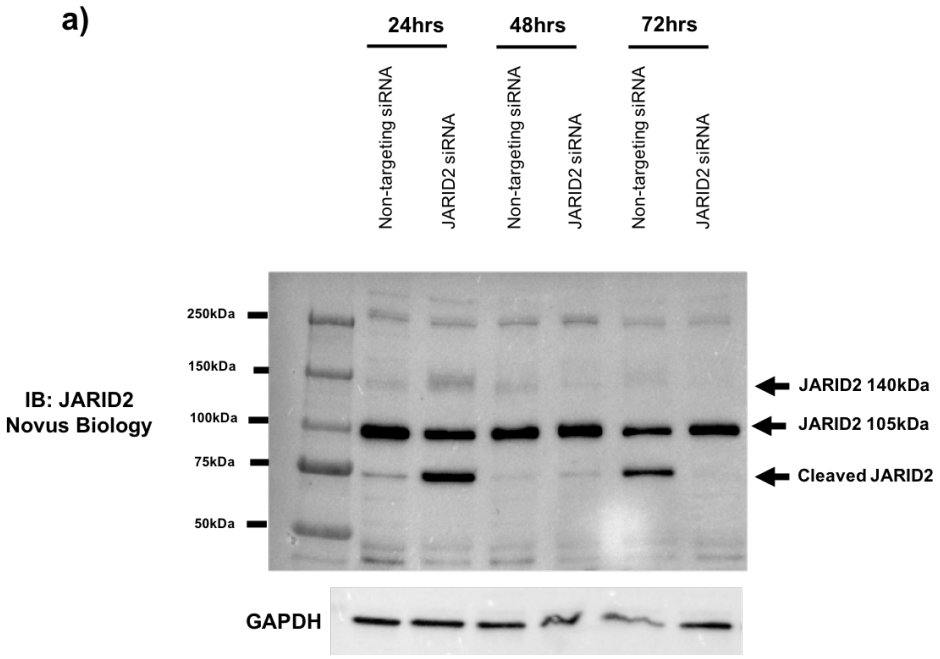


Figure 3-5: Schematic diagram represent the isoforms of JARID2 protein with their functional domain. Three different isoforms of JARID2 protein was identified with different molecular weight (140, 120 and 106kDa) along with the cleaved product of full-length JARID2 (ΔN-JARID2, ~80 kDa). The diagram shows the main functional domain of JARID2: JumonjiN (JmjN, red), AT-rich DNA binding domain (ARID, blue) and JumonjiN (JmjC, green). Regions with different protein sequences from the canonical isoforms are represented in yellow. Reproduced from https://www.uniprot.org/uniprotkb/Q92833/entry by UniProt consortium, 2023.

To assesses the knockdown efficiency on the protein level and to see if the knockdown in the protein level reflects the qPCR results, I immunoblotted the protein lysates from M059K cells transfected with JARID2 siRNA and non-targeting siRNA for 24, 48 or 72 h. The immunoblot was first performed on a nitrocellulose membrane using JARID2 antibody (Novus biology) that recognizes the N-terminal region of JARID2 (expected to detect the full length 140kDa form of JARID2). Protein expression analysis did not show any reduction in JARID2 signal in the same order as the transcript levels (**Figure**

3.6 a), with JARID2 isoforms 1 and 3 quantified via normalisation to the loading control, GAPDH (**Figure 3.6 b&c**). I did not observe the expected change in protein level between JARID2 siRNA transfected cells and non-target siRNA transfected cells. In addition, non-specific bands were observed.



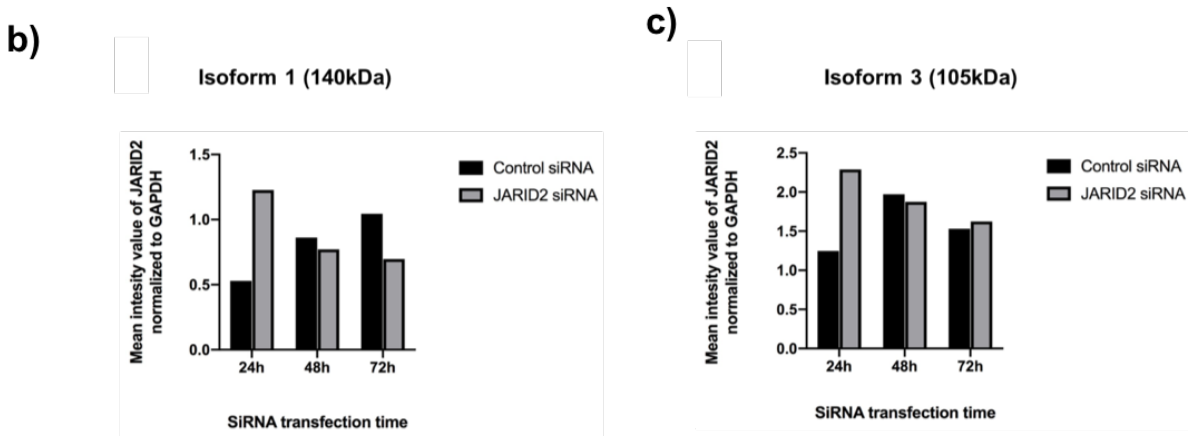
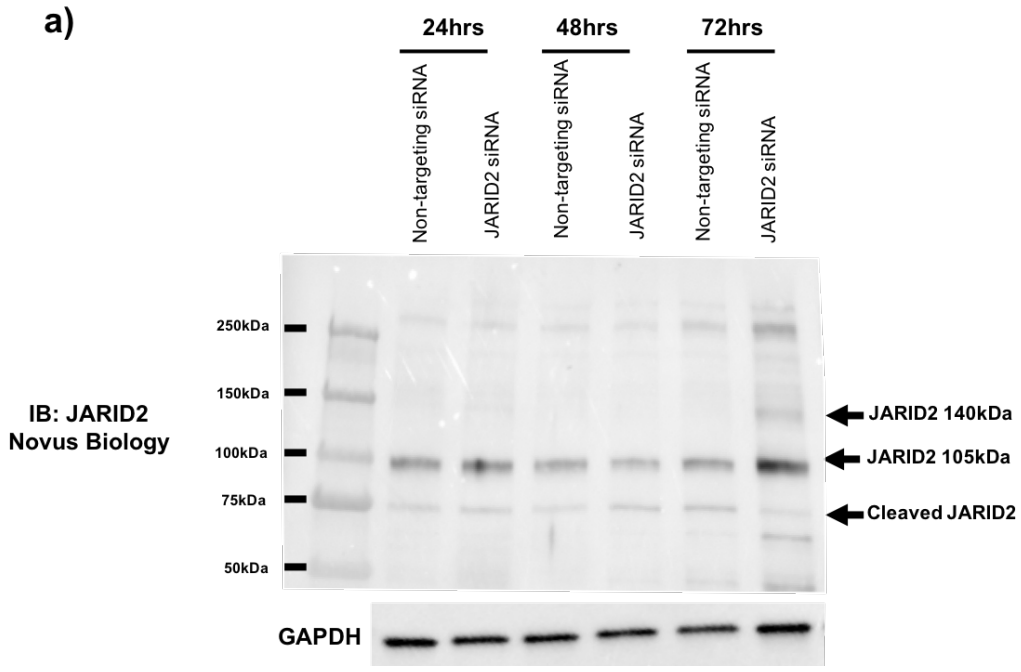


Figure 3-6: Western blot analysis of the efficiencies of siRNA knockdown of JARID2 in M059K cells. Total protein was extracted from M059K cells transfected with either JARID2 siRNA and non-target (control) siRNA for 24 hrs, 48 hrs and 72hrs **(A)**. Western blot was carried out using a JARID2 antibody that recognizes the N-terminal region of JARID2 protein with GAPDH (37kDa) used as loading control. western blot analysis did not show any reduction in JARID2 signal in the same order as the transcript levels. Also, multiple non-specific bands were observed. The blue arrows show the possible JARID2 140kDa and 105kDa bands. The bar graphs represent the quantification of the western blot results of JARID2 signal of **(B)** isoform 1 (140kDa) and **(C)** isoform 3 (105kDa) normalized to GAPDH using image lab.

The western blot in **Figure 3.6** had used nitrocellulose membrane as a transfer substrate. There is some evidence that antibody binding can be dependent on the type of membrane and so this was initially repeated using polyvinylidene difluoride (PVDF) membrane instead. Unfortunately, this resulted in complete absence of JARID2-assoicated bands at 140 kDa (**Figure 3.7 a**). Interestingly, the predominant band was observed at ~ 105 kDa in this case and a faint band was observed at ~ 75 kDa which could represent the cleaved N-terminus of JARID2 in the truncated protein isoform. Signal intensities of isoform 3 (105kDa) was quantified and normalized to GAPDH (**Figure 3.7 b**). Different results were generated using nitrocellulose and PVDF membrane, thus, it was difficult to evaluate the knockdown of JARID2.



b)

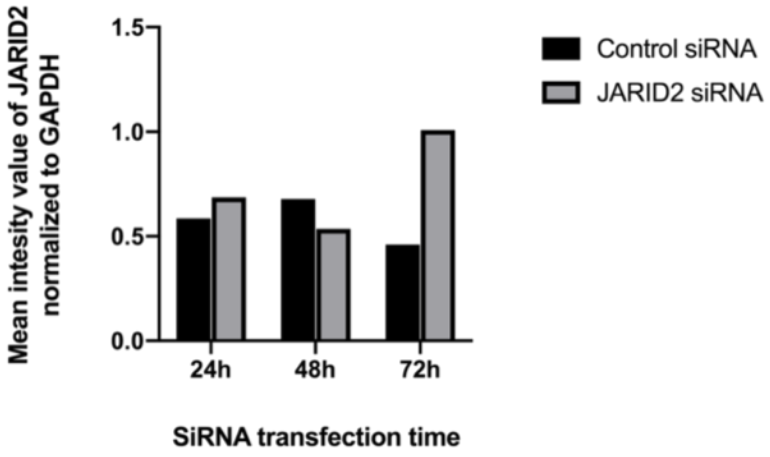
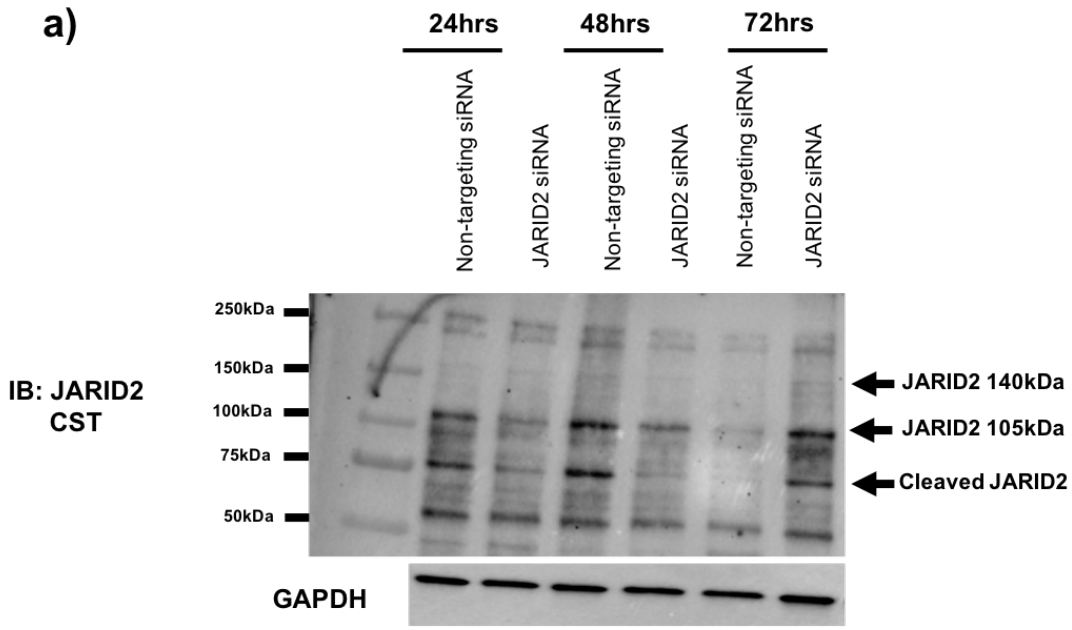


Figure 3-7: Western blot analysis of the efficiencies of siRNA knockdown of JARID2 expression in transfected M059K cells. Total protein was extracted from M059K cells transfected with JARID2 siRNA and non-targeting siRNA and western blot was carried out using PVDF membrane and JARID2 antibody that recognizes the N-terminal region (Novus biology) of JARID2 protein with GAPDH (37kDa) used as loading control **(A)**. A complete absence of JARID2 isoform 1 (140kDa) band in all lanes was observed and strong bands were observed for all samples at 105kDa. The bar graphs represent the quantification of the western blot results of JARID2 signal of **(B)** isoform 3 (105kDa) normalized to GAPDH using image lab.

I reverted to the nitrocellulose membrane and attempted the second antibody (i.e. from Cell signaling technology), which recognizes the C-terminal region of JARID2 protein sequence and may be expected to detect the full length (140kDa), isoform 3 and the cleaved C-terminus form (80kDa) of JARID2 (**Figure 3.8**). An almost complete absence of 140 kDa band was noticed for all samples and therefore, it was difficult to evaluate the knockdown for this canonical isoform at the protein level. However, clear bands (strong signal) were observed at 105kDa and ~75kDa for all samples. Using this antibody, more non-specific bands were observed in comparison with the first (N-terminal detecting) antibody. Signal intensities of isoform 3 (105kDa) and the cleaved form were quantified and normalized to GAPDH (**Figure 3.8 b&c**).



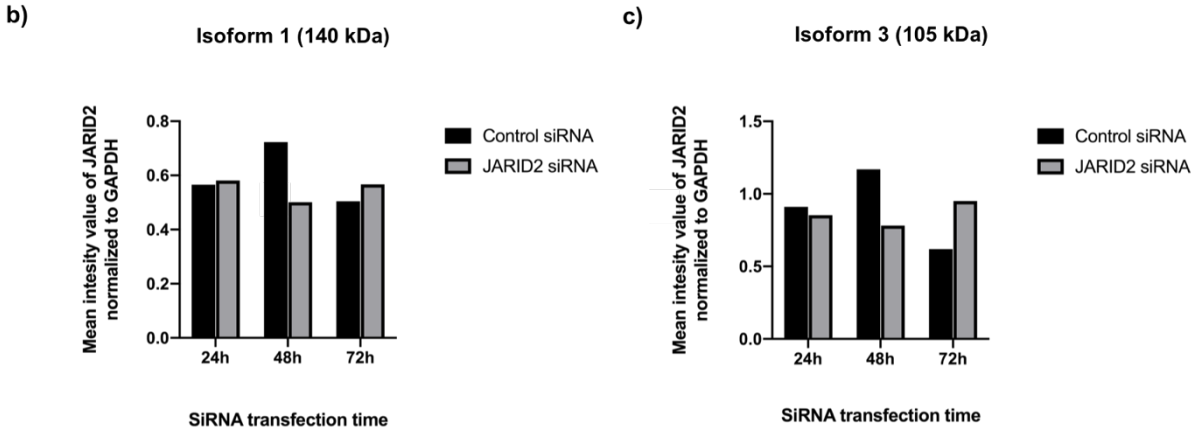


Figure 3-8: Western blot analysis for validation of knockdown of JARID2 using siRNA in M059K cells. (A) Western blot of M059K cells transfected with JARID2 siRNA and non-target siRNA for 24h, 48h and 72h carried out using nitrocellulose membrane and anti-JARID2 (CST) that recognizes the C-terminal region of JARID2 protein sequence. A complete absence of the canonical JARID2 protein (140 kDa) was observed and clear bands were observed in all lanes at ~75kDa and 105kDa. The bar graphs represent the quantification of the western blot results of JARID2 signal of **(B)** isoform 1 (140kDa) and **(C)** isoform 3 (105kDa) normalized to GAPDH using image lab.

Western blot analysis was then attempted on GBM63 and HEK293T lysates in 3 technical replicates with two different antibodies that recognize the N-terminal (ab192252) and the C-terminal regions (ab93288), and similar results were obtained. No reduction in the intensity of the canonical (140kDa) isoform (Figure 3.9 a-b) was observed. Fewer non-specific bands were observed for the N-terminal antibody (i.e. ab192252) in comparison with all antibodies that were used in this study. These findings might imply issues with antibody specificity, therefore, the N-terminal antibody (ab192252) was selected as optimum for the remaining experiments.

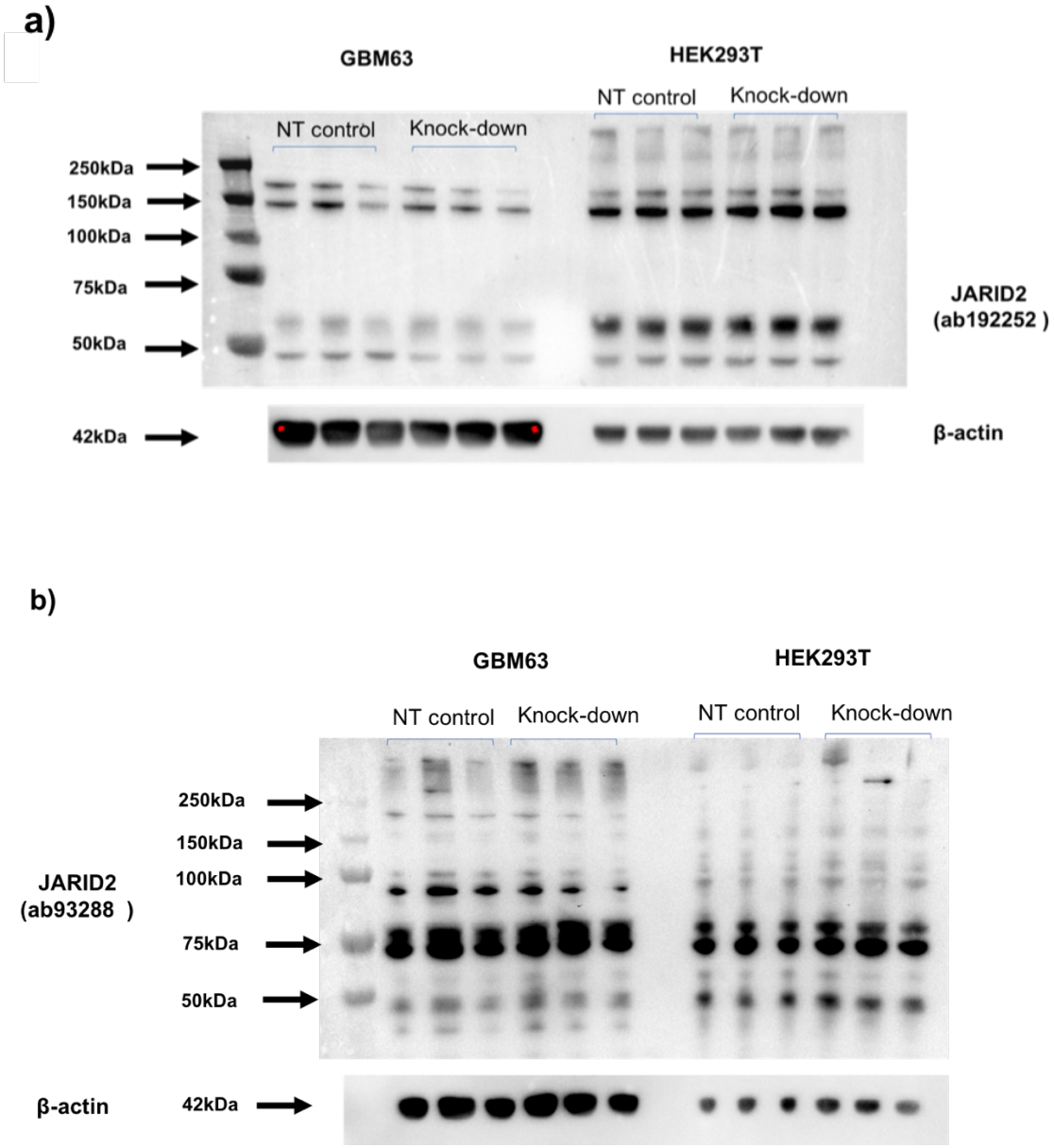


Figure 3-9: Representative image of western blot of siRNA transfected samples. Western blot of siRNA transfected cell lines (i.e. GBM63 and HEK293T) probed for two JARID2 antibodies: anti-JARID2 ab192252 **(A)** and ab93288 **(B)** with β-actin displayed as a loading control. Same lysate was

divided into 3 technical replicates and each replicate was loaded in each well. There is no reduction in JARID2 intensity after 72 hours.

3.2.3 The specificity of the selected JARID2 antibodies was verified via overexpression of the full length tagged JARID2 and co-immunoprecipitation (Co-IP) assays

3.2.3.1 Construction of JARID2 expressing plasmids

Next, I aimed to verify the specificity of the selected JARID2 antibody through overexpression of JARID2 protein and examine the presence of the exogenous version of this protein in the transfected cells. In this experiment, the design and construction of a full length JARID2 plasmid bearing 3xFlag tags (3xflag-JARID2) was achieved using LR cloning gateway methodology (**see Chapter 2, Section 2.2.4.6**; this was completed by Dr Marilena Elpidorou, a postdoc in the Stead group). The maps of each plasmid are provided in Appendix D. The resultant expression plasmid was purified and verified by Sanger sequencing using CMV, M13 and SV40 primers along with internal primers designed within the JARID2 sequence (**see table 2.5 in chapter 2, section 2.2-4.8**).

3.2.3.2 Immunodetection of the exogenous flag-tagged JARID2 protein via western blot

The specificity of the selected JARID2 antibodies were verified via the overexpression of the tagged version of the gene followed by western blot assay. The experiment was performed on HEK293T cells transfected with the experimental flag-tagged plasmid, 3xflag-JARID2 and an empty GW306 control plasmid (**see Chapter 2, Section 2.2.4.9**). Cells were grown to 60-70% confluency, lysed, protein extracted and analysed by western blotting to characterize JARID2 expression.

Protein expression analysis of HEK293T plots confirmed the recognition of an overexpressed N-terminally flag-tagged JARID2 protein in the cells that have been altered to overexpress JARID2. As shown in **Figure 3.10**, a higher molecular weight band, present only in the transfected cells was visualized using a flag-tag antibody (**Figure 3.10a**) and JARID2 (ab192252) antibody (**Figure 3.10b**). The flag-tag is only 8 amino acids long which corresponds to approximately 1 kDa in molecular weight. Also, a JARID2-associated band at 140 kDa was detected using JARID2 antibody (**Figure 3.10b**), which was not detected in a membrane probed with flag-tag antibody. This indicates that flag-tagged JARID2 was successfully expressed in HEK293T cells as a result of the overexpression experiment. On the other hand, no band of flag-tagged JARID2 protein was observed in non-

transfected cells or cells transfected with control plasmid. The absence of this band in these cells confirms that the signal is specific to transfection with flag-tagged JARID2 plasmid. It confirms that anti-JARID2 antibody (i.e. ab192252) is target-specific.

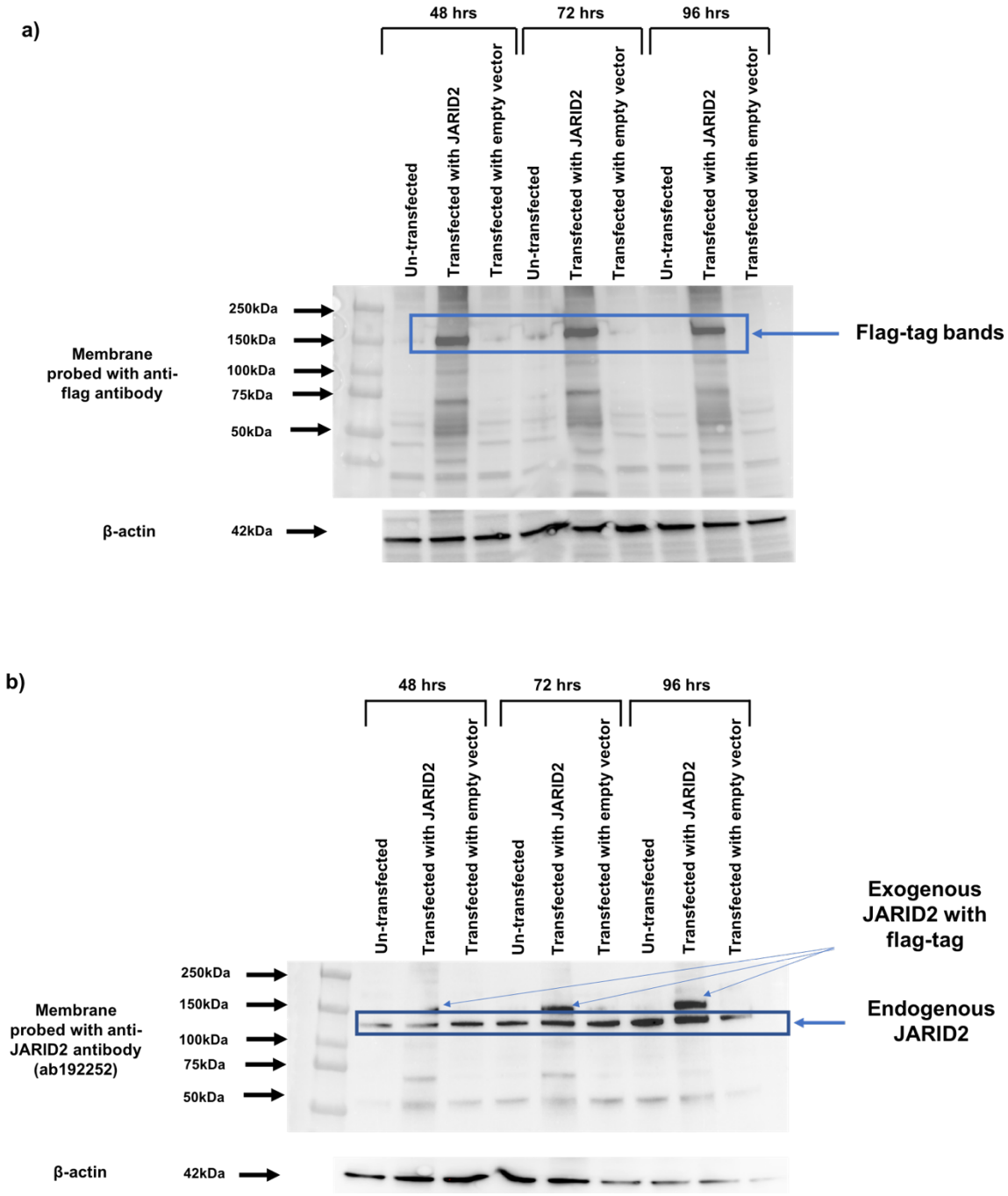


Figure 3-10: Overexpression of flagged-tagged JARID2 in HEK293T cells. (a&b) Western blot analysis of HEK293T cells transfected with flagged-tagged JARID2 and an empty vector revealed the

expression of exogenous and 3xflag-tagged JARID2 in cells transfected with flagged-tagged JARID2 construct only using anti-flag and anti-JARID2 (ab192252) antibodies. β-actin was used a loading control.

To verify this observation further, an immunoprecipitation (IP) experiment was carried out on lysates transfected with plasmids encoding either flag-tagged JARID2 or GW306 control as a control (**see Chapter 2, Section 2.2.5.4**). To perform the IP, the transfected cells were subjected to IP with JARID2 (ab192252) or flag antibodies. The resulted pull-down protein complexes were run on western blot along with whole cell extracts and probed with either JARID2 (ab192252) or flag antibodies. Results obtained from western blot analysis showed that flag-tagged JARID2 protein was detected and bound by anti-JARID2 antibody and efficiently pulled down using protein A-agarose (**Figure 3.11 a-b**). A clear strong signal at ~143 kDa was observed in cell transfected with flag-tagged JARID2 construct and in IPs pulled with JARID2 and flag antibodies. No bands of flag-tagged JARID2 protein were observed in non-transfected cells or cells transfected with control plasmid. Additionally, JARID2-associated band at 140 kDa was observed in a membrane immunoblotted with JARID2 antibody (**Figure 3.11 b, Lane 1,2,3,5 and 7**). These results together, demonstrated that overexpression studies and IP make use of epitope tags (i.e. flag tag) for studying the specificity of the antibody. In general, in the above tested applications and conditions, the specificity of JARID2 antibody and more specifically ab192252 was validated and proved.

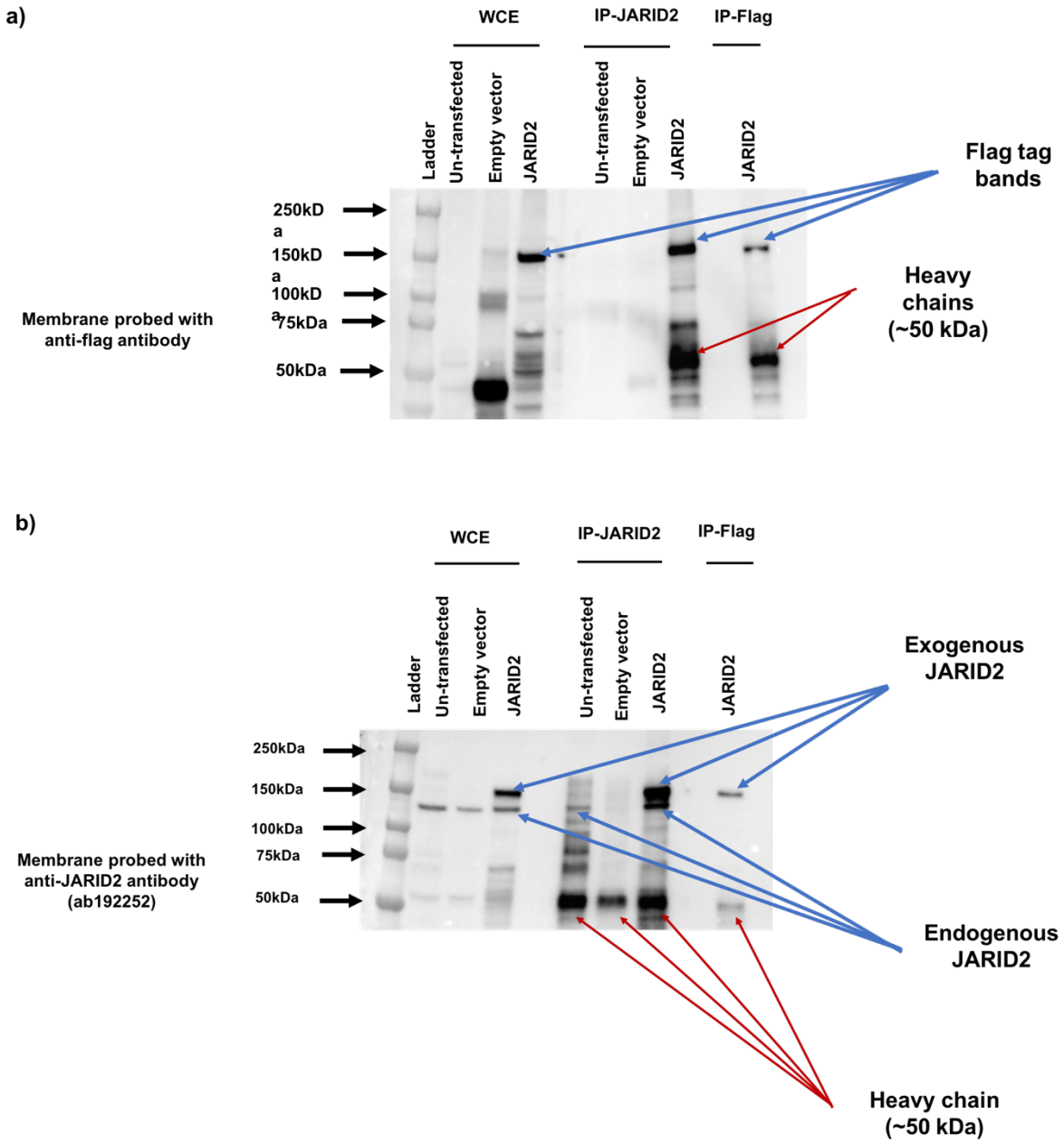


Figure 3-11: Western blot of overexpressed protein after co-immunoprecipitation assay. HEK293T cells were transfected with either 3xflag-tags or GW306 plasmids for up to 96 h. Co-immunoprecipitation of purified 3xflag-tag JARID2 protein using antibodies against flag-tag and JARID2 (ab192252). The western blot was developed with anti-flag **(a)** and anti-JARID2 antibody (ab192252) **(b)** and both confirmed the expression of flagged-tagged JARID2 protein in the cells transfected with 3xflag-tag JARID2 plasmid compared to non-transfected and cells transfected with empty vector. Red arrows represent the IgG heavy chain at a molecular weight of ~ 50 kDa.

3.3 Discussion

Antibodies are among the most widely used tools for protein detection, however, proper validation of their applicability for a given application is required before use. The specificity and the sensitivity of an antibody determines its usefulness (189). Several studies concentrate on recently identified proteins, such as in this project, hence an antibody to such a protein is not likely to have any proper validation. In this chapter I aimed to validate the selected JARID2 antibodies for the use in ChIP-seq and CUT&RUN assays as it is central to my project. I purchased several JARID2 antibodies from various suppliers that were sold as JARID2-specific (**Table 2.1 in Chapter 2, Section 2.2.2**).

Several methods have been established to quantitatively evaluate the performance of the antibodies. This includes western blot (WB), siRNA, immunoprecipitation (IP), immunofluorescence (IF), knockdown or knock out of the target protein, immunoprecipitation with an epitope-tagged version of the protein and immunoprecipitation followed by mass spectrometry (138). For ChIP-seq, ENCODE consortia suggested a primary and a secondary mode of assessments to characterize the specificity of antibodies (134). In the context of this project, I investigated the feasibility of these assays in characterizing the specificity of the antibodies. As an initial attempt, I used siRNA knockdown assay to determine the specificity of JARID2 antibodies in different cells lines using SMARTpool siRNA reagents that target different JARID2 regions (**Figure 3.3**). I demonstrated that siRNA knockdown of JARID2 caused significant reduction in the mRNA level as early as 24 h post transfection (**Figure 3.4 a&b**) but not at the protein level. The signal for JARID2 was still present (**Figures 3.6a, 3.7a, 3.9a and 3.9a**). Also, a complete absence of JARID2 band at the expected molecular weight of 140 kDa was observed when a PVDF membrane was used (**Figure 3.7a**) and when the membrane was probed with CST antibody (**Figure 3.8a**). In addition, several additional bands above and below the expected molecular weights for different JARID2 isoforms was observed and this also raises concerns about nonspecificity. Beside this, the presence of JARID2 protein after attempted knockdown might indicate that the protein is quite stable, therefore, remain relatively constant for a long period of time. Lack of specificity with commercial antibodies is relatively common; for example, this was reported by other group who tested two antibodies against HoxA1 and phospho-4EBP1 on lysates from ten cell lines using a simple western blot assay. They reported a complete absence of H0xA1 band at the expected molecular weight of 37 kDa in 9 cell lines out of 10. Likewise, they showed the presence of nonspecific bands at lower signal value level for both tested antibodies (138). On the contrary, Shuaib et al., examined the specificity of AGO1 antibody

for ChIP-seq application using the same approach. They showed that siRNA knockdown of AGO1 in HepG2 cells resulted in a successful degradation of AGO1 mRNA and a subsequent loss of the encoded protein in the cells that were transfected with siRNA against AGO1 compared to the cells that were transfected with control siRNA (191). Similarly, a successful knockdown in the mRNA and protein level was observed for siRNA-mediated WDR18 and EZH2 at 48 h post transfection (192).

It has been noted that mRNA level and the protein level do not always correlate with each other and mRNA measurement can overestimate knockdown of genes whose protein products have long half-lives or present in abundant quantities (184). To address this limitation, I applied an alternative approach to assesses the specificity JARID2 antibodies. Overexpression of an epitope tagged version of protein followed by western blot was used. In the present study, FLAG-tagged JARID2 expression plasmid was constructed first using LR gateway cloning technology prior to the overexpression experiment. I emphasized the success of LR cloning gateway experiment and the creation of 3x Flagged-tagged JARID2 expression plasmid via Sanger sequencing. After verification, plasmid transfection was performed aimed at overexpressing the exogenous Flagged-tagged version of JARID2 in HEK293T. HEK293T cells transfected with an empty vector and un-transfected cells were used as reference. The success of exogenous JARID2 transfer into the cells was confirmed by a simple western blotting experiment. Western blot analysis shows a clear band for the fused Flag-tag at a molecular weight of ~143 kDa (**Figure 3.10a**) when the membrane was probed with anti-Flag antibody and this indicates the success of the transfection experiment. The detection of the overexpressed Flag tag along with JARID2 (**Figure 3.10b**) reveals that JARID2 antibody (i.e. ab192252) is target-specific. Also, I confirmed the applicability of the epitope-tags approach in evaluating antibody specificity through its capability in distinguishing the endogenous from the overexpressed (i.e. exogenous) proteins. Despite this, only few studies reported the use of this approach for antibody validation (183, 193, 194).

The specificity was further verified through immunoprecipitation (IP) assay. In the present study, I performed an IP experiment to evaluate the use of the JARID2 antibody (i.e. ab192252). Lysates from HEK293T cells overexpressed with plasmids encoding either flag-tagged JARID2 or GW306 control as a control were subjected to IP with JARID2 (ab192252) and flag antibodies followed by western blot. Western blot analysis indicated that this antibody is target-specific due to the presence of strong signal at the expected molecular weight of JARID2 (i.e. 140 kDa) in the samples that were immunoprecipitated with anti-JARID2 antibody (**Figure 3.11 B, Lane 7**). Together, these

findings showed that epitope tags, such as flag tag, are used in overexpression experiments and IP to investigate the antibody's specificity. Generally speaking, the specificity of the JARID2 antibody, and more particularly ab192252, was confirmed and proven in the aforementioned evaluated applications and settings. Instead of being utilized for validation, gene overexpression and co-IP were used to characterize protein-protein interactions (195-197).

Collectively, the results indicated that the specificity of the available commercial antibodies is not as advertised and they often yield misleading results. Also, even if the antibody was designed to detect specific target protein, they may not always be successful in doing so in all applications. I concluded that among all the selected JARID2 antibodies, anti-JARID2 antibody (ab192252) is target-specific and performs well in all applications that have been tested in this work.

Chapter 4
Developing the computational approach

This chapter summarizes the development of ChIP-seq data analysis pipeline and the development of a promoter status calling approach to classify whether multiple DNA binding factors or histone modifications were present in a pre-defined promoter region or not. The chapter outlines the major steps involves in a typical ChIP-seq computational analysis workflow based on the ENCODE consortium. In addition, it presents comprehensive details on the performance of the developed approaches in calling and characterizing the promoter states across the promoter regions. The developed approaches were applied on two datasets: an external dataset published in Liau et al (198) and in-house dataset derived from DNA from a fresh frozen pair of primary and matched recurrent GBM samples. The results from both approaches were compared to see which approach is producing results that make the most biological sense. The selected approach was used for further downstream analysis aiming to call the promoter status for H3K4me3, H3K27me3, JARID2 and EZH2 binding in the above-mentioned datasets.

4.1 Introduction

As described earlier, epigenetics is the field of biology that studies how cells regulate gene activity without involving alterations in DNA sequence in a manner that is usually reversible. The well-understood phenomenon of DNA methylation, histone modifications, chromatin remodeling and non-coding RNAs are the main molecular mechanisms that mediate epigenetic phenomena (69, 90, 199). These mechanisms impact chromatin condensation, nuclear organization and the transcriptional state of the associated DNA, and therefore play a prominent role in modulating gene activity. Studying these mechanisms is important to discover regulatory regions and their cell type-specific activity patterns and for interpreting disease-association studies (200, 201). Several studies have demonstrated that these epigenetic modifications take place mainly in the gene promoter which is commonly referred to as a genomic region at which the transcription of the gene is initiated (202, 203). It contains the transcription start sites (TSS) (+1bp) and it is often located directly upstream of the gene or at the 5’ end of the transcription starts sites. As shown in **Figure 4.1**, the promoter is divided into three parts: (1) the core promoter which serves as a binding platform for the transcription machinery and it contains an RNA polymerase binding site typically situated ~ 34bp up stream of the transcription start site (TSS); (2) the proximal promoter a region that is normally found at ~ 250bp up stream of the TSS and contain several primary regulatory elements; (3) the distal promoter which is located upstream of the proximal promoter and contains transcription

factor binding sites along with additional regulatory elements such as enhancers and silencers (178, 204).

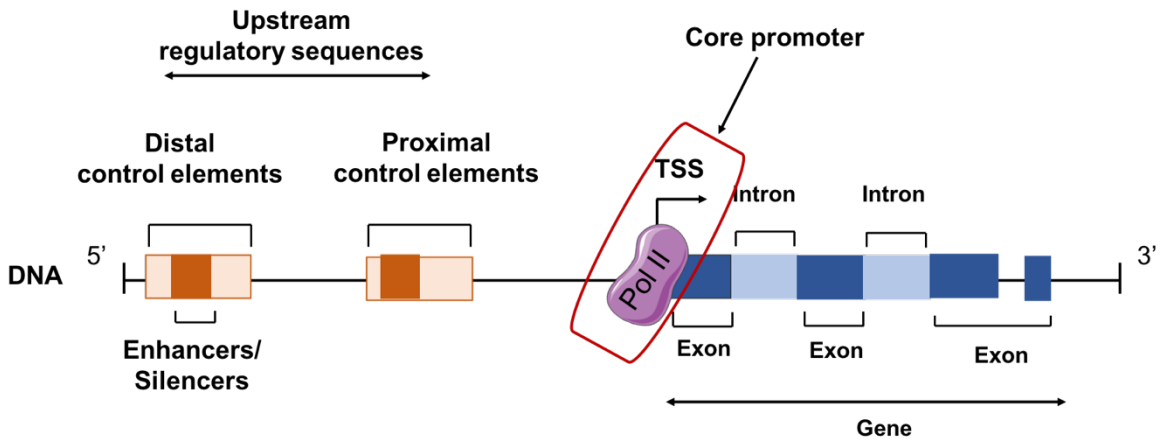


Figure 4-1: Schematic representation of the upstream regions that contribute to the full promoter region. The promoter region composed of the core promoter where transcription is initiated at the transcription start site (TSS, represented by black arrow) which is located at the 5' end of a gene RNA polymerase II (Pol II). The proximal region is located upstream of the core promoter (~250 bp) and facilitates transcription through the binding of the transcription factors. Upstream of the proximal region is the distal promoter region. Different regulatory elements, including silencers, enhancers, and cis-elements (orange rectangular boxes), which control gene expression at the transcriptional level, are present in these two locations.

The definition of the promoter region is not yet set. In this study, I used the definition of the promoter which is +/- 1kb around the TSS because it has been used in our group previously and found that a subset of genes is dysregulated in GBM's patient following standard treatment due to the epigenetic remodeling of their promoters via mechanisms involving JARID2 as an adaptive response that mediates tumour recurrence.

Genome-wide chromatin state annotations ("Chromatin state maps") provide a rich source of information about the epigenomic landscape and how it contributes to cell identity, development, lineage specification and disease, yielding insights beyond what is typically obtained by RNA expression profiling. In recent years, genome wide mapping of chromatin states in humans has been via high throughput sequencing technologies (205-207). Chromatin immunoprecipitation (ChIP)

coupled with sequencing is the most widely used technique to generate genome-wide map of histone modifications associated with diverse regulatory and epigenetic functions, including H3K27me3, H3K4me3, H3K4me1, H3K9me3 and H3K27ac (207, 208). Nowadays, multiple consortia such as NIH Roadmap, Epigenomics, Encode, blueprint and DEEP are providing genome-wide maps of histone modifications generated by ChIP-seq (209, 210).

The pattern of histone binding in different regions of the genome was originally identified using different peak finding tools as described in **Chapter 1, Section 1.6.1.2**. These tools separate the genome into regions of high and low enrichment of binding signal irrespective of the genomic function of these regions (152, 153, 211). The most commonly used tool is MACS2 and it has been used extensively to identify peaks in different studies (155, 212). However, these tools summarize only one ChIP-seq dataset (one experiment with a single antibody) at a time (212, 213). To integrate information across multiple datasets, researchers have shifted to unsupervised clustering approaches such as k-means and Hidden Markov Models (HMMs). These approaches enable the researchers to study various combinations of histone modifications and assign different regions of the genome into classes, or states, depending on the presence, absence or even the strength of signal for these modifications (157, 214). One of the most prominent approaches is ChromHMM which has proven its usefulness in determining the combinatorial patterns of multiple epigenetic marks as explained in **Chapter 1, Section 1.6.1.2**. It is a completely unsupervised HMM approach that defines the chromatin state through the presence or absence of histone marks or TF binding within specific segments of the genome but, these segments cannot be predefined by the user (157, 214, 215). I decided to utilize ChromHMM to output promoter status calling because it has been widely used to annotate the epigenome in Roadmap and in the ENCODE projects. it has been applied on 111 Roadmap primary cell lines and 16 ENCODE cell lines with six histone marks. In addition, it was used to characterize cancer subtypes according to their chromatin profile (216-218). In this study, I wished to be able to specifically call each promoter region's status according to the presence or absence of binding or modification signal from multiple ChIPseq experiments on the same sample.

One way to do this is was via adoption of ChromHMM to output promoter calling status so I could characterize the chromatin states in those specific regions in our datasets. I developed a bespoke java programming approach to use the standard ChromHMM output to assign a call for each promoter region, specifically, as per my objective 2. I denote this approach Approach 1 and I will

use this term in the context. To identify alternative methods, I also looked into the literature and tried to find if, and how, others had attempted to solve this problem. I found an approach in which a chromatin call was generated by scoring the enrichment of signal in defined regions compared to the background across the same sized windows across all genomic regions. I denote this approach Approach 2. I decided to apply both approaches and compare the outputs to see which is producing results that make the most biological sense.

In this chapter, I present the general ChIP-seq analysis workflow step-by-step based on ENCODE consortia guidelines, from quality assessment to chromatin-state annotation (**see Figure 2.6 in Chapter 2, Section 2.2-8.1**). I then describe how I developed the promoter calling status approaches and compared the output from them (**Figure 4.2**). I conclude with which approach I decided to adopt for application to my cell lines and patient data.

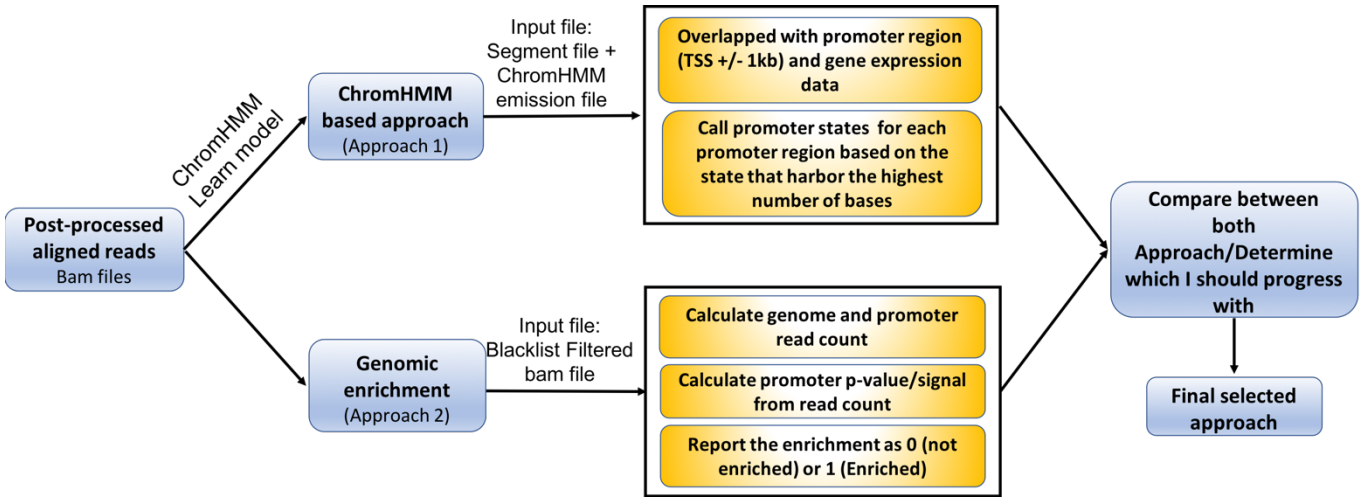


Figure 4-2: Schematic workflow of the proposed promoter calling status approaches. Two approaches have been developed. Approach 1 was developed based on ChromHMM output which was adapted via a bespoke java programming to assign a call for each promoter region. Approach 2 was developed by scoring the enrichment of signal as 0 (not enriched) or as 1 (enriched) in defined regions compared to the background across the same sized windows. Both approaches were compared to determine which approach I should proceed with for my cell lines and patient datasets.

4.2 Results

4.2.1 Identification of datasets for optimizing promoter status calling approaches

As described in **Chapter 2, Sections 2.2-8.1**, I developed the ChIP-seq pipeline and the promoter calling approach using an external dataset, published in Liau et al, as the work discussed in this paper is very central to my project. This study investigated the changes in H3K27me3 and H3K4me3 between GSC8 (an untreated patient-derived GBM stem cell line) and GSC8per (which are GSC8 cells that persist following a long-term treatment). These data are, therefore, relevant for the pipelines I need to develop but also enable me to analyse biologically relevant samples. I will be able to both compare the results obtained with those from the paper final results but also incorporate or consider the findings when analysing my own samples.

In addition, DNA from a fresh frozen pair of primary and matched recurrent GBM samples in our lab underwent ChIPseq (performed externally at Active Motif, Inc) to assess EZH2, H3K27me3 and H3K4me3 in both samples and JARID2, though unfortunately this was only successful in the recurrence. These in-house ChIPseq data were also used to aid the development and optimisation of the ChIPseq data pipeline. None of these datasets have any biological replicates.

4.2.2 ChIP-seq data pre-processing and read mapping

The pipeline was developed and optimized using the external dataset first. In general, more than 15 million 38bp single-end ChIP-seq reads were reported for this dataset except for one sample which has less than 10 million (**see Table 4.1**). The initial step in the proposed ChIP-seq pipeline was the quality evaluation of raw sequencing reads using FastQC. The program provided a simple checkpoint for the quality of the obtained data. **Table 4.1** summarizes the main statistics of each analysed sample. The key quality metrices in the generated report including per sequence quality score, total number of sequences processed, sequence duplication levels and adapter content.

Sample name	Sample description	Mean Sequence quality (phred score)	Total sequence	GC content
SRR4420628	Input_GSC8per	35	30206748	40%
SRR4420631	Input_GSC8	32	21983479	38%
SRR4420639	H3K4me3_GSC8per	35	19437096	50%

SRR4420644	H3K4me3_GSC8	33	5142695	50%
SRR4420649	H3K27me3_GSC8per	35	21435974	41%
SRR4420654	H3K27me3_GSC8	32	16932524	40%

Table 4-1: Main quality metrices for the external dataset from FastQC program.
Table includes the sample name of the external dataset, their description, mean quality score of the sequence, total number of sequence and the percentage of GC content across the reads

The quality scores at each position for all reads were high with a median quality score above 30. This indicated that the likelihood of incorrect base call is 1 in 1000 and the base call accuracy is 99.9%. A better base call is usually associated with higher score. In general, the quality of the calls is divided into three categories: reads with very good quality score (> 28), reads with reasonable score (between 20 and 28) and reads with poor quality score (< 20). The proportion for each of the four nucleotides was relatively constant across all reads and this explains the absence or lower existence of overrepresented sequence. The GC composition pattern showed a slight deviation from the theoretical along the read length. No adapter content and overrepresented sequences were found in these samples; therefore, I proceed directly to the mapping step. The main alignment statistics are listed in **Table 4.2**. Almost all reads of all samples were mapped properly with an average mapping percentage of 99.9% with a uniquely mapped read of 93%. These values indicating high mapping efficiency. Mapped reads were processed further using Picard tools to remove unmapped and duplicated reads.

Sample name	Total number of mapped reads	Alignment percentage	Uniquely mapped reads (%)
Input_GSC8per	30205279	100%	93.8%
Input_GSC8	21981769	99.99%	93.7%
H3K4me3_GSC8per	19436522	100%	93.7%
H3K4me3_GSC8	5142495	100%	94.0%
H3K27me3_GSC8per	21434808	99.99%	93.3%
H3K27me3_GSC8	16931104	99.99%	93.3%

Table 4-2: Mapping statistics of the external datasets.
Table includes the total number of mapped reads, the alignment percentage and the percentage of uniquely mapped reads of each sample of the external dataset.

4.2.3 Assessment of library complexity and ChIP enrichment

After processing, library complexity in terms of non-redundant fraction (NRF), relative strand cross-correlation coefficient (RSC), normalized strand cross-correlation coefficient (NSC) and PCR bottleneck coefficient (PBC1 and PBC2) (**see Chapter 1, Section 1.6.1.2 for definitions**) was calculated and evaluated according to ENCODE guidelines as shown in **Table 4.3**. An NRF fraction of 0.9 was observed suggesting a high complexity of the sequencing libraries for these samples. Also, higher enrichment of ChIP fragment (i.e. NSC & RSC > 1) around the targeted sites (i.e. around H3K27me3, H3K4me3, EZH2) over the background was observed. In general, this dataset showed an ideal library complexity and this confirms the absence of overrepresented/ duplicated reads in this dataset as described above.

Sample name	Non-redundant fraction (NRF)	Complexity	Relative Strand Cross-correlation coefficient (NSC)	Normalized Strand Cross-correlation coefficient (RSC)	PBC1/PBC2	Bottlenecking level
Input_GSC8per	0.9	Ideal	2.01	1.5	0.9/87003.3	None
Input_GSC8	0.9	Ideal	2.33	2.0	0.9/63549.1	None
H3K4me3_GSC8per	0.9	Ideal	1.75	1.42	0.9/38331.6	None
H3K4me3_GSC8	0.9	Ideal	1.09	1.22	0.9/102282.7	None
H3K27me3_GSC8per	0.9	Ideal	1.88	1.63	0.9/83346.8	None
H3K27me3_GSC8	0.9	Ideal	2.08	1.38	0.9/61149.2	None

Table 4-3: Summary of library complexity and ChIP enrichment of the external dataset.
Table summarized the library complexity of each sample of the external dataset in terms of NRF, NSC, RSC, PBC1, PBC2 and PCR bottlenecking

4.2.4 Peak identification

Initial identification of enriched regions (peaks) was performed using MACS2 pairing each ChIP-seq sample with its input control. Despite the fact that I need to call the promoter status as described above, I performed peak calling because it is considering a standard step that is done with ChIP-seq data analysis after read mapping and it is widely used by the researchers to identify true areas of enrichment along the genome. Additionally, I called the peaks in order to check and compare my findings with the published results. This will help me to ensure that the bam files processing is working correctly.

Using the default or recommended parameters for broad peaks with a broad cut-off value of 0.1 as mentioned in **Chapter 2, Section 2.2-8.1**, peak models for H3K27me3 histone mark and EZH2 datasets were generated. Peak models for H3K4me3 histone mark datasets were generated using the default parameters for narrow peak (**see Chapter 2, Section 2.2-8.1 for more details**) with a q-value of 0.01. The total number of significant peaks for H3K27me3 and H3K4me3 across the whole genome is plotted in **Figure 4.3**.

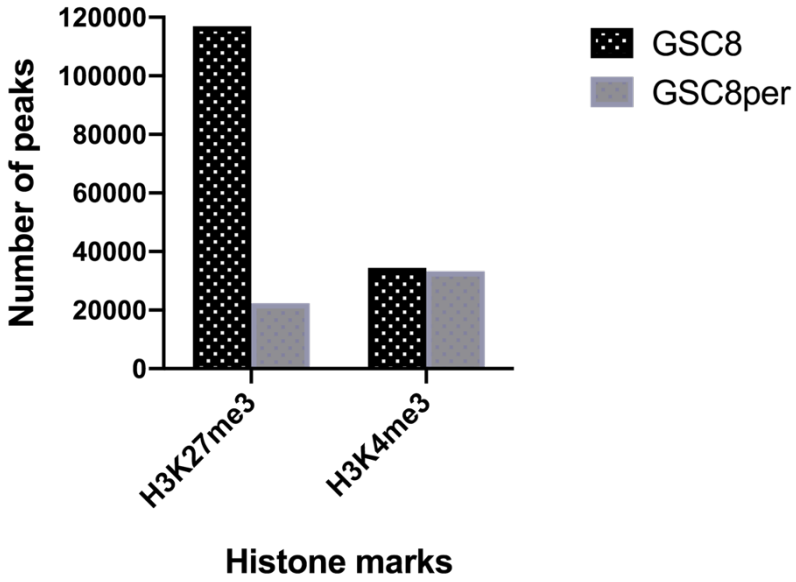


Figure 4-3: Identifications of H3K27me3 and H3K4me3 in the external dataset across the whole genome. Bar plots of the number of identified H3K27me3 and H3K4me3 peaks in GSC8 and GSC8per samples across the whole genome.

A significant reduction of H3K27me3 mark by approximately 68% was observed in GSC8per cell state in comparison with GSC8 cell state. This finding agreed with Liau et al, which identified a global loss of H3K27me3 that was suggested to cause the transition of cell from naïve to persister state. Liau et al identified a complete loss of H3K27me3 in GSC8per of *HEY1*, *FABP7* and *DLX2* genes in comparison with GSC8, whereas, much smaller peaks were observed in GSC8per for *SALL2* and *METAP1D* genes in comparison with the GSC8 state. My peak calling method also reproduced these more specific, gene-level findings (**Figure 4.4**) and this indicated that the processing of bam files is working well.

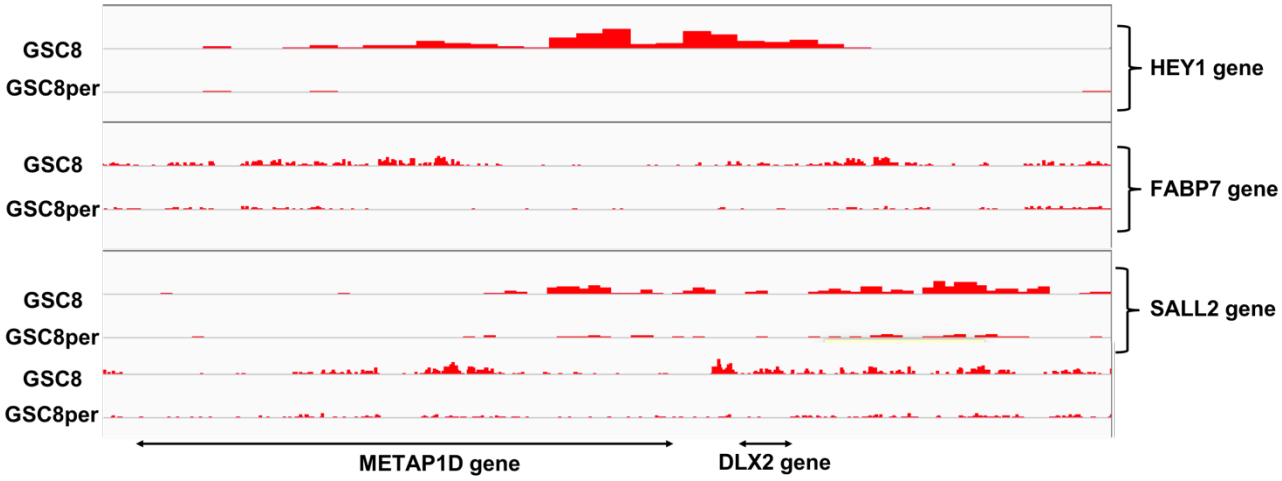


Figure 4-4: ChIP-seq profiles of H3K27me3 at genomic loci of HEY1, FABP7, SALL2, MATAP1D and DLX2 according to my data processing. Examples of a complete absence of H3K27me3 peak in HEY1, FABP7 and DLX2 genomic loci in GSC8per in comparison to GSC8. A global reduction in H3K27me3 peaks was observed in SALL2 and METAP1D genomic loci.

4.2.5 Chromatin state analysis

4.2.5.1 Chromatin state discovery (Approach 1)

To characterise JARID2 and EZH2 binding profiles and the location of specific histone marks (H3K27me3 and H3K4me3) in primary and recurrent GBM samples, I employed the most widely used tool ChromHMM. I originally developed and implemented the promoter calling status approach on the aforementioned external dataset from Liau et al.

Prior to generating the chromatin state model, the resulting post-processed bam files were converted to bed files using *bamtobed* function of Bedtools (version 2.30). All bed files were binarized using the *binarizeBed* function of ChromHMM with default parameters. By default, the interval parameter value is set to 200bp to divide the genome into segments. All input samples were used as controls to adjust the binarization locally as described in **Chapter 2, section 2.2-8.2.1**. The binarized data was used as input to train the model using the *learnModel* function of ChromHMM. The common model of chromatin states was generated by effectively concatenating multiple cell types (**see Figure 2.7 in Chapter 2, Section 2.2-8.2.1**) corresponding to two histone mark (H3K27me3 and H3k4me3) resulting in one shared model for all cell types with cell-type-specific annotations. A model with 4 functionally distinct chromatin states were generated based on the selected learn model parameters as shown in **Figure 4.5**. These states are Null (neither mark), repressed (i.e. H3K27me3 only), bivalent (i.e. both H3K27me3 and H3K4me3 exists and active (i.e. H3K4me3 only).

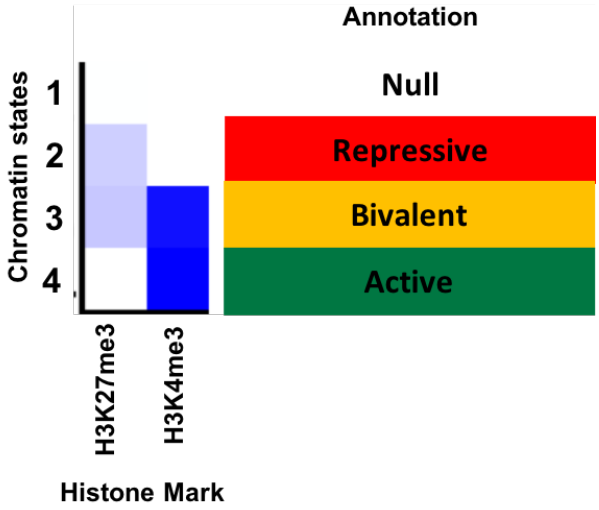


Figure 4-5: ChromHMM model based on an external dataset from Liau et al. Emission profile from a 4-State LearnModel based on the two histone modifications studied. Each state is represented by a separate row, and each mark is represented by a different column. ChromHMM identifies functionally distinct chromatin states representing null (state 1), repressive (state 2), bivalent (state 3) and active (state 4). The probability of each state is indicated by a different shade of blue; a darker shade of blue indicates a higher likelihood of seeing the mark in that condition.

The learned emission parameters were returned in a four-column segmentation file (.txt format) and a heatmap image. The segmentation file contains segments (partitions of the genome) along with corresponding emission states (i.e. E1, E2, E3, etc). The image file is the representation of the emission file in which each row is a state and each column is an input data file ("histone mark" or EZH2 or any other mark). In each state, the probability vector represents the likelihood of finding each mark in that particular state. The darker the blue colour the greater the probability of observing the mark in that state. I then annotated each candidate state based on the probability of each mark in that state. The corresponding annotation of the observed states along with their emission probability were given in Appendix E.

As we were interested in characterizing chromatin state in promoter regions, we used our pre-defined promoter regions which is +/- 1kb either side of the TSS. I used the gencode annotation file version 27 that contains comprehensive gene annotation of the reference chromosomes, scaffolds, assembly patches and alternate loci (haplotypes) to create the promoter file as described in **Chapter 2, Section 2.2-8.2**. I developed the promoter status calling approach using Java programming language in NetBeans IDE. In brief, four different source files with .java extension were created with one main method that contain some statements and information about the methods, variable and constants that I used to generate the desired output. The pre-defined promoter file, the resulting segmentation file and the emission file from *learnModel* function of ChromHMM was used as inputs to create the promoter calling file in .txt format. The resulting file was formatted to include the information of the promoter region in terms of chromosome number, start position, end position, transcript ID and gene ID. In addition, the number of overlapped bases for each state in each promoter region and the final promoter call/emission was included in the resulted file. The final promoter call was assigned based on the state that harbor the highest number of bases as shown in **Figure 4.6**. The promoter calling files produced by Approach 1, were compared with Approach 2. The latter is described below and the analysis of the comparison is in **Section 4.2.4.3**.

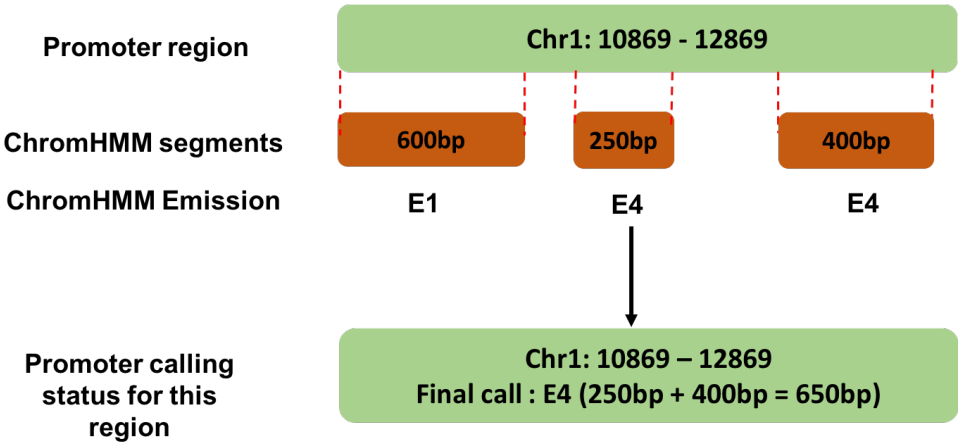


Figure 4-6: Schematic representation of promoter calling status assignment based on approach 1. The pre-defined promoter file, the resulting ChromHMM segment files and emission file were used as input to generate the promoter calling status file. The final state call was assigned based on the state that harboured the highest number of bases.

4.2.5.2 Promoter enrichment method (Approach 2)

In parallel with the above-mentioned approach, I inspected the literature to see how others tried to define promoter regions in terms of multiple histone marks, and how they score/quantify them. I identified an alternative approach in which a group defined enriched regions based on read counts against a background model (**Figure 4.7**).

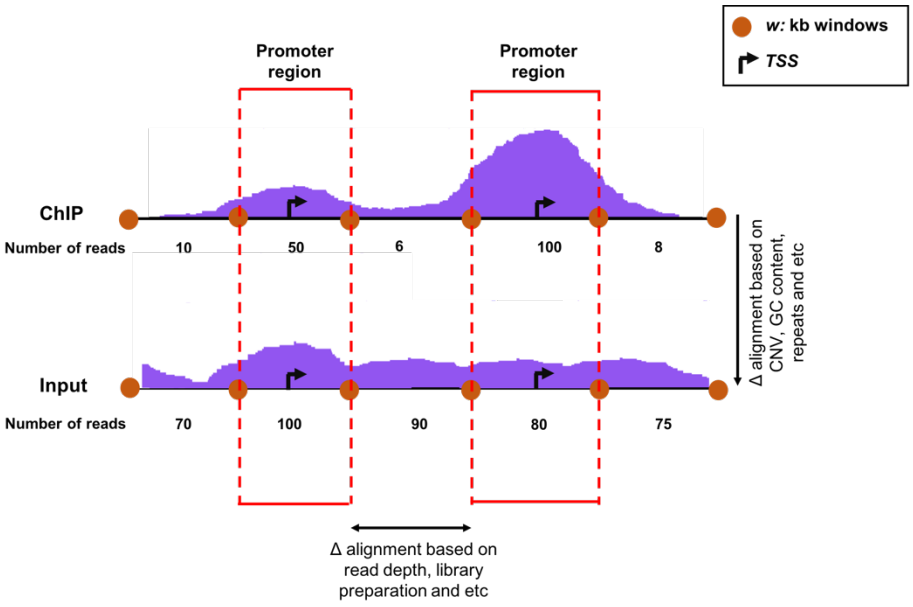


Figure 4-7: Schematic diagram of the development of promoter enrichment method (Approach 2). Enriched intervals were identified by comparing the mean fragment count in a fixed window size (distance between orange circles) against background. Read count intersecting within each fixed size window was first counted for a local enrichment of ChIP or control (i.e. input DNA, fragmented but not immunoprecipitated) signal across the genomic region and then across the promoter region (Identified within a red rectangle). For any given window, the normalized enrichment is calculated from the significance of the read count compared with the null hypothesis of no enrichment, using a Poisson test. The score given is either 0 (not significantly enriched) or 1 (significantly enriched).

Genomic regions enriched for each of the histone mark were detected as explained in **Chapter 2, Section 2.2-8.2.2** by comparing the mean fragment count of the aligned reads in a fixed size window (*w*) of 2-kb. In short, the number of reads intersecting with each fixed-size window across the whole genome were counted for a local enrichment of ChIP versus control/background signal. Under the assumption that random read alignment follows a Poisson distribution with parameter $\lambda_{ChIP,}$ calculated across the whole tiled genome first and then this value was used within significance testing of the read count in each promoter window. The calculation used the read depth within the corresponding window in the matched input control, as a factor of the average window read count in the same control, to weight $\lambda_{ChIP.}$ In our case, the p-value threshold was initially set to a default of 1×10^{-5} because this was selected within the publication I based this approach on.

This approach was developed into a programme called GBMProm in partnership with AD Bioinformatics, in parallel to me coding Approach 1, as explained above. GBMProm calculates a p-value for enrichment of each histone mark or DNA binding factor within a promoter region. The default p-value threshold of 1×10^{-5} was used to classify a binary score (i.e. the program gives a score of 0 if the adjusted p-value for each of the histone mark is higher than the selected threshold or 1 if the adjusted p-value is lower than the threshold at the promoter region). By combining binary scores from all experiments, defined in a certain order, the overall chromatin state of each promoter was ascertained. For example, the first score represents the enrichment state of H3K27me3 and the second score represent the enrichment state of H3K4me3 (and the third for EZH2, which was not studied in the external

dataset). So, '000' means zero enrichment of all marks, '100' means the enrichment of H3K27me3 only, '110' means the enrichment of both marks and '010' means the enrichment of H3K4me3 only (**see Table 4.4 below**). I then characterized these combined binary annotations to label promoters as shown in **Table 4.5** as active, repressive or bivalent based on the existence of each element (**see Section 4.2.4.3 for more detail**).

4.2.5.3 Comparing and contrasting approaches

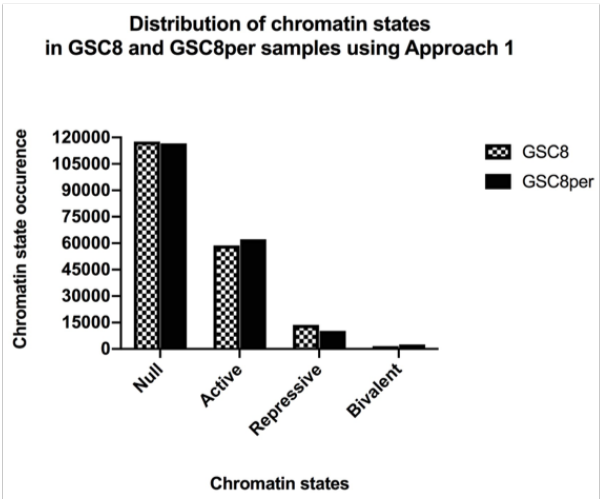
4.2.5.3.1 Quantitative comparison between emission/call occurrence across the promoter regions between ChromHMM and enriched pipeline approaches

First, I compared how each approach characterizes the observed chromatin states. Despite the fact that each approach calls the final chromatin states for each promoter region differently as described above, both approaches gave a similar characterization and description of the resulting chromatin states for each sample (**Table 4.4**). An almost similar enrichment pattern of each state was observed for both approaches (**Figure 4.8 a&b**).

State (emission order)	ChromHMM emission	Enriched pipeline Call	Enriched pipeline state
E1	Null	000	Null
E2	Repressed	100	Repressed
E3	Bivalent	110	Bivalent
E4	Active	010	Active

Table 4-4: Chromatin states calls from Approach 1 and Approach 2 for the external dataset. Table includes ChromHMM emission, enriched pipeline call and state

a)



b)

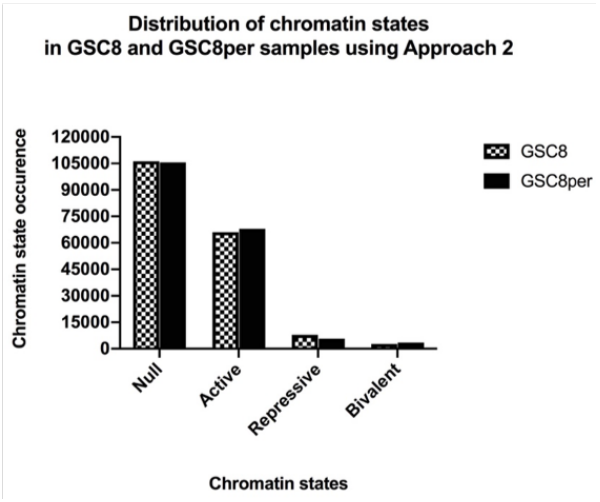


Figure 4-8: Comparison of the characterized chromatin states between Approach 1 (a) and Approach 2 (b) in GSC8 and GSC8per samples. Bar plots shows the occurrence of each state across the promoter regions. The X-axes represent the chromatin states, while the Y-axes represent the number of chromatin states occurrence in the GCS8 and GCS8per samples.

4.2.5.3.2 Integration of RNA-seq data to optimize promoter calling parameters

To determine which approach was giving the most accurate promoter call, I integrated RNA-seq data with the output of each approach. The advantage of this integration is to see which approach best links changes in the chromatin states with the observed changes in the gene expression, based on what we know of the activating or repressive roles of these marks on gene expression. The integration was performed by intersecting RNA-seq-derived expression data for the samples in question with the promoter calling file from Approach 1 and Approach 2, respectively. The output files included, the promoter region, the final chromatin calls for both the pre- and post-treatment sample, the associated gene expression data in terms of FPKM and the change in gene expression through treatment as $\log_2$FC.

First, using the external dataset, I selected several promoter regions, at random, where there was a discrepancy between the promoter call from the two approaches. I loaded the bigwig

files from bam files for each histone mark in IGV to visually assess the enrichment of each mark at the selected region in comparison with the input (i.e. control sample with no mark) (**Figure 4.9 a-d**). Also, I loaded the segment file from ChromHMM in IGV to visualize the state segmentations across the selected promoter region. Then, I compared the final promoter status call from both approaches based on their algorithms (**See sections 4.2.5.1 and 4.2.5.2 for explanations of the algorithms**). Next, I integrated RNA-seq data to assess the correlation of the final call with gene expression (**Table 4.5**).

Promoter region	Final Approach 1 promoter call	Approach 2 promoter call	Notes for ChromHMM high resolution	Notes from IGV and RNA-seq data
Cell state: primary cell (GSC8)				
chr2:164621848	E2: Repressed	110 – Bivalent	E1:152 E2:1048 E3:800	The call from approach 1 is inconsistent with IGV result. It agrees that the signal is present but not the majority. The call from approach 2 is consistent with IGV result (**Figure 4.9 a**).
chr12:132610543	E2: Repressed	010- Active	E2:1400 E3:600	The call from approach 1 is inconsistent with IGV result. It agrees that the signal is present but not the majority. The call from approach 2 is consistent with IGV result due to the presence of high signal of H3K4me3 (**Figure 4.9 b**).
chr1:10694479	E3- Bivalent	010- Active	E2:879 E3: 1121	The call from approach 1 is inconsistent with IGV result. It agrees that the signal is present but not the majority. The call

				from approach 2 is consistent with IGV result due to the presence of high signal of H3K4me3.
chr8:22579646	E2- Repressed	110-Bivalent	E2:1200 E3:800	The calls from both approaches are inconsistent with IGV result. There is a clear signal of H3K4me3.
chr8:23084403	E2: repressed	000- Null	E2:2000	The calls from both approaches are inconsistent with IGV result. IGV analysis indicated the presence of H3K4me3 signal and the gene expression revealed higher gene expression for this cell state.
Cell state: Persistent cell (GSC8per)				
chr7:151814840	E1: Null	010- Active	E1: 1400 E4: 600	The call from approach 1 is inconsistent with IGV result. It agrees that the signal is present but not the majority. The call from approach 2 is consistent with IGV result as there is a clear signal of H3K4me3 (**Figure 4.9 c**). However, gene expression analysis indicated a reduced expression of genes as it goes from GSC8 to GSC8per.
chr1:925738	E2: Repressed	110: Bivalent	E2: 1138 E3: 862	The calls from both approaches are inconsistent with IGV result as it supposed to be

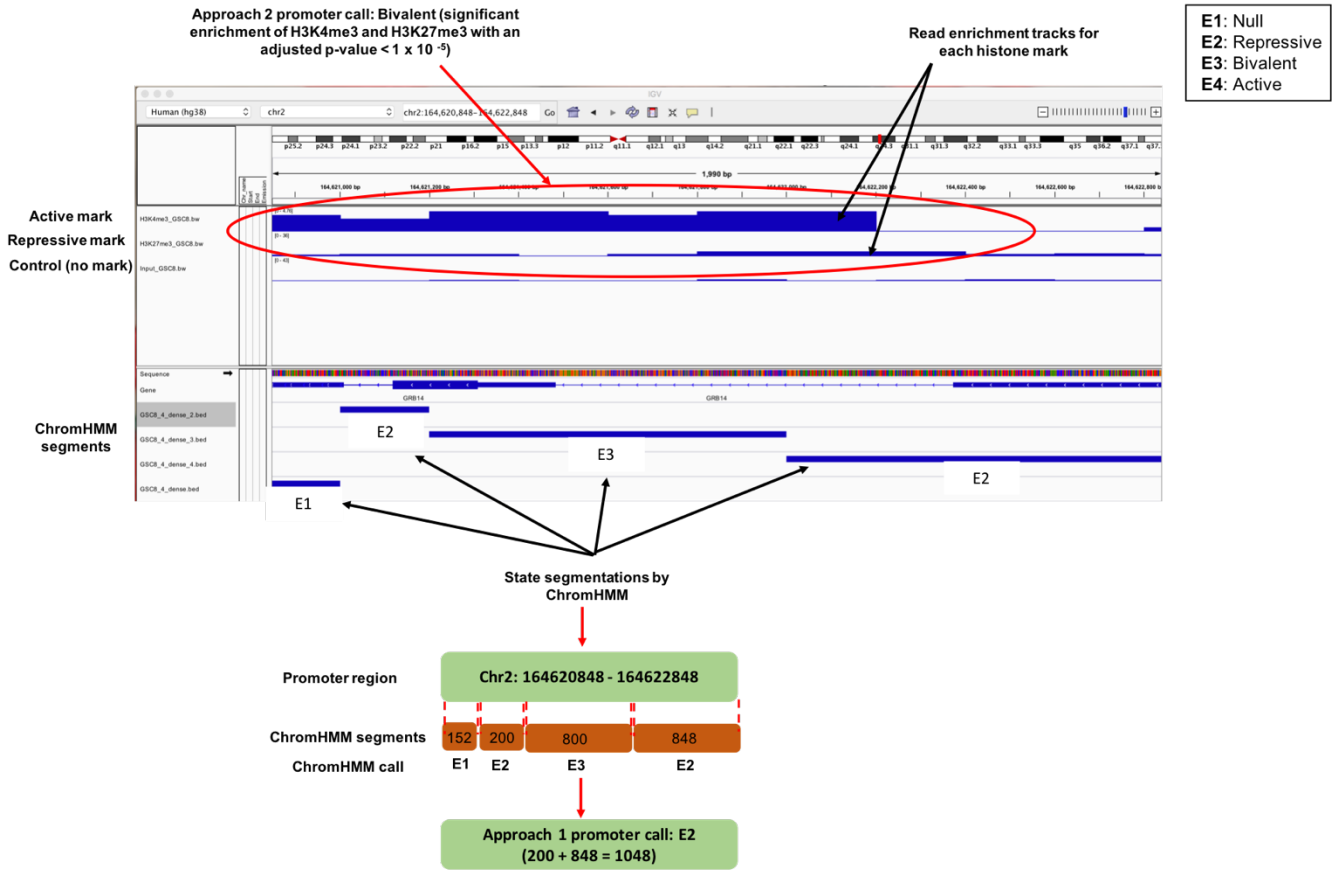
				called active based on IGV analysis (**Figure 4.9 d**). There is a clear signal for H3K4me3.I tried to correlate this with gene expression but the results indicated a reduced expression of gene as it goes from GSC8 to GSC8per.
chr2:164621848	E4: Active	110- Bivalent	E1: 648 E4: 1352	The call from approach 1 is inconsistent with IGV result. It agrees that the signal is present but not the majority. The call from approach 2 is consistent with IGV result as there is a clear signal for both marks (**Figure 4.9 f**).
chr12:132610543	E3: Bivalent	110- Bivalent	E1: 543 E2: 257 E3: 1000 E4: 200	Both approaches called this promoter as bivalent which is consistent with IGV result. However, the signal from H3K4me3 is apparent and the bivalency is just starting. This finding is nicely correlated with gene expression analysis which showed an increase in the signal due the high enrichment of H3K4me3.
chr13:20192898	E3: Bivalent	010- Active	E2: 600 E3: 1400	The call from approach 1 is inconsistent with IGV result. The call from approach 2 is

				consistent with IGV result as there is a clear signal for H3K4me3 mark. Gene expression analysis showed an increase in the gene expression in this cell state.

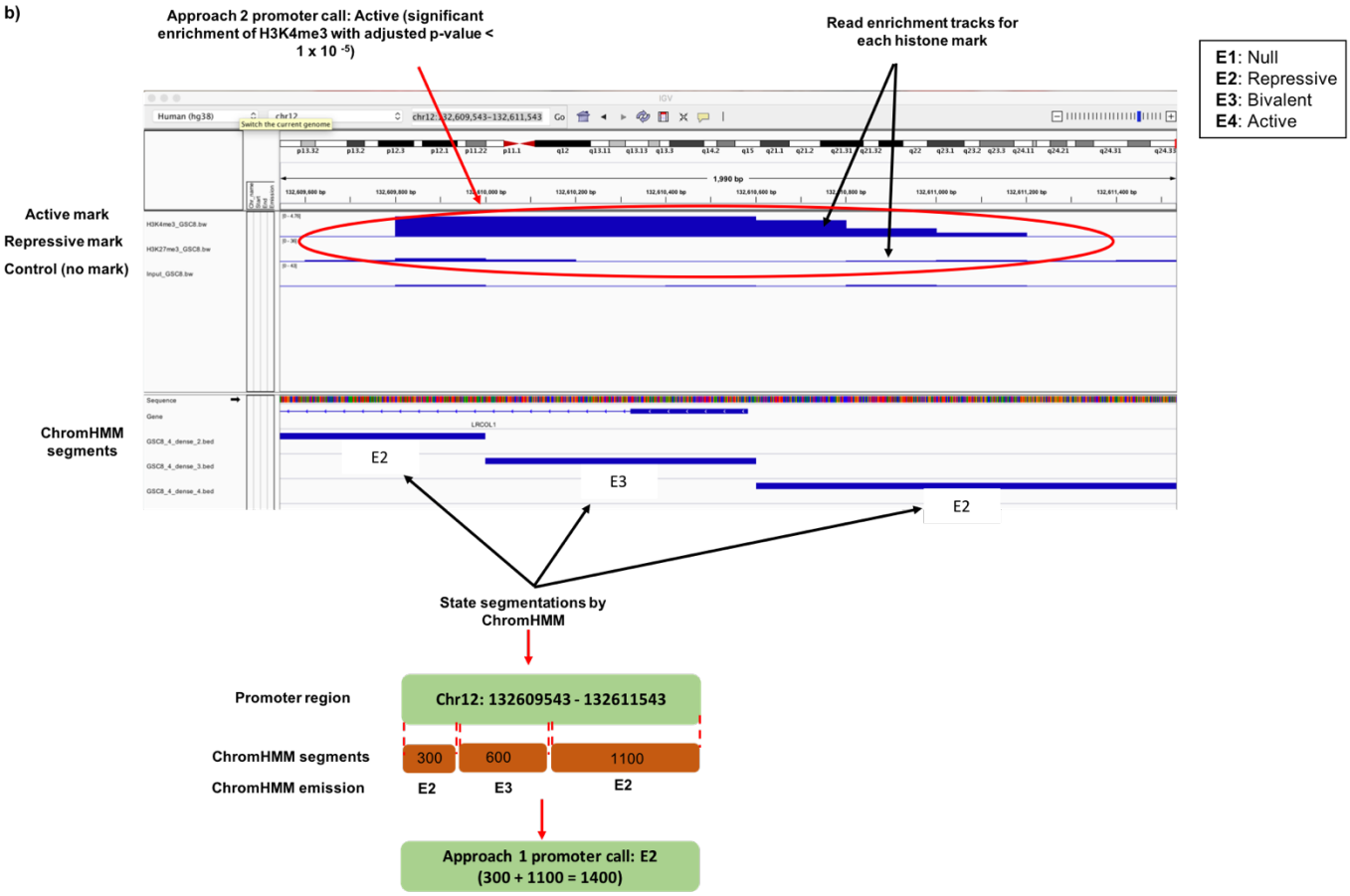
Table 4-5: IGV results of the called chromatin states using Approach 1 and Approach 2 for the external dataset.

Table includes the promoter region, the final call from Approach 1 and approach 2, notes from ChromHMM high resolution and notes from IGV and RNA-seq data for the external dataset

a)



b)


Approach 2 promoter call: Active (significant enrichment of H3K4me3 with adjusted p-value < 1 x 10 -5)
Read enrichment tracks for each histone mark
E1: Null
E2: Repressive
E3: Bivalent
E4: Active
Active mark
Repressive mark
Control (no mark)
ChromHMM segments
E2
E3
E2
State segmentations by ChromHMM
Promoter region
Chr12: 132609543 - 132611543
ChromHMM segments
ChromHMM emission
300
600
1100
E2
E3
E2
Approach 1 promoter call: E2
(300 + 1100 = 1400)

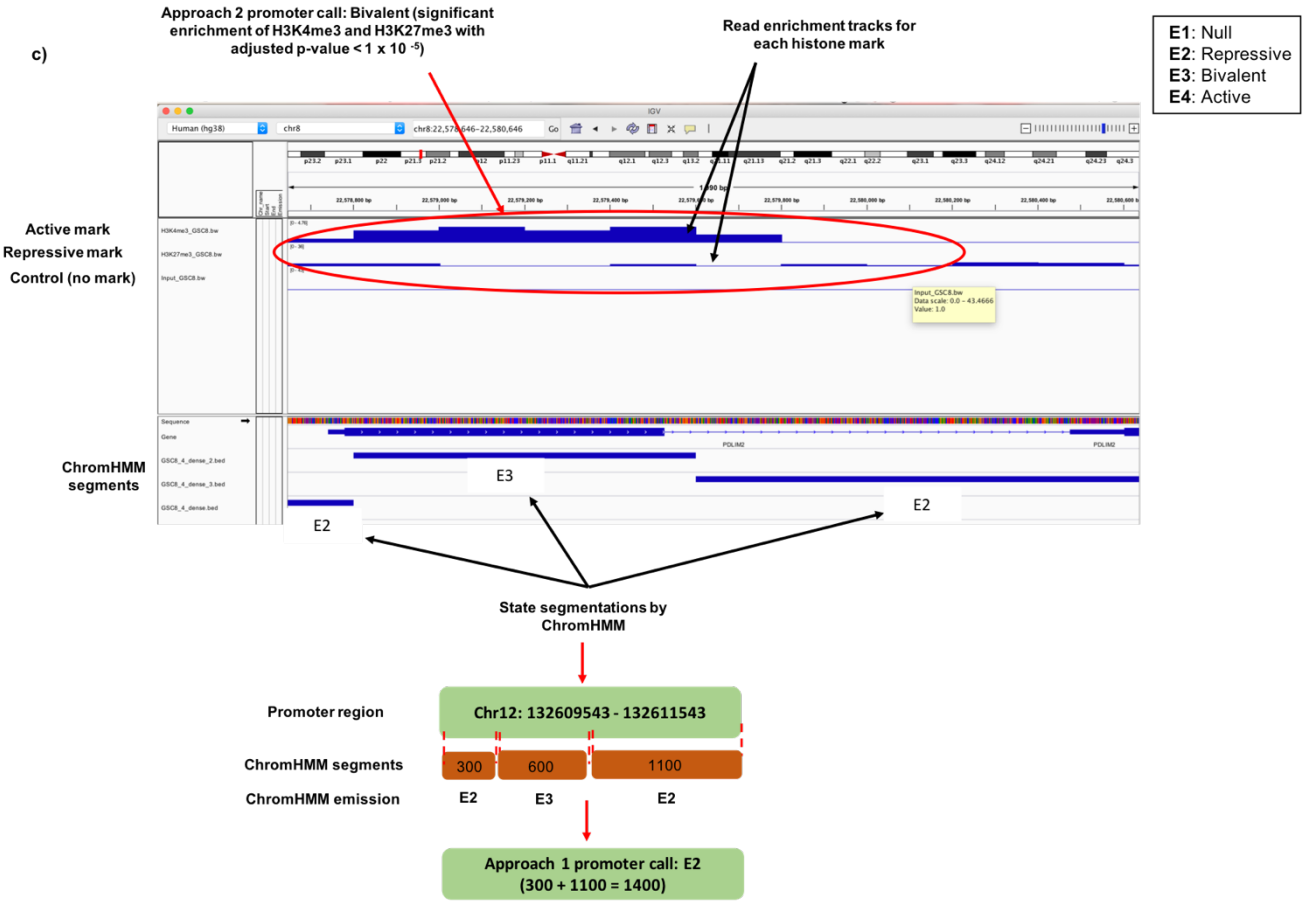


Figure 4.9: Integrative genomic viewer (IGV) browser tracks of H3K4me3, H3K27me3 aligned to a human reference genome in comparison to the input (i.e. control) sample and the state segmentations from ChromHMM. The upper blue tracks represent read enrichments of H3K4me3 and H3K27me3 in comparison to the input sample across three different promoter regions (**a,b** and **c**). The lower tracks at the bottom of the image represent the state segments resulted from ChromHMM. The tracks are aligned to a human reference genome (hg38). E1, E2, E3 and E4 represent ChromHMM emissions for each segment. The annotation of each emission is determined based of the probability of observing each mark in that state (**See figure 4.5 and table 4.4 for emission annotations**).

Based on the above results, there is a clear disagreement between both approaches at the selected promoters; however, Approach 2 seems to best annotate the promoters in comparison to Approach 1. This might be due to the fact that Approach 1 picks up signals in that region, but the choice of 'most prevalent state' across the promoter is removing that signal (**Figure 4.5**). For example, in **figure 4.9a** and **table 4.5**, ChromHMM indicates that H3K4me3 and H3K27me3 signals are present at that promoter region, but because

ChromHMM is sub-sectioning the promoter region into separate segments with its corresponding state, it was difficult to assign one final call for that promoter region using the information given by ChromHMM. The developed approach (i.e. Approach 1) was designed to call the promoter status in our regions of interest (i.e. +/- 1kb either side of the TSS) and the final call was assigned based on the state that harbors the highest number of bases, as shown in **Figure 4.6.** However, based on the given examples (**Figure 4.9 a-c**), the assignment of the final call based on the 'most prevalent state' is neglecting the signal of the other mark.

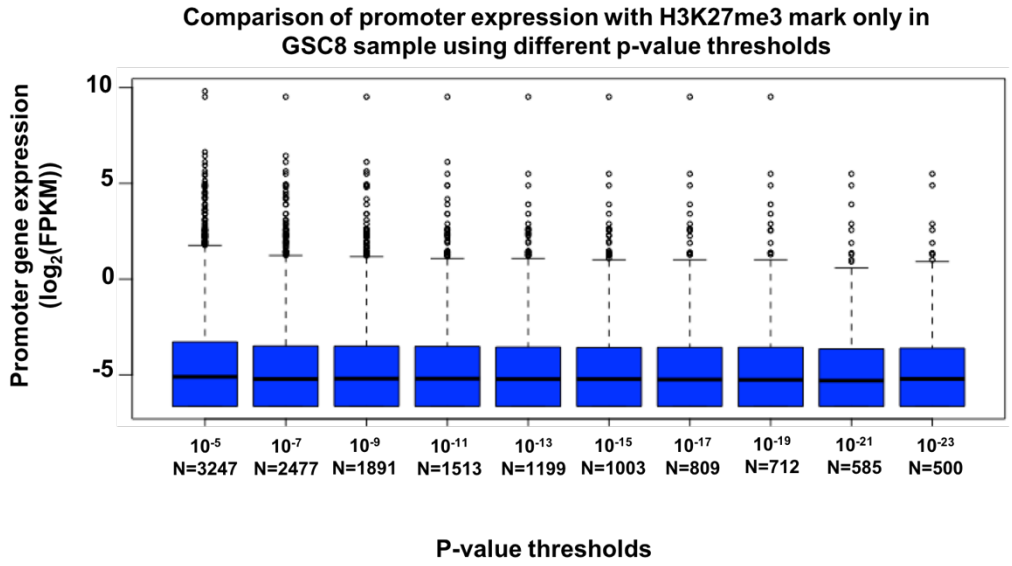
In general, the quantitative analysis, the assessment of chromatin transition and the correlation of the chromatin states with its gene expression indicated that Approach 2 is more accurate than Approach 1.

4.2.5.3.3 Optimization of p-value threshold

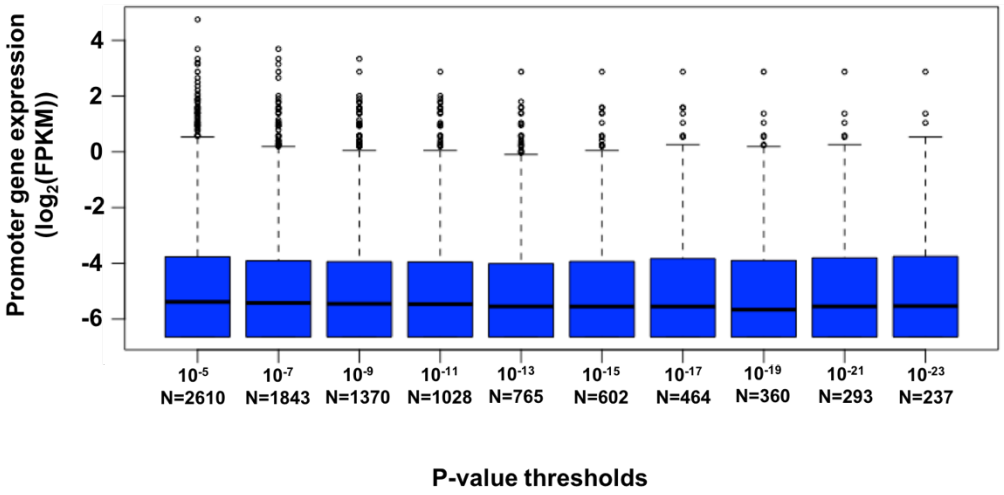
Despite the fact that Approach 2 calls the promoter status accurately, it still required further optimization by increasing the stringency. This was achieved by changing the p-value threshold at which significance is reached and a promoter is scored as containing the mark (1) versus not doing (0). Promoters in the GSC8 samples for which the H3K4me3 p-value is > 0.05 (i.e. that do not contain the active mark) were filtered. Then, I created boxplots of the gene expression of promoters with H3K27me3 at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$ as shown in **Figure 4.10a**. I counted the number of the promoters marked as having the H3K27me3 mark at each p-value threshold to see how increased stringency of calling affected scored as '1'. I quantified the percentage of the reduction in these promoters as the p-value threshold decreased from $1x10^{-5}$ to $1x10^{-23}$. The idea is to select a suitable p-value where we see a plateau in gene expression for what are considered 'repressed' promoters. Changes in the gene expression were observed to plateau for promoters with H3K27me3 mark at a p-value of $1x10^{-15}$. I repeated the above approach for the GSC8per samples (**Figure 4.10b**) and a change in gene expression was observed at $1x10^{-13}$ and $1x10^{-15}$. Next, I applied the above approach for those promoters in GSC8 that do not contain the repressive mark (i.e. H3K27me3 p-value is > 0.05) and I plotted the gene expression of promoters with H3K4me3 at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$ (**Figure 4.10c**). Similar steps were applied for H3K4me3 promoters in GSC8per samples and I plotted their gene expression (**Figure 4.10d**). No changes were observed at any selected p-values in the promoters with H3K4me3

mark in both cell states. In view of the above results, I decided to select a p-value of $1x10^{-15}$ for further downstream analysis for both marks to ensure ease of use of the programme.

a)



b)



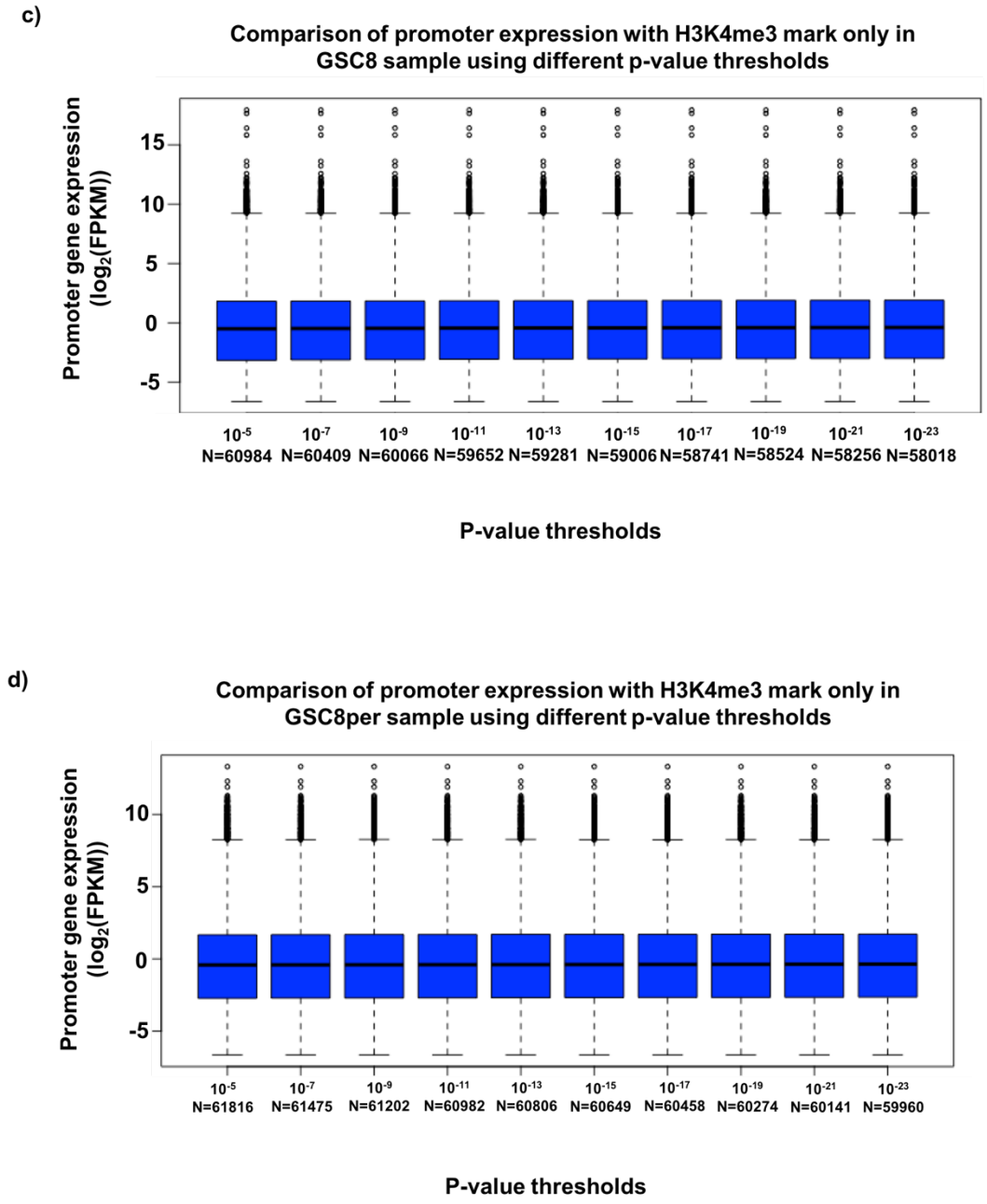


Figure 4-10: The box plots of log2-transformed gene expression of promoters with each histone mark in GSC8 and GSC8per samples using different p-value thresholds. (a) Box plots showing the gene expression of GSC8 promoters with H3K27me3 mark only and do not contain H3K4me3 mark (i.e. H3K4me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. **(b)** Box plots showing the gene expression of GSC8per promoters with H3K27me3 mark only and do not contain H3K4me3 mark (i.e. H3K4me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. **(c)** Box plots showing the gene expression of GSC8 promoters with H3K4me3 mark only and do not contain H3K27me3 mark (i.e. H3K27me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. **(d)** Box

plots showing the gene expression of GSC8per promoters with H3K4me3 mark only and do not contain H3K27me3 mark (i.e. H3K27me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. The X-axes represent the p-value thresholds, while the Y-axes represent the log2-transformed gene expression of the promoters in GSC8 and GSC8per samples. The horizontal lines in each column of the plot represents the mean gene expression value.

In order to ensure that the selection of this p-value is suitable for the analysis, I decided to assess the enrichment of H3K27me3 in GSC8 naïve and GSC8per using it, in comparison to some of the specific findings in the published paper. This was to check that my bespoke method would yield those same results. It was stated in the paper that Notch 1 intracellular domain-associated genes (N1ICD) that have the H3K27me3 mark in GCS8 tend to lose it in GSC8per. I examined these genes and I found either a reduction (**Figure 4.11 a**) or complete loss (**Figure 4.11 b**) of H3K27me3 signal in GSC8per in comparison to GSC8 naïve. The gene expression was studied and I found clear changes in the gene expression in GSC8 naïve and GSC8per due to the loss of H3K27me3 mark (**Figure 4.12**). The reduction of H3K27me3 causes an increase in the gene expression of NTRK2 and SGK4. These findings suggested that Approach 2 along with the optimized p-value of $1x10^{-15}$ is generating comparable results to those in the paper.

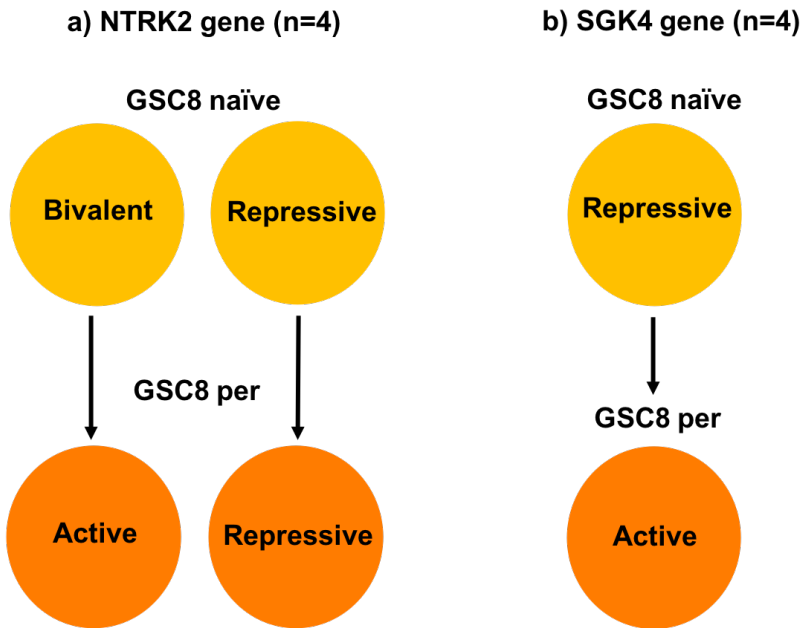


Figure 4-11: Distribution of H3K27me3 mark in N1ICD-associated loci between GSC8 naïve and GSC8per. Chromatin state transition of H3K27me3 in the promoter regions of NTRK2 (a) and SGK4 (b) genes between GSC8 naïve and GSC8per. SGK4 gene has only one reported state transition across its promoters which is repressive to bivalent. Whereas, NTRK2 gene has two reported state changes: two promoters with bivalent-active transition and two promoters with repressive-repressive transition. n refers to the number of promoters for each gene.

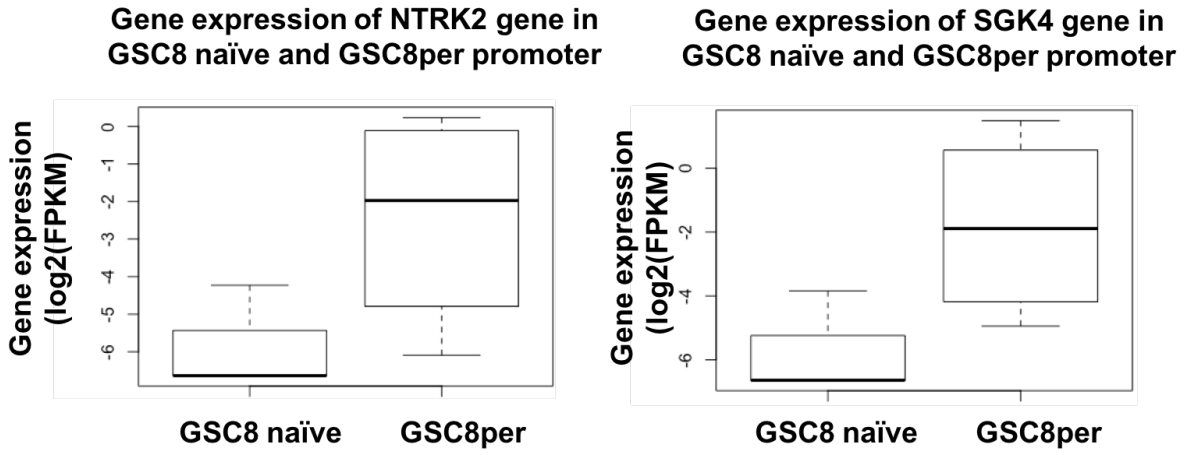


Figure 4-12: Gene expression analysis of N1ICD-associated loci between GSC8 naïve and GSC8per shows the effect of a reduction in H3K27me3 signal. Increases in the expression of gene promoters in NTRK2 (a), SGK4 (b) due to the global reduction in H3K27me3 enrichments in GSC8per in comparison to GSC8 naïve. A decrease in the expression of gene promoter in DPF3 (D) was also observed. The X-axes represent the cell types, while the Y-axes represent the log2-transformed gene expression of the promoters in GSC8 and GSC8per samples. The horizontal lines in each column of the plot represents the mean gene expression value.

4.3 Implementation and optimization of the developed ChIP-seq pipeline and promoter calling status approaches on an in-house ChIPseq dataset

An in-house dataset (**see Section 4.2.1**) was also used to develop and implement ChIP-seq pipeline and optimize the promoter calling approaches. This is useful to ensure the reproducibility of the results and that the developed pipeline is designed to be applicable for different datasets. I applied the above described steps on this dataset typically from quality assessment of raw reads to promoter calling approach comparison.

More than 30 million 75bp single-end ChIP-seq reads were collected from high throughput sequencing for all samples except the JARID2 experiment (**Table 4.6**). The quality of each sample was evaluated: the mean quality score per read was > 35 in all cases. Quality assessment indicated the existence of overrepresented sequence and adapters. The GC composition is more than 40% and this amount of GC content in the library causes a deviation in the distribution of GC content from the normal distribution level. In our case, the sum of deviations from the normal distribution is between 15% and 30%.

Sample name	Sample description	Mean Sequence quality (phred score)	Total sequence	GC content
EZH2_P	EZH2-ChIP of primary fresh frozen GBM tissue	35	37947592	46%
EZH2_R	EZH2-ChIP of recurrent fresh frozen GBM tissue	35	37334507	42%
H3K4me3_P	H3K4me3-ChIP of primary fresh frozen GBM tissue	35	37423744	61%
H3K4me3_R	H3K4me3-ChIP of recurrent fresh frozen GBM tissue	35	38895833	63%
H3K27me3_P	H3K27me3-ChIP of primary fresh frozen GBM tissue	35	47456449	53%
H3K27me3_R	H3K27me3-ChIP of recurrent fresh frozen GBM tissue	35	35582578	51%
JARID2_R	JARID2-ChIP of recurrent fresh frozen GBM tissue	35	49455157	45%
Input_P	Input-ChIP of primary fresh frozen GBM tissue	35	34625927	40%
Input_R	Input-ChIP of recurrent fresh frozen GBM tissue	35	31639635	43%

Table 4-6: Main quality metrices of the in-house dataset from FastQC program.
Table summarized the quality metrices of an in-house dataset in terms of mean quality score of the reads, total number of sequence and the percentage of GC content across the reads

Due to the presence of adapter content and overrepresented sequences, I trimmed these sequences and I filtered low quality reads using Cutadapt. Trimmed reads were re-assessed again using FastQC as described above and these trimmed reads were mapped to the human reference genome. The main alignment statistics was obtained (**Table 4.7**) and the results suggested that the reads for each sample were mapped properly with a uniquely mapping percentage of more than 80%.

Sample name	Total number of mapped reads	Alignment percentage	Uniquely mapped reads
EZH2_P	3040728	82%	87%
EZH2_R	35067574	94%	88%
H3K4me3_P	35483357	96%	80%
H3K4me3_R	34957858	96%	80%
H3K27me3_P	41776595	91%	90%
H3K27me3_R	33188331	94%	92%
JARID2_R	47543358	96%	88%
Input_P	33123148	96%	87%
Input_R	30607299	97%	87%

Table 4-7: Mapping statistics of the in-house datasets.
Table summarizes the mapping statistics of each sample of in-house dataset in terms of total number of mapped reads, the alignment percentages and the percentages of uniquely mapped reads

Then, the library complexity was evaluated and an NRF value between 0.6 and 1 (**Table 4.8**) was observed for all samples except the H3K27me3-ChIP of the primary sample which has a lower library complexity (NRF < 0.4). The classification of library complexity was based on ENCODE guidelines in which the complexity is called ideal if an NRF value is > 0.9, acceptable if it is between 0.7 and 0.9 and concerning if it is < 0.7.

Sample name	Non-redundant fraction (NRF)	Complexity	Relative Strand Cross-correlation coefficient (NSC)	Normalized Strand Cross-correlation coefficient (RSC)	PBC1/PBC2	Bottlenecking level
EZH2_P	0.8	Compliant	1.05	3.61	0.8/4.1	Mild
EZH2_R	0.7	Acceptable	1.05	4.57	0.7/3.5	Mild
H3K4me3_P	0.7	Acceptable	2.03	1.17	0.7/3.6	Moderate
H3K4me3_R	0.6	Acceptable	2.34	1.20	0.6/2.8	Moderate
H3K27me3_P	0.4	Concerning	1.38	1.31	0.4/1.4	Moderate
H3K27me3_R	0.6	Acceptable	1.23	1.51	0.6/2.2	Moderate
JARID2_R	0.7	Acceptable	1.02	1.5	0.7/2.5	Moderate
Input_P	1.0	Ideal	1.01	1.01	1.0/30.3	None
Input_R	1.0	Ideal	1.01	1.01	1.0/33.4	None

Table 4-8: Summary of library complexity and ChIP enrichment of the in-house dataset.
Table summarizes the library complexity of each sample of in-house dataset in terms of NRF, library complexity, NSC, RSC, PBC1, PBC2 and PCR bottlenecking level

The ENCODE consortium endorses an NRF of > 0.8 for 10 million reads. Therefore, I downsampled the read depth to 10 million for this dataset and I reassessed the library complexity again. An NRF of > 0.9 was observed for all samples (**See Table 4.9**). As shown in **Table 4.8**, higher enrichment of ChIP fragment (i.e. NSC & RSC > 1) around the targeted sites (i.e. around H3K27me3, H3K4me3, EZH2 and JARID2) over the background was observed. In general, this dataset showed an acceptable library complexity.

Sample name	**NRF_subsample to 10 million reads**	**Complexity**
EZH2_P	0.9	Ideal
EZH2_R	0.9	Ideal
H3K4me3_P	0.9	Ideal
H3K4me3_R	0.8	Compliant
H3K27me3_P	0.8	Compliant
H3K27me3_R	0.8	Compliant
JARID2_R	0.8	Compliant
Input_P	1.0	Ideal
Input_R	1.0	Ideal

Table 4-9: Library complexity of the in-house dataset after down-sampling the reads to 10 million.
Table includes the NRF value of subsampled samples and the library complexity for each sample of an in-house dataset.

Then, I assessed the applicability of Approach 1 on this dataset. A model with 8 chromatin states was characterized by ChromHMM as shown in **Figure 4.13**. These states are in order null (neither mark), weakly repressed (i.e. weak H3K27me3 signal), repressed (i.e. H3K27me3 signal only), bivalent (i.e. both H3K27me3 and H3K4me3 signal exists) with higher signal of H3K27me3, bivalent with higher signal of H3K4me3, active (i.e. H3K4me3 only), bivalent with higher signal of H3K4me3 along with EZH2 signal, and weak bivalent along with EZH2 signal.

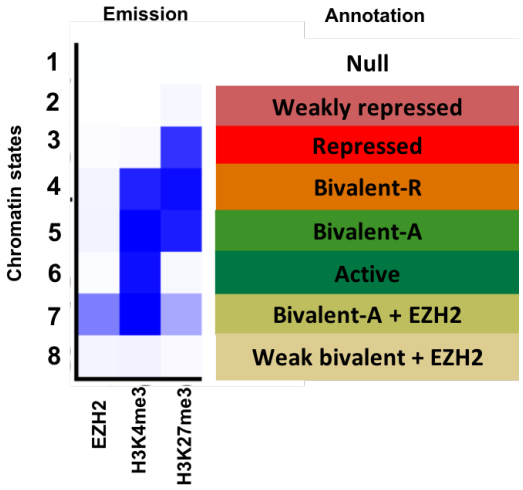


Figure 4-13: ChromHMM model based on an in-house dataset. Emission profile from an 8-State LearnModel based on the two histone modifications and EZH2. Each row corresponds to a different state, and each column corresponds to a different mark. ChromHMM identifies functionally distinct chromatin states representing null (state 1), weakly repressed (state 2), repressed (state 3), bivalent-R (both marks are present but with higher probability of the repressive mark, state 4), bivalent-A (both marks are present but with higher probability of the active mark, state 5), active (state 6), bivalent-A + EZH2 (state7, both marks are present but with higher probability of active mark in addition with EZH2), weak bivalent + EZH2 (state 8, both marks are present but both occur with lower probabilities in addition to EZH2). The probability of each state is represented by blue colour, the darker blue colour corresponds to a greater probability of observing the mark in the state.

The resulting segment file from ChromHMM was used to generate the final promoter calling file as described above. In addition, I applied Approach 2 to call the final promoter status and this was compared with the resulting file from Approach 1 to further determine and confirm

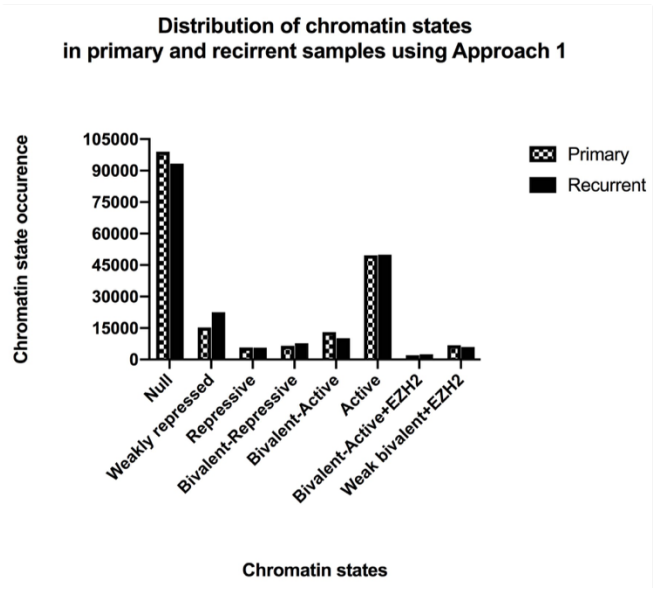
which I should progress with. The characterization of the resulting promoter states was determined as shown in **Table 4.10**. A noticeable difference in the characterization of the promoter states between the two approaches were observed. The biggest difference seems to be in the loss of EZH2 signal using Approach 1.

State (emission order)	ChromHMM emission	Enriched pipeline Call	Enriched pipeline state
E1	Null	000	Null
E2	Weakly repressed	010	Active
E3	repressed	110	Bivalent
E4	Bivalent-Repressive	111	Bivalent+EZH2
E5	Bivalent-Active	100	Repressive
E6	Active	011	Active+EZH2
E7	Bivalent-active + EZH2	001	EZH2
E8	Weak bivalent + EZH2	101	Repressive+EZH2

Table 4-10: Chromatin states calls based on approach 1 and approach 2 for an in-house dataset.
Table includes the emission state from ChromHMM and the promoter call from enriched pipeline

In order to further compare and contrast between the two promoter calling approaches, I performed a quantitative analysis by comparing the occurrence of the chromatin states calls that were resulted from Approach 1 and Approach 2. A considerable difference in the characterization of the chromatin state between both approaches was found (**Figure 4.14 a-b**). For example, approach 2 identifies some promoters with EZH2, whereas, an absence of promoters with EZH2 was noticed in approach 1. Similarly, promoters with bivalent state were found using approach 2 but not in approach 1.

a)



b)

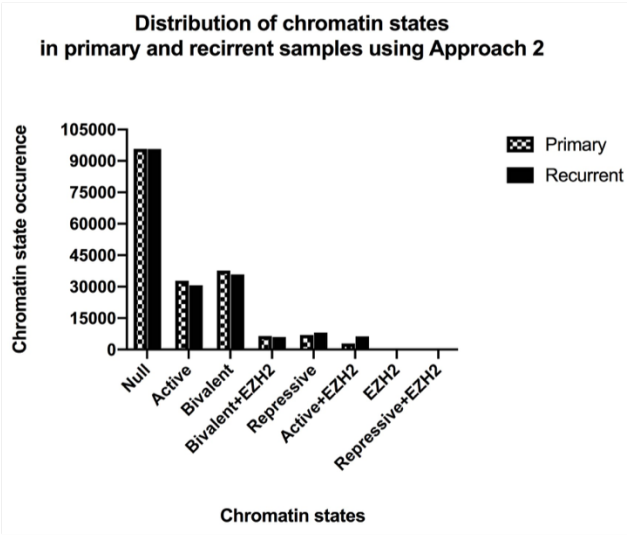


Figure 4-14: Comparison of the characterized chromatin states between approach 1 (a) and approach 2 (b) in the primary and recurrent samples of the in-house dataset. Bar plots shows the occurrence of 8 chromatin states across the promoter regions from an in-house dataset. The X-axes represent the chromatin states, while the Y-axes represent the number of chromatin states occurrence in the primary and recurrent samples.

Then, I visually assessed the chromatin state calls across randomly selected promoters as described above and I correlated the chromatin state call with its gene expression to determine which approach annotates the promoter correctly (**Table 4.11**).

Promoter region	Final approach 1 promoter call	Approach 2 promoter call	Notes for ChromHMM high resolution	Notes from IGV and RNA-seq data
Cell state: primary cell				
chr1:911435	E3: Repressed	111 – Bivalent+EZH2	E3: 1035 E4: 965	IGV analysis indicated the presence of high signal of H3K27me3 followed by H3K4me3 and EZH2 which means that this Promoter should called bivalent with the presence of EZH2.

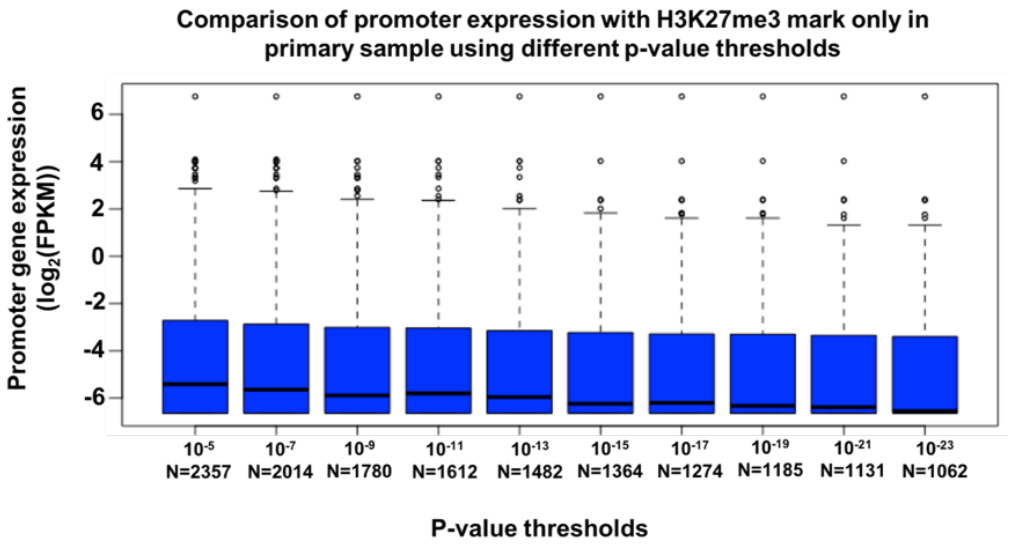
				Approach 1 called this promoter as repressed despite the fact that the bivalent state is present but the signal for this state is not the majority. Approach 2 called this promoter as bivalent with the presence of EZH2 which is correct.
chr1:19923617	E1: Null	001: EZH2		According to IGV analysis, the call from approach 1 is consistent with IGV result as there is a clear signal for EZH2. There is a loss in EZH2 signal using Approach 1.
chr8:22141902	E6- Active	110- Bivalent	E5:400 E6: 1298 E8:302	According to IGV, the call from approach 1 is inconsistent with IGV result. It agrees that the signal for both marks are present but not the majority, therefore it was called active. The call from approach 2 is consistent with IGV result due to the presence of high signal of H3K4me3 and H3K27me3.
chr19:40750448	E7- Bivalent-active + EZH2	011-Active + EZH2	E6:800 E7:1200	According to IGV, the call from approach 2 is consistent with IGV result due to the presence of clear signals of H3K4me3 and EZH2 and the absence of H3K27me3 signals. There is a clear loss of EZH2 signal using Approach 1.
chr3:51978080	E8: Weak bivalent + EZH2	010- Active	E6:320 E8: 1680	The call from approaches are inconsistent with IGV result despite the fact that there is signals for H3K4me3 and EZH2. Approach 1 called it bivalent, however, there is no signal for H3K27me3. Approach 2 called it active

				but it loses the call of EZH2 which is clearly present.
Cell state: Recurrent				
chr3:78670540	E1: Null	001- EZH2	E1: 2000	According to IGV, there is a clear signal for EZH2, therefore, the call from approach 2 is consistent with IGV result. There is a clear loss of EZH2 call using Approach 1.
chr10:80208209	E1: Null	100: Repressive	E1: 1009 E4: 200 E5: 191 E6:200 E8: 400	According to IGV, the call from approach 2 is consistent with IGV result as there is a clear signal of H3K27me3 in the recurrent sample.
chr17:42745049	E6: Active	110- Bivalent	E3: 49 E4: 400 E5:400 E6:1151	According to IGV, the call from approach 2 is consistent with IGV result due to the presence of H3K4me3 and H3K27me3 signals. Approach 1 agrees that the signal for both marks is present (i.e. bivalent) but not the majority.
chr9:127399964	E8: Repressed	010- Active	E6: 436 E8: 1564	The call from approach 2 is consistent with IGV result due to the presence of H3K4me3 signal. Approach 1 agrees the signal for H3K4me3 is there but not majority
chr5:173235206	E4: Bivalent-Repressive	111- Bivalent + EZH2	E4: 2000	Both approaches called this promoter region as bivalent, however Approach 1 indicated that there is a high signal of H3K27me3 in comparison to H3K4me3 therefore it called it bivalent + repressive. Approach 2 called it bivalent with EZH2 which is more accurate due to the presence of EZH2 signal.

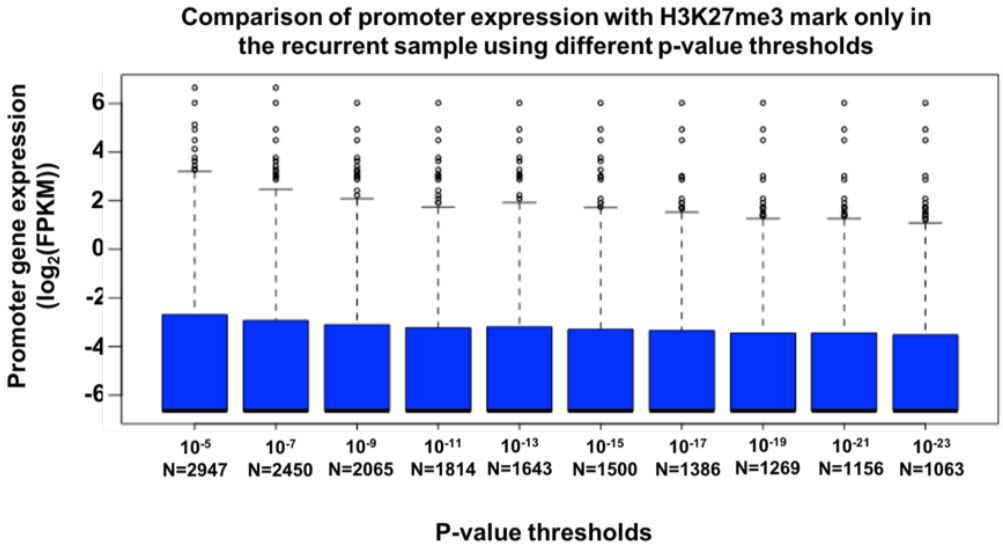
Table 4-11: IGV results of the called chromatin states using approach 1 and approach 2 for an in-house dataset.
Table includes the promoter region, the final call from Approach 1 and approach 2, notes from ChromHMM high resolution and notes from IGV and RNA-seq data for an in-house dataset

Based on the above results, there is a clear disagreement between both approaches, however, Approach 2 annotates all the promoter regions accurately with a better resolution in comparison to Approach 1. In addition, and as I mentioned above, the biggest difference seems to be in the loss of EZH2 signal using Approach 1. I further optimized Approach 2 by increasing the stringency of the p-value and I repeated the same steps as in **section 4.2.4.4** to determine the suitable p-value where we see a plateau in gene expression for what are considered either 'repressed' promoters (**Figure 4.15 a-b**) or 'active' promoters (**Figure c-d**).

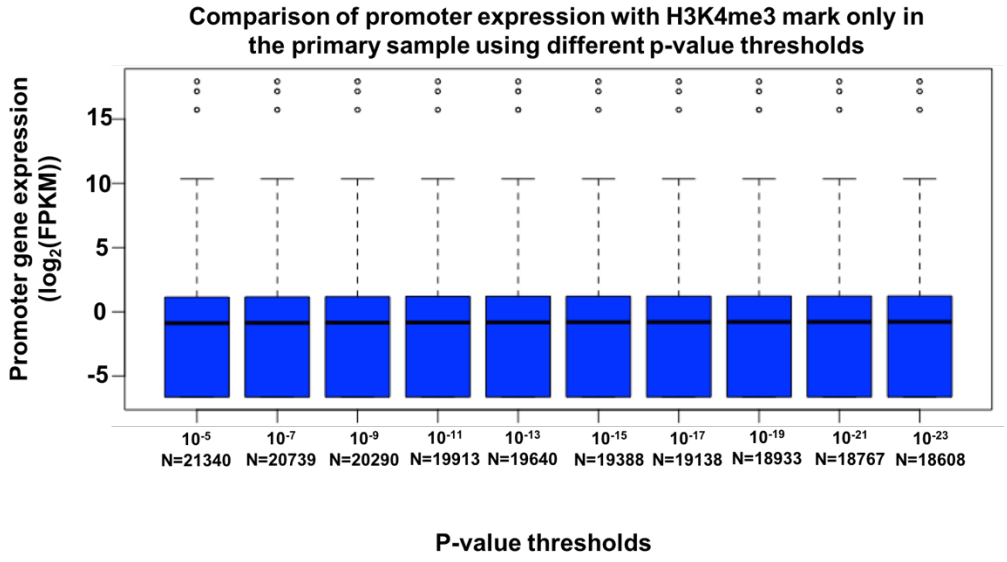
a)



b)


Comparison of promoter expression with H3K27me3 mark only in the recurrent sample using different p-value thresholds
Promoter gene expression (log2(FPKM))
6
4
2
0
-2
-4
-6
10-5 N=2947
10-7 N=2450
10-9 N=2065
10-11 N=1814
10-13 N=1643
10-15 N=1500
10-17 N=1386
10-19 N=1269
10-21 N=1156
10-23 N=1063
P-value thresholds

c)


Comparison of promoter expression with H3K4me3 mark only in the primary sample using different p-value thresholds
Promoter gene expression (log2(FPKM))
15
10
5
0
-5
10-5 N=21340
10-7 N=20739
10-9 N=20290
10-11 N=19913
10-13 N=19640
10-15 N=19388
10-17 N=19138
10-19 N=18933
10-21 N=18767
10-23 N=18608
P-value thresholds

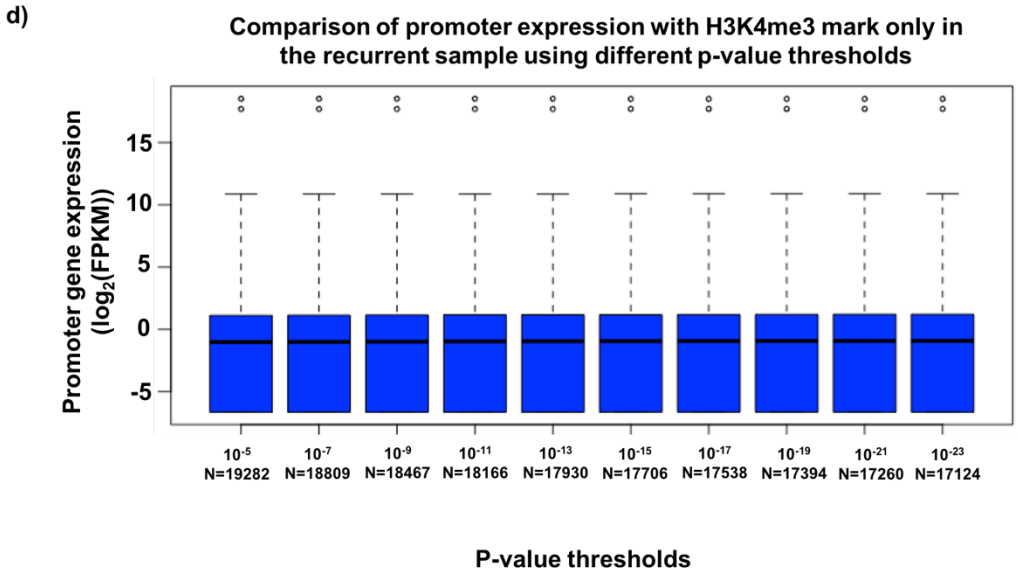


Figure 4-15: The box plots of log2-transformed gene expression of promoters with each histone mark in the primary and recurrent samples of the in-house dataset using different p-value thresholds. (a) Box plots showing the gene expression of the primary promoters with H3K27me3 mark only and do not contain H3K4me3 mark (i.e. H3K4me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. **(b)** Box plots showing the gene expression of the recurrent promoters with H3K27me3 mark only and do not contain H3K4me3 mark (i.e. H3K4me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. **(c)** Box plots showing the gene expression of the primary promoters with H3K4me3 mark only and do not contain H3K27me3 mark (i.e. H3K27me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. **(d)** Box plots showing the gene expression of the recurrent promoters with H3K4me3 mark only and do not contain H3K27me3 mark (i.e. H3K27me3 p-value is > 0.05) at p-values thresholds ranging from $1x10^{-5}$ to $1x10^{-23}$. The X-axes represent the p-value thresholds, while the Y-axes represent the log2-transformed gene expression of the promoters in the primary and recurrent samples. The horizontal lines in each column of the plot represents the mean gene expression value.

A plateau in the gene expression was observed in the promoters with H3K27me3 at a p-value threshold of $1x10^{-15}$. It was difficult to confirm this selection for the promoters with H3K4me3 mark only, therefore, a p-value of $1x10^{-15}$ was selected to be used for further downstream analysis.

4.4 Discussion

High throughput sequencing technology have made it feasible to identify several epigenome markers across the genome in various cell lines (128). It is useful to annotate specific genomic regions with regards their distinct chromatin states to give more detailed characteristics of epigenetic signatures (e.g., poised promoter, active enhancer) for each cell. Chromatin immunoprecipitation (ChIP) followed by sequencing is one of the most commonly used technologies to create genome-wide chromatin-state maps (219).

This chapter aimed to assemble a ChIP-seq data analysis pipeline and adopt or develop a promoter status calling approach to classify whether multiple DNA binding factors or histone modifications were present in a pre-defined promoter region or not. For the purpose of this work, I developed a ChIP-seq analysis pipeline based on the ENCODE consortium. A typical ChIP-seq computational analysis workflow consists of five main steps namely raw data quality assessment, trimming low quality reads and adapter sequences, sequence alignment, removal of the duplicated reads and peak calling. I demonstrated the applicability of the proposed ChIP-seq analysis pipeline on an external dataset from Liau et al (198). I showed that the results generated from this pipeline was similar to those in the paper in which a global reduction in H3K27me3 peaks was observed in the GSC8per cells in comparison to GSC8 cells (**Figure 4.3**). This finding suggested that the bam file processing is working correctly. Numerous studies implement this general ChIP-seq pipeline to generate global epigenomic profiles of different histone marks across multiple cell lines (198, 220, 221).

As this study aimed to call and score each promoter region's status according to the presence or absence of binding or modification signal from multiple ChIPseq experiments on the same sample, two promoter status calling approaches were developed as described in **Sections**

4.5.2.1 and 4.5.2.2. Approach 1 was developed via adoption of ChromHMM to output promoter calling status. A bespoke java programming approach was developed to use the standard ChromHMM output to assign a call for each promoter region, specifically, as per my objective 2. Approach 2 was developed by scoring the enrichment of signal in defined regions compared to the background across the same sized windows across all genomic regions (**Figure 4.2**). The success of this approach in detecting and scoring the enrichment of transcription factors and histone modifications across different cell lines was reported in several studies (206, 222-224).

The performance of each approach was compared to see which is producing results that make the most biological sense. An inspection of raw data determined that Approach 1 does detect signal in the promoter region but the choice of 'most prevalent state' across that region removes that signal from the call, leading to a loss of resolution (**Table 4.10**). One was to resolve this may have been to set a different threshold in deciding which states to call. Approach 2 worked best upon inspection at default setting, however, in some promoters, the final promoter calls were inconsistent with IGV results and this might be linked to the p-value that was used as default in this study (i.e. $1x10^{-5}$).

The issue of inconsistency between the promoter call and the IGV results was solved by optimising the threshold. I tested different p-value thresholds that can be used for further analysis and the results suggested the use of a p-value threshold of $1x10^{-15}$. I was able to show that Approach 2 with the optimized p-value is generating reproducible results in which a significant reduction or a clear absence of H3K27me3 mark in GSC8per in comparison to GSC8 naïve was observed (**Figure 4.11 and 4.12**). This finding was similar to those reported in the paper (198).

In view of the above reported findings, I concluded that Approach 2 enables us to score the enrichment of histone marks along with EZH2 in the promoter regions as per our objective 2. Therefore, I employed the developed pipeline to detect and score the enrichments of the two histone marks along with along with EZH2 and JARID2 in our in-house dataset and the analysis is provided in the next chapter (**see Chapter 5**).

Chapter 5

Genome-wide profiling of H3K4me3, H3K27me3 and EZH2 and their roles in gene transcription in a primary and recurrent GBM sample

5.1 Introduction

As described in **Chapter 1, section 1.3.4**, the N-terminal tail domains of histone proteins are subjected to numerous post-translational modifications including methylation, acetylation, ubiquitination, and phosphorylation. These modifications alter the structure of the chromatin either directly or indirectly, leading to alterations in gene transcription (74, 77, 80). These transcriptional changes in gene expression have been linked to tumour progression and metastasis of several cancer forms, including GBM (81, 101). H3K27me3 and H3K4me3 are the most well-characterized H3 methylation and their role in cancer development has been extensively studied (102). In addition to these marks, EZH2, a subunit of Polycomb Repressive Complex 2 (PRC2), generally mediates H3K27me3, which suppresses stem cell differentiation while mediating stem cell maintenance and self-renewal by trimethylating H3K27 (101). Numerous studies showed that regulation of histone modifications by PRC2 is a key factor of tumour cell plasticity and it is necessary for glioblastoma cells to survive and adapt to their microenvironment. Poor prognosis in GBM patients is caused by disruption of PRC2 function, which is caused by overexpression of its enzyme component EZH2, underscoring the significance of this histone modification in glioblastoma biology. Numerous genes involved in cell-cycle control, cell differentiation, proliferation, and self-renewal were discovered to be suppressed by EZH2 and PRC2 (69, 103)(**see Chapter 1, section 1.3.4.1 for more detail**).

Recent studies demonstrated the role of JARID2 in regulating gene expression through its colocalization with EZH2, which promotes the recruitment of PRC2 (110). Contribution of EZH2 and JARID2 was found to be associated with tumour progression and carcinogenesis (112). The role of JARID2 was emphasized in our group and the preliminary data suggest that gene expression changes associated with JARID2 occur during GBM recurrence. The work highlighted JARID2 as a potential master regulator of transcriptional changes through treatment, but furthermore showed that these changes were taking place at genes that are commonly found to be bivalent in both normal brain and glioma tissue (67). I decided to

investigate this further by characterising and comparing the binding site profiles of EZH2 (as the catalytic subunit of PRC2) and the two histone marks in a matched pair of primary and recurrent fresh frozen GBM samples. I used genome-wide approach for this purpose. It has been demonstrated that profiling of histone modifications provides precise descriptive data that can be used to infer the regulatory effects of histone modifications on gene expression (**see Chapter 1, section 1.3.4.1**) (117). I had also planned to profile the binding of JARID2 but the limiting factor was the amount of tissue available, meaning a full repertoire of antibodies could not be used. The timing of this aspect of the work was during COVID19, when the complete validation of the JARID2 antibodies were to complete. Given that ChIP validated antibodies existed for EZH2 and the two histone marks, these were prioritised.

This chapter summarizes the generation and the analysis of genome-wide chromatin states of H3K27me3, H3K4me3 and EZH2 binding for a matched pair of primary and recurrent fresh frozen GBM samples from an Up responder (one in which the genes dysregulated through treatment are *increased* in expression from primary to recurrence). It outlines the analysis that was performed to quantify the presence of the histone marks, along with EZH2 binding, in gene promoter regions using the genomic enrichment approach. Also, it provides a comprehensive comparative analysis of chromatin states between the primary and recurrent samples. The difference in the prevalence of all marks between samples was statistically examined using Chi-squared tests. The analysis was first performed globally across all promoter regions. Then, I focused on analysing a subset of genes that was found to be dysregulated in GBM's patients following standard treatment. These genes are connected by containing JARID2 binding sites, according to publicly available ChIPseq datasets, and have been coined the JARID2-Binding Site genes (JBS genes). The JBS genes were included in gene set enrichment analysis applied to a gene expression data from a cohort of paired GBM samples, and were found to be the most significant gene set that is dysregulated through treatment in GBM patients. Within these JBS genes, I looked specifically at those JBSgenes that were in the leading edge (i.e. had driven the gene set's enrichment in the gene expression data) of at least 50% of patients (denoted as LE50) and in the leading edge of at least 70% of patients (denoted as LE70). Next, I examined the state transitions through treatment across all promoters, in JBS, LE50 and LE70 genes (**see chapter 1, section 1.3.4.1 for more details**).

To gain insight into the role of H3K4me3 and H3K27me3 in transcriptional regulation of the genes, I integrated the chromatin state maps with gene expression data. I examined the impact that the amount of H3K4me3, H3K27me3 and EZH2 binding at each promoter region had on gene expression across all gene subsets (i.e. JBS, LE50, LE70 and non-JBSgenes). I used a penalised regression approach called ridge regression to see how well the changes in the binding of each histone mark of ChIP experiments can predict the changes in gene expression.

Regression analysis is an important statistical method for describing and characterizing the relationship between a single dependent variable (Y) and one or several independent variables (X). The analysis generates a model that predicts the effect of one or more explanatory variables on the response variable (225, 226). Regression has been widely employed in every scientific and technological discipline, as well as in finance and economics. In genetic studies, ridge regression has been used to accurately predict the level of gene expression (227). The most common types of regression analysis are simple linear regression, multiple regression, least absolute shrinkage and selection operator (Lasso) regression and ridge regression (226). Simple regression, as its name implies, is the most basic type of regression. It is used to study the effect of one independent variable on one dependent variable only when the relationship is linear (226). Linear regression is specified by the equation:

$$\mathbf{Y = a \times X + b}$$

Where Y is the dependent variable, a is the slope of the line, b is the y-intersect of the line and X is the independent variable (i.e. features).

The slop a is called regression coefficient and it gives an indication of how much the independent variable X contributed to the explanation of the dependent variable Y. In many cases, the contribution of a single independent variable is insufficient to fully explain the dependent variable Y. Therefore, multivariable regression analysis should be performed (226, 228).

Multiple regression is similar to simple linear regression except that it involves one or more independent variables to predict the outcome of dependent variable (229). Multiple linear regression equation can be expressed as follow:

$$\mathbf{y = a1x1 + a2x2 + a3x3 + ... + anxn + b}$$

Where Y is the dependent variable, a1, a2, a3, …., an are the coefficients, b is the y-intersect of the line and x1, x2, x3 , …, xn are the independent variable (i.e. features).

Multiple linear regression enables the analysts to characterize the variation in the model and the relative contributions of each independent variable to the overall variance (230). However, as the number of variables increases, the model becomes complex and the analysts must carefully consider how many features to keep and which ones to eliminate. The process of feature selection in machine learning is an important step to preventing overfitting (231). LASSO is one of the regularized linear regression that can be used for feature selection and parameter elimination by adding the penalty term L1 regularization to the absolute value of magnitude of coefficients to lower the coefficient to zero. LASSO uses the advantage of feature selection to build models for datasets that suffer from the multicollinearity problem (i.e. independent variables are highly correlated) (232, 233). Ridge regression (RR) is another type of the regularized regression model that is used to simplify model complexity, but instead of performing feature selection, it estimates and shrinks the coefficients of correlated predictors toward zero. RR penalizes the total squared regression coefficients with an L2 penalty like regular sum of square residuals methods (184, 186, 233) as follows:

$$\mathbf{RSS + \lambda \Sigma \beta_j^2}$$

Where RSS is sum of the square residuals**,** λ lambda is the shrinkage penalty, Σ a Greek symbol that means sum and β is the weights of the coefficients of the independent variables

The sum of square residuals is calculated as follows:

$$\mathbf{RSS = \Sigma(y_i - \hat{y}_i)2}$$

Where y_i is the actual response value for i^{th} observation and $\hat{y}_i$ is the predicted response value based on the multiple linear regression model.

The shrinkage factor which is "$\lambda\Sigma\beta_j^2$ "and the amount of the penalty can be controlled using λ. Choosing a suitable λ value is critical. For instance, the penalty term is ineffective when λ =0, and ridge regression will result in the conventional least squares coefficients. However, as λ approaches infinity, the shrinkage penalty becomes more significant and the ridge regression coefficients approach zero (234). It has been shown that in terms of the prediction errors, RR perform the best among the other regression approaches. This is because RR estimates a regression coefficient for each predictor variable rather than performing variable selection (188). Despite this, the main challenge in RR applications is the selection of the ridge parameter thatcontrols the level of shrinkage of the regression coefficients (189).

5.2 Results

5.2.1 Chromatin states in a primary versus matched recurrent GBM sample differ most at the genes for which expression is most commonly dysregulated through treatment

I initially focused on comparing the epigenetic landscape of a matched pair of primary and recurrent fresh frozen GBM samples as described in **Chapter 4, section 4.2.1**. This is essential to explore the change in epigenetic marks in response to treatment. I performed this comparison by profiling histone H3 lysine 4 trimethylation (H3K4me3; an active mark responsible for transcriptional initiation), H3 lysine 27 trimethylation (H3K27me3, a repressive mark responsible for transcriptional inhibition) and Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit (EZH2, the catalytic subunit of PRC2 that is responsible for trimethylating lysine-27 of histone 3 i.e. creation of the repressive H3K27me3 mark) and calling their status using the genomic enrichment approach with optimised parameters (see Chapter 4) at our pre-defined promoter regions. In summary: the chromatin state map was

generated by scoring the enrichment of signal in defined regions compared to the background across the same sized windows (i.e. 2-kb) across all genomic regions.

I started the analysis by annotating the chromatin state of each promoter into readily interpretable annotation based on the giving score of each element (i.e. H3K4me3, H3K27me3 and EZH2). The program gives a score of 0 if the adjusted p-value for each of the histone mark is higher than the selected threshold or 1 if the adjusted p-value is lower than the threshold at the promoter region. The results suggested the presence of 8 distinct chromatin states (**Table 5.1**) across the promoter regions of each sample as shown in **Figure 5.1**.

State (emission order)	Enriched pipeline Call	Enriched pipeline state
E1	000	Null
E2	010	Active
E3	110	Bivalent
E4	111	Bivalent+EZH2
E5	100	Repressive
E6	011	Active+EZH2
E7	001	EZH2
E8	101	Repressive+EZH2

Table 5-1: Chromatin states calls based on approach 2 for an in-house dataset.
Table includes the emission state order, the state call and its corresponding annotations based on enriched pipeline (approach 2 from Chapter 4)

Figure 5.1 indicates that the distribution pattern of the chromatin states across the promoter regions is similar for the primary and recurrent samples with the dominant state of null (i.e. no histone mark or EZH2 signal) followed by active (H3K4me3 alone) and then bivalent (H3K4me3 and H3K27me3). I assessed whether the prevalence of all marks differed between samples statistically using a 2x8 Chi-squared test. Results indicated that there is no significant difference in the distribution (Chi-square P-value > 0.05). Then, I examined the significance of the individual states that showed a visual difference in their enrichment (i.e. active and bivalent) between the primary and recurrent samples using 2x2 contingency tables (e.g. a 2 x 2 being 'active x all other states' in 'primary x recurrent'). Again, I found no significant difference in their distributions (Chi-square P-value > 0.05).

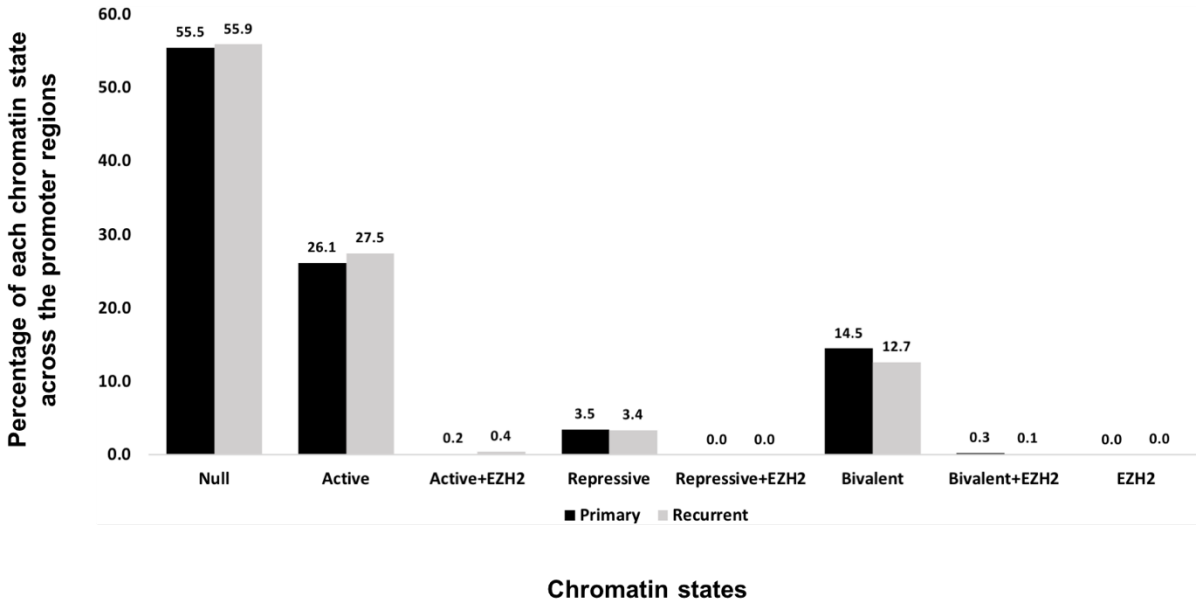


Figure 5-1: Distribution of 8 distinct chromatin states across the promoter regions of the primary and recurrent samples of our in-house dataset. Bar plots show the percentage of each state across the promoter regions resulted from approach 2. The x-axis represents the chromatin states

The Stead group have discovered that a subset of genes is dysregulated in GBM patients following standard treatment, hypothesising this being due to the epigenetic remodeling of their promoters via mechanisms involving JARID2 (67). Hence, I decided to closely explore the extent to which the distribution of the chromatin states vary in this subset of genes (JARID2 binding site genes – JBS genes) vs non-JBS gene promoters (**Figure 5.2**). Whilst the distributions of the chromatin states between the primary and recurrent samples appear to be the same when investigating all genes, the distributions are substantially different in JBS gene versus non-JBS gene promoters. According to **Figure 5.2**, the difference between JBS genes and non-JBS genes pertains to the active and bivalent promoter states. There are more active promoters in non-JBS genes and more bivalent promoters in JBS genes. The difference in the distribution of these two marks was assessed statistically via chi-squared test using 2x2 contingency table and the result was significant for the primary sample (Chi-squared P-value is 4.7×10^{-75}) showing that JBS genes are significantly more likely to be bivalent. This is also true in the recurrent GBM (Chi-squared P-value is 3.3×10^{-05}).

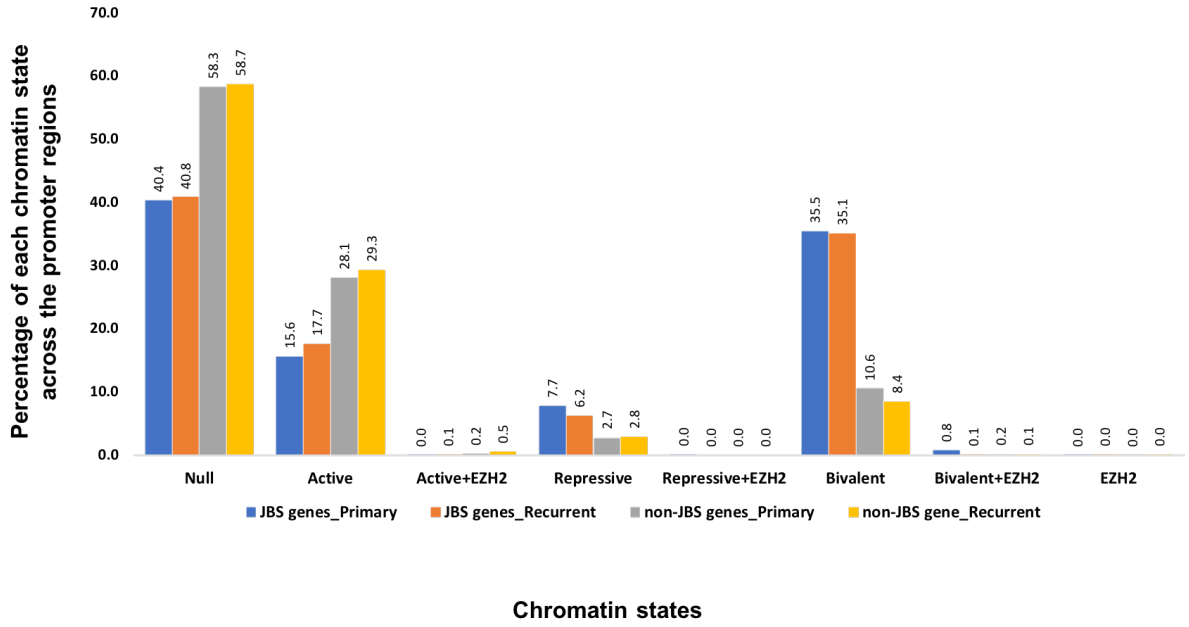


Figure 5-2: Distribution of the chromatin states in JBS and non-JBS gene promoters of the primary and recurrent sample. Bar plots show the percentage of each state in JBS and non-JBS gene promoters. The x-axis represents the chromatin states.

To expand the analysis, I again explored the distribution of chromatin states in the JBSgenes, but also specifically those JBSgenes that were in the leading edge of more than 50% of patients (denoted as LE50) and in the leading edge of more than 70% of patients (denoted as LE70) as explained in **Chapter 1, section 1.3.4.1**. Results are shown in **Figure 5.3**. These gene subsets are our genes of interest and previous work in our group showed that they are most commonly and significantly dysregulated in GBM samples through treatment, thus, we think they are candidate drivers of treatment resistance when they change in expression.

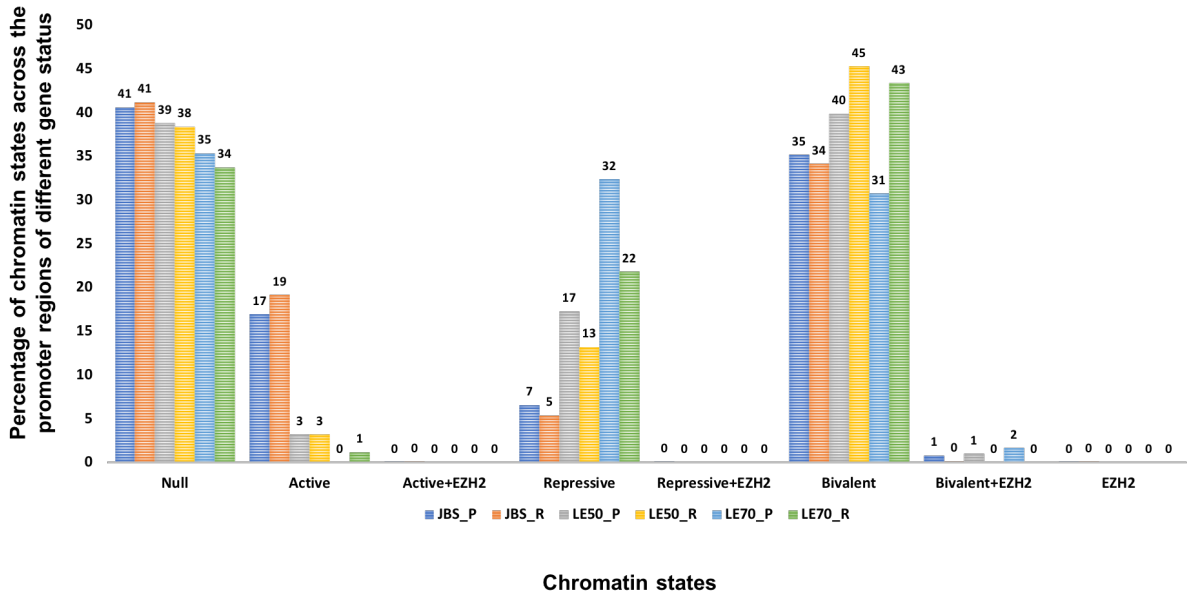


Figure 5-3: Distribution of the chromatin states in JBSgenes, LE50 and LE70 gene sets in the primary and recurrent sample. Bar plots show the percentage of each state in JBSgenes, LE50 and LE70 in the primary and recurrent samples. The x-axis represents the chromatin states.

The distribution patterns of the chromatin states are visually the same in each set of genes except for: 1) the repressive mark for which there appeared to be a larger reduction through treatment compared to all JBS genes, though this did not prove to be significant (Chi-squared P-value is >0.05); and 2) the bivalent state which is noticeably gained in response to treatment, compared to the full JBS gene set (Chi-squared P-value is 3.4E-05).

5.2.2 Chromatin state transition analysis revealed that JBS gene promoters tend to be bivalent through treatment

To investigate whether the relative gain in proportion of bivalent promoters at leading edge genes (**Figure 5.3**) is caused by a shift away from other states, while the bivalent state remains stable, or transit to a bivalent state, I studied the chromatin state transitions through treatment at each individual promoter. I plotted the percentage of each observed transition for all promoters and for JBS, LE50 and LE70 genes (**Figure 5.4**).

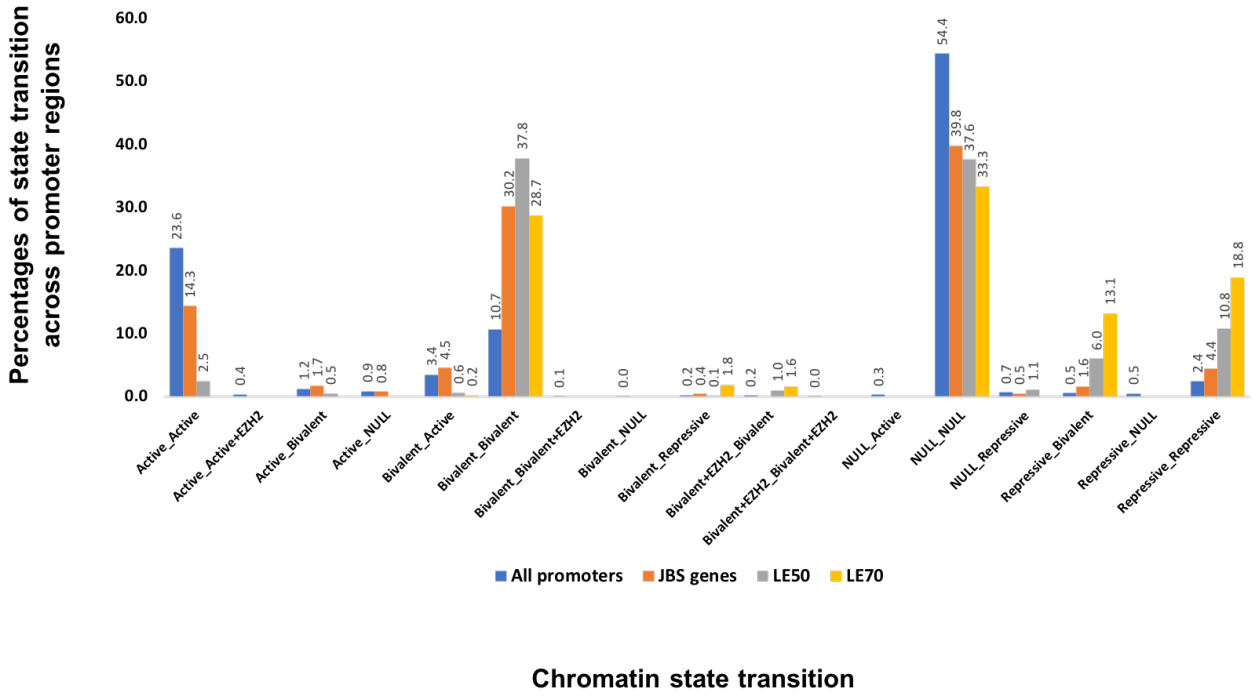


Figure 5-4: Genome-wide chromatin state transition in an in-house dataset. The bar plots show the percentage of state transitions across all promoters, JBS gene, LE50 and LE70 gene promoters.

The most popular state transition is null-null (promoters that are null [000] in the primary and remain null in the recurrence) in all cases, except LE50 genes where it is almost equally prevalent as bivalent-bivalent. However, the second most prevalent transition is active-active across 'all' promoters whereas it is bivalent-bivalent when looking specifically in the JBS, LE50 and LE70 gene subsets. There is a significant increase in the proportion of repressive-bivalent promoter transitions (promoters that only harbour the H3K27me3 mark in the primary tumour, making them repressive, but additionally gain the H3K4me3 mark in the recurrence, making them bivalent) in the LE50 (Chi-squared P-value is 0.043) and LE70 (Chi-squared P-value is 0.0005) gene subsets compared to when all genes are analysed. This suggests that the enrichment in bivalency in our genes of interest is partially driven by changes from repressive to bivalent state, but is mostly caused by retention of bivalency within these gene subsets where we know expression increases during treatment.

These findings might indicate that the change in gene expression specifically in JBS gene promoters might not be a consequence of epigenetic remodelling based on these specific histone marks, or alternatively that it cannot simply be explained by our binary classification of chromatin state. To interrogate the latter, I decided to perform a more in-depth analysis of the exact level of the mark present at each promoter (i.e. H3K4me3, H£K27me3 and EZH2), especially those which are bivalent. The hypothesis here is that, although the promoter remains in a bivalent state, the balance between the repressive (H3K27me3) and active (H3K4me3) mark may alter in ways that elicit changes in gene expression. Therefore, I decided to examine the strength of each signal at promoters.

5.2.3 The level of the repressive mark (H3K27me3) and the active mark (H3K4me3) in JBS gene promoters associates with the changes in the gene expression

The genomic enrichment program gives a score of 0 or 1 to report the absence or presence of the mark. However, this output does not account for the amount of H3K27me3, H3K4me3 and EZH2 binding at each promoter. The above findings suggest that the change in gene expression at JBSgenes is not driven on the presence or absence of the mark. However, it could still be impacted by the amount of the mark present at that promoter. Therefore, I decided to score the amount of mark at each promoter as described below.

The program determines enrichment of the ChIP signal across a specified region (in our case, promoters) based on a Poisson test p-value: the smaller the p-value the greater the signal for a given histone mark or protein binding (**Figure 5.5a**). As shown in **Figure 5.5a**, it is possible for two peaks to be significant (i.e. less than a 0.05 threshold) despite the fact that they differ in the size. To increase the resolution of our investigation, I wanted a continuous score that more intuitively relays the strength of ChIP signal. This can be achieved using -log10 (adjusted p-value), to make a more human-readable scale in which higher scores indicate greater enrichment; this facilitates the analysis of studying the amount of mark binding at each promoter.

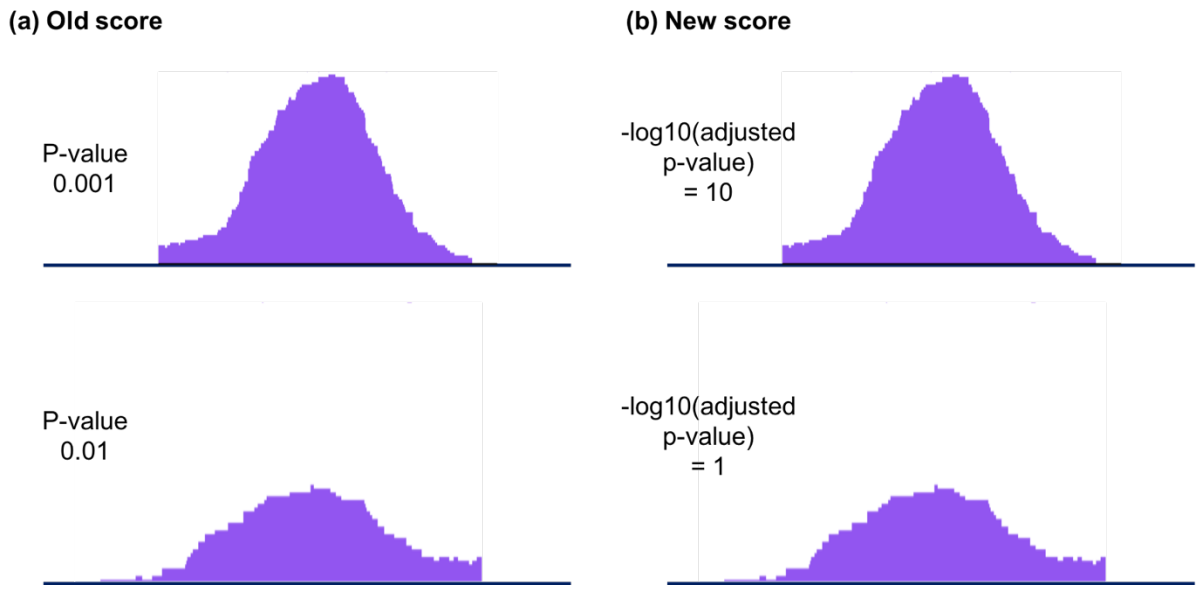


Figure 5-5: Schematic representation of the old and new scoring system of the mark signal to assesses the level of mark/binding (i.e. signal) at each promoter. (a) The significance of the signal was reported based on the p-value in which the smaller the p-value the greater the signal for that mark. The new scoring system **(b)** is to calculate the -log10(adjusted p-value) which flips the data so that the higher the score the more signal.

I calculated the change in the amount of signal through treatment by subtracting the score for each ChIP'd mark in the recurrent from the score of the mark in the primary (to give the delta score). I plotted the delta scores of H3K27me3, H3K4me3 and EZH2 for promoters in each gene group (i.e. JBS, LE50, LE70 and non-JBS genes) to give **Figure 5.6 (a-c)**.

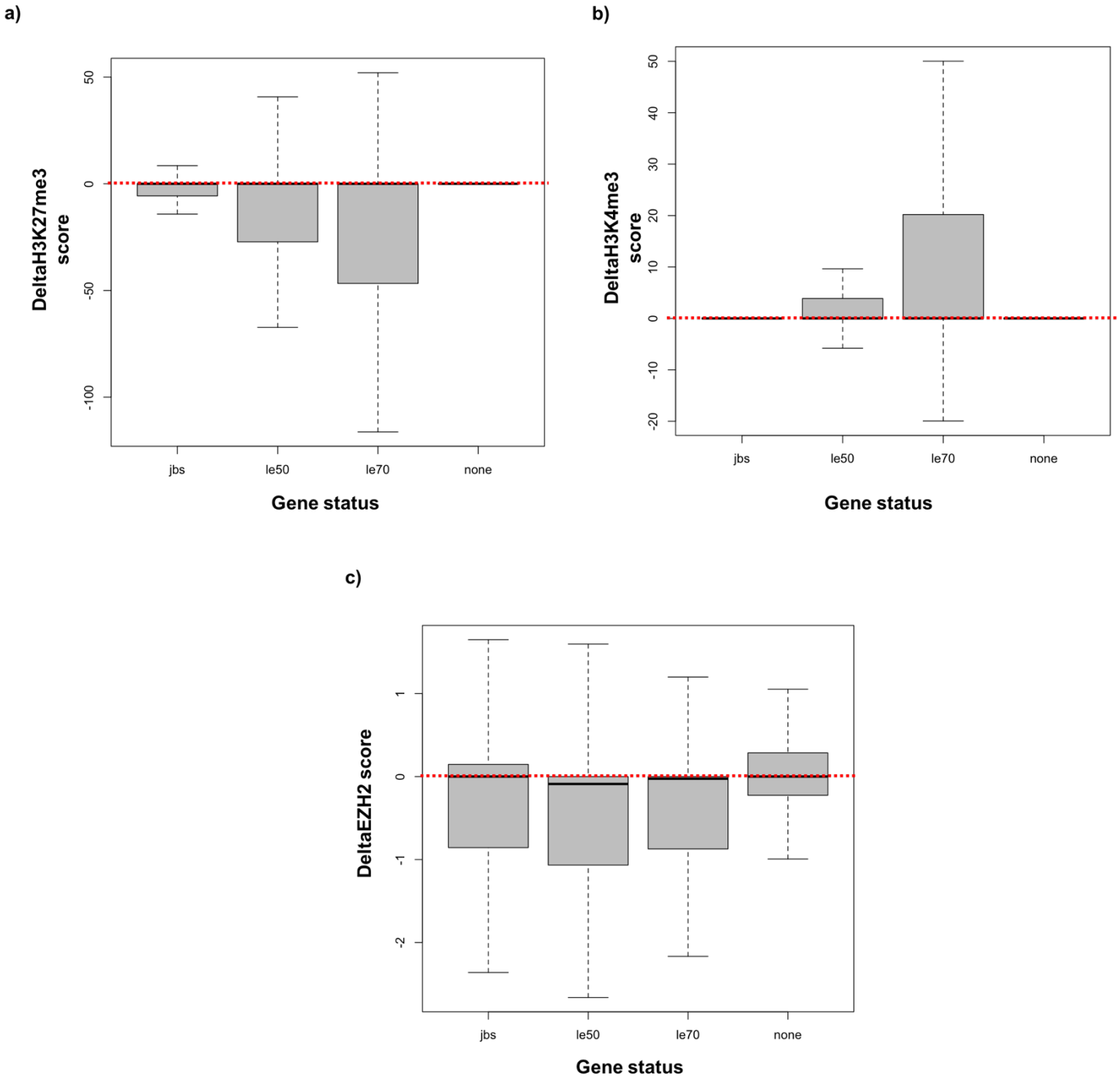
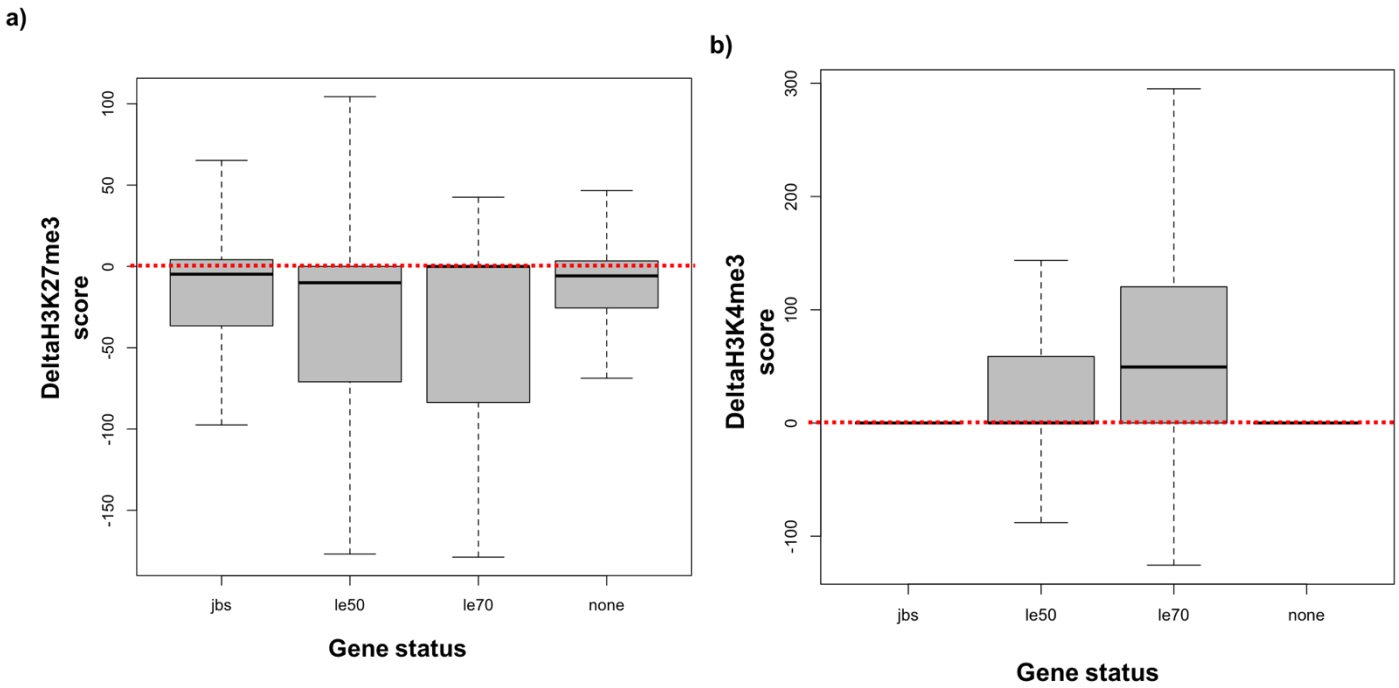


Figure 5-6: Box plot of the changes in scores (delta score) for (a) H3K27me3, (b) H3K4me3 and (c) EZH2 in the JBS, LE50, LE70 and non-JBS gene sets. Each boxplot represents the changes in the amount of signals at promoters for genes in each gene set through treatment. The red dotted line denotes a delta score of zero i.e. no changes in the ChIP signal.

As shown in **Figure 5.6a**, the H3K27me3 signal predominantly changes through treatment, except at non-JBS gene promoters. JBS genes have greater loss of H3K27me3, with the LE50 and LE70 gene subsets, which we think are driving the treatment resistance when they change in expression, undergoing larger reductions in signal from this mark indicating a greater loss of H3K27me3 at these genes. On the other hand, the of H3K4me3 signal increases more through treatment (**Figure 5.6b**) specifically in our genes of interest. So, across LE50 and LE70, despite there being retention of a bivalent state (**Figure 5.4b**) there is a reduction of H3K27me3, and an increase of H3K4me3, which coincides with the expression of these genes being upregulated during treatment in this Up-responder patient. With regard EZH2, LE50 and LE70 are more likely to reduce EZH2 signal, as may be expected because there is a clear reduction of H3K27me3. EZH2 is responsible for catalysing H3K27 trimethylation so a reduction in EZH2 binding is directly linked with a reduction in the H3K27me3 mark.

I then further examined the signals specifically in the genes that stay bivalent through treatment, and similar findings were observed. There is a clear reduction of H3K27me3 (**Figure 5.7a**) and EZH2 binding signal (**Figure 5.7b**) in JBS, LE50 and LE70 genes in comparison to non-JBS genes and a noticeable increase in signal from the H3K4me3 mark (**Figure 5.7c**).



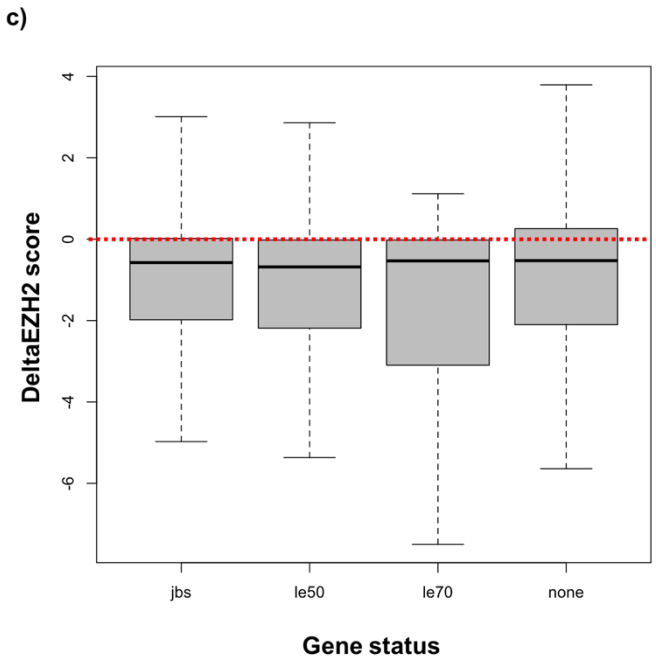


Figure 5-7: Box plot of the changes in (a) H3K27me3, (b) H3K4me3 and (c) EZH2 delta score for JBS, LE50, LE70 and non-JBS gene sets for genes that stay bivalent through treatment. Each boxplot represents the changes in the amount of mark binding at each gene set for genes that stay bivalent through treatment. The red dotted line represents the delta score of zero where there are no changes in the amount of mark binding.

To investigate whether these (changes in) epigenetic scores directly associate with (changes in) gene expression, I used ridge regression. I built several models using gene expression in the primary or recurrent sample, or log2FC in gene expression through treatment, as response variable and either the per-sample score, or delta score, for H3K4me3, H3K27me3 and EZH2 as predictor variables. In ridge regression, the R^2 value indicates how much of the variability in the response variable is captured in the model i.e. how accurately the predictor variables can predict the response. Gene expression is a hugely complicated biological phenomenon, resulting from orchestration of many players, so a simplified model is not expected to be able to adequately predict gene expression, or changes therein, but I wanted to use this approach to compare R^2 value across models and see if the histone marks, and EZH2 binding, were more predictive in the gene sets in question. For the purpose of this work, I used ridge regression to build several models. In the first model, I used the score (i.e. -log10

(adjusted p-value)) of two histone marks (i.e. H3K4me3 and H3K27me3) along with EZH2 of the primary sample as predictor variables and the gene expression as the response variable and I fitted these parameters in a simple regression model as follow:

$$\mathbf{y = a1x1 + a2x2 + a3x3 + ... + anxn + b}$$

Where Y is the response variable (i.e. gene expression), a1, a2, a3, …., an are the coefficients, b is the y-intersect of the line and x1, x2, x3 , …, xn are the independent variable (i.e. features).

I repeated the model and this time I used the score (i.e. -log10 (adjusted p-value)) of two histone marks (i.e. H3K4me3 and H3K27me3) along with EZH2 of the *recurrent* sample as predictor variables. I build the third model using the delta score of H3K27me3, H3K4me3 and EZH2 as predictor variables and the log_2FC as response variable. I applied each model for each gene subset (i.e. non-JBSgenes, JBSgenes, LE50 and LE70 genes) and in each model, I calculated the R^2 value. In all models, the R^2 value which describes how well the model is predicting the response variable was very low. This is likely because gene expression is such a complicated phenomenon that these marks alone cannot suitably model it. But, interestingly, we found the R^2 values are higher for our gene sets of interest (**Figure 5.7**) which implies that those marks are more important for driving gene expression in those genes than the others.

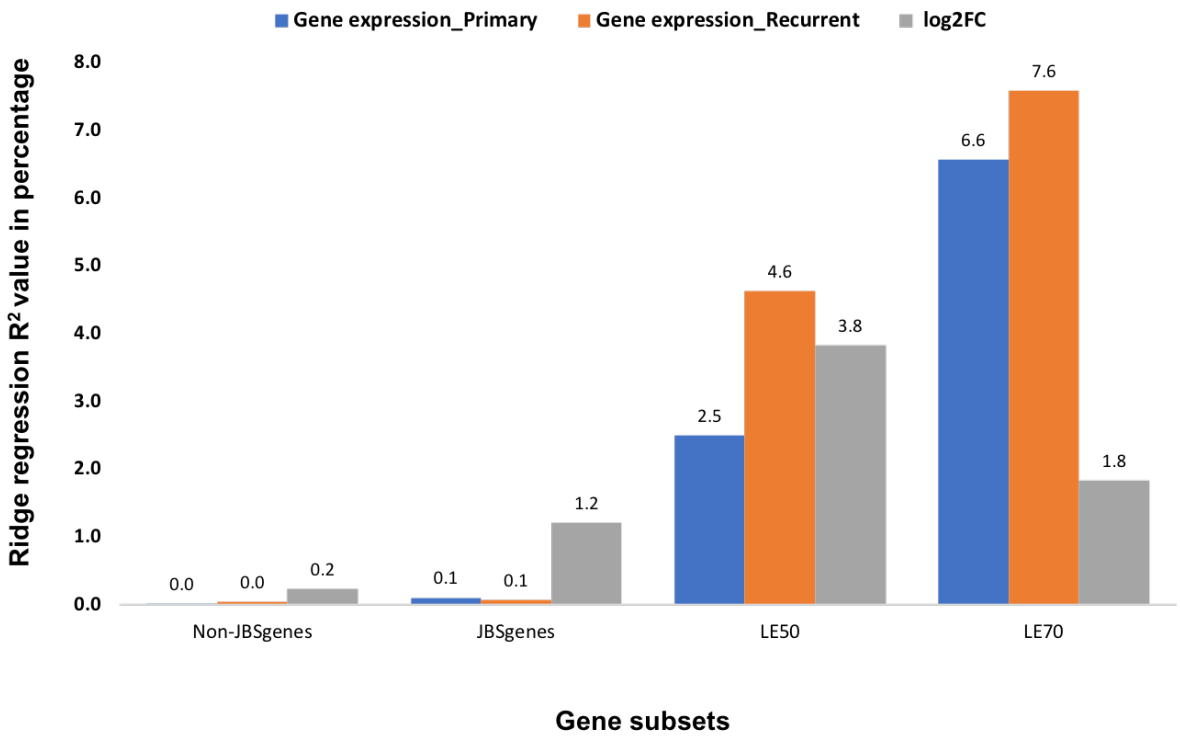


Figure 5-8: Ridge regression analysis of different regression models cross our genes of interest (i.e. Non-JBSgenes, JBSgenes, LE50 and LE70). The bar plots showed the R^2 values of different regression models using the two-histone marks along with EZH2 as predictors and the gene expression of the primary or recurrent samples, or the the log_2FC, as the response variable across non-JBSgenes, JBSgenes, LE50 and LE70 genes.

5.3 Discussion

Protein-DNA interactions control a wide range of biological activities, including the regulation of gene expression, DNA replication, packaging of chromosomal DNA, and disease states. Genotype and expression analyses are strengthened by epigenetic data (235). Genome-wide mapping of chromatin states including histone variations and post-translational modifications have become key research focuses, therefore, tremendous efforts have been made to understand these interactions (75). The main hypothesis for my research is that histone remodeling is driving the changes in the gene expression observed in GBM through treatment, and herein I was trying to acquire evidence for (or against) that. To look into this, I generated a genome-wide chromatin landscape for H3K27me3, H3K4me3 and EZH2 binding from matched fresh frozen pair primary and recurrent GBM samples of our in-house dataset.

I called the promoter status of the primary and recurrent samples using the developed promoter calling status from Chapter 4. Then, I performed an integrative analysis on histone marks along with EZH2 by correlating their modifications with the changes in gene expression. I found that the pattern of distributions and the occurrence of chromatin states between both samples were insignificant (Chi-square P-value > 0.05). Similar results were obtained for human colorectal cancer sample and paired normal mucosa in which the profiles of histone modifications between these two samples were similar (236).

The chromatin state distribution was generated for subsets of genes (i.e. JBSgenes, LE50 and LE70 genes) that are found to be dysregulated in GBM patients following standard treatment due to the epigenetic remodeling of their promoters (67). I found that the amount of H3K4me3, H3K27me3 and EZH2 binding at JBS, LE50 and LE70 caused a change in gene expression. Interestingly, I found that the bivalent state is predominant in these genes.

Primary-to-recurrent chromatin state transitions were studied to investigate whether JARID2 binding genes stay bivalent through treatment or acquire the bivalency through state transition. The results suggested that these genes tend to be bivalent through treatment and this bivalency drives the recurrence of the tumour. However, for LE50 and LE70 genes, the enrichment in bivalency was found to be partially driven by changes from repressive to bivalent state. Despite the fact that bivalent promoters do not poise genes for rapid activation and that bivalent genes are transcriptionally inactive (106), upregulation of these gene sets was observed in U response subtype patients. This might indicate that the change in gene expression specifically in the promoter of these two sets of genes might not only be a consequence of epigenetic remodeling based on these specific histone marks, or alternatively that it cannot simply be explained by our binary classificationof chromatin state. It has been noted that as cancer develops, bivalent promoters lose their histone modifications while gaining DNA methylation (237). There is a strong correlation between the H3K27me3/H3K4me3 ratio and the DNA hypermethylation of bivalent promoters in cancers (238). Evidence suggests that the orchestration of gene expression throughout cancer development and tumorigenesis may depend on DNA hypermethylation and histone modification (239). This correlation was further studied by another group who reported the

presence of two different classes of bivalent promoters: promoter with low H3K27me3:H3K4me3 ratio (loBiv) and was substantially enriched for the activating marks H3K4me3 and H2AZ and promoters with high H3K27me3:H3K4me3 ratio (hiBiv) and had increased occupancy by PRC1/2 components. The study also showed that loBiv genes might be more compatible with transcription. These findings might explain the reason behind the upregulation of these genes while they are in the repressed state (238, 240). In general, bivalent regions are found to be associated with chemo-resistance and facilitate GBM tumorigenicity. Several studies reported that a substantialnumber of genes in tumours have bivalent promoters and this bivalency alters the expressionof genes and confers phenotypic plasticity (192-194). Bivalent genes offer new therapeutic opportunity for the management of several types of cancers including GBM in the future (241).

Also, the analysis suggested that JARID2 genes promote tumour recurrence through transcriptional reprogramming in GBM patients following standard therapy. JARID2, in general, contributes to GBM malignancies through several cancer-related signaling pathways such as cancer cell epithelial-mesenchymal transition and stem cell maintenance (242). Deregulation of JARID2 was found to be associated with tumour initiation and progression in different types of cancers. For instance, high expression of JARID2 promoted epithelial and mesenchymal transition (EMT) in hepatocellular carcinoma (HCC) tissues, lung and colon cancers (243, 244). In addition, alteration of JARID2 expression in bladder cancer patients was found to be positively associated with cell invasion and sphere forming ability. Similarly, alteration in JARID2 expression promotes the proliferation and invasion of ovarian cancer through the PI3k/AKt signaling pathways (245). Several knockdown experiments were performed in order to understand the role of JARID2 in tumour initiation and progression and the results suggested that knockdown of JARID2 lowered the population of tumour-initiating cells and inhibits the proliferation, invasion and the EMT in different types of cancers (243, 244). Collectively, these studies provide insights into the possible role of JARID2 in the proliferation, invasion and metastasis of cancers, nevertheless, its role remains unclear (244). Also, it provides evidence that JARID2 can be used as a potential therapeutic target in the treatment of cancers. There is a need for a therapeutic agent (i.e. small molecular inhibitors) that can directly target JARID2 and inhibit its expression in GBM patients (242, 246). To

effectively treat GBM patients, the exact pathways need to be investigated in order to develop potent targeted chemotherapeutic drugs.

I concluded that histone remodeling is associated with the changes in gene expression, however, this work hasn't proved that this remodeling is *causing* the changes in gene expression, it could be that gene expression changes result in histone remodelling. However, my data justifies the need to study this further e.g. by looking into histone modifiers and to show that when histone methylation is stopped, it stops the gene expression changes in order to provide a causal link.

Chapter 6

Experimental optimizations and computational analysis of CUT&RUN experiments

6.1 Introduction

6.1.1 Overview of CUT&RUN

Over the past few decades, Chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) has become a well-established approach for genome-wide profiling of chromatin associated proteins and chromatin states, including post-translational modifications (PTMs) (136). ChIP-seq typically involves formaldehyde fixation (cross-linking) of bulk chromatin inside the cells which is followed by mechanical shearing or enzymatic cleavage into shorter fragments. The cross-linked DNA is then immunoprecipitated with the protein-factor of interest (128, 131). However, the solubilization of chromatin by sonication causes the disruption of the cells and their nuclei, and often results in capture of/contamination by non-immunoprecipitated fragments meaning very deep sequencing is needed in order to resolve the targeted protein binding sites (247). In general, current standard ChIP-seq protocols have some limitations and are not free from artifacts. The primary reported limitations as described in **Chapter 1, Section 1.6.1** are the need for abundant starting materials, in the range of 1-20 million cells per immunoprecipitation, and the cost. Additionally, ChIP-seq suffers from poor resolution and low signal to noise ratio (131-133). To address some of these limitations, an *in-situ* method called Cleavage Under Targets and Release Using Nuclease (CUT&RUN) was developed by Skene and Hanikof as an alternative to ChIP-seq. This method is an improved chromatin immunocleavage (ChIC)-targeted nuclease strategy (248). It isolates protein-DNA complexes on native chromatin by binding the transcription factors or histone modifications with target-specific primary antibodies that have been tethered to micrococcal nuclease (MNase) tagged protein A (pA). The latter interacts with immunoglobulin G (pAG-MNase) and cleaves DNA either side of the antibody binding location, with the fragment then eluting out of the cell (**Figure 5.1**) (249). Since only the targeted fragments enter into solution, with the majority of DNA left behind in the cell/nucleus, CUT&RUN has extraordinarily low background levels and this require less depth in sequencing in comparison to ChIP-seq. Hence, the CUT&RUN assay outperforms

ChIP-seq in resolution and signal-to-noise ratio. In addition, CUT&RUN requires low cell input with a simple cost-effective workflow which can be completed within 1-2 days (248, 250). Comparison of CUT&RUN to ChIP-seq protocols are summarized in **Table 6.1**.

Method / Parameters	ChIP-seq	CUT&RUN
Number of input cells required	10^7 - 10^8 cells	5,000 – 250,000
Input sample	Fragmented chromatin	Intact cells or nuclei
Resolution	Poor	High
Signal-to-noise ratio	Low	High
Fragmentation bias	GC bias	No
Protocol time (Cells to DNA)	3-5 days	1-2 days
Chromatin fragmentation	Sonication	pAG-MNase digestion
Sequencing depth required	>30 million	3 -5 million

Table 6-1: Summarizes the main differences between ChIP-seq and CUT&RUN protocols (96, 196)
Table summarizes the main differences between ChIP-seq and CUT&RUN assays

CUT&RUN has been tested and applied widely on different cell lines including human embryonic stem cell lines and diffuse midline glioma (DMG) patient derived cell lines to profile different histone marks such as H3K4me3, H3K4me3 and H3K27ac (251, 252). It has been also been applied widely to map yeast transcription factor (TF) binding sites and several GATA1 motifs. It became clear that CUT&RUN had higher enriched areas in a smaller window at the peak center in comparison to ChIP-seq. This reflects the high resolution of this technique as mentioned above (248). The method was modified by other groups and tested on Drosophila tissues, fresh frozen mouse tissue and xenograft tissue. The adapted protocol reliably generates an efficient epigenome profile of transcription factor binding sites and histone modifications (248, 253).

Several CUT&RUN protocols have been developed making it possible to use on frozen or fresh tissues and cells with low starting numbers of intact cells or nuclei. The improved CUT&RUN protocol and its workflow is presented below (253).

6.1.2 CUT&RUN experimental workflow

CUT&RUN is a simple and straightforward technique that can be completed in 1-2 days using general laboratory equipment (253). A typical CUT&RUN workflow is presented in **Figure 6.1**.

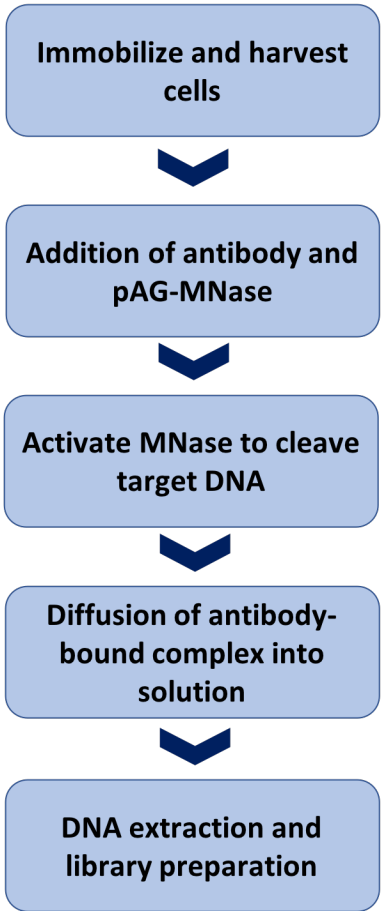


Figure 6-1: Typical CUT&RUN workflow. Workflow of CUT&RUN assay from cell harvesting to library preparation and sequencing.

In order to isolate the targeted DNA from the protein-DNA complex, intact cells or unfixed nuclei are first harvested, washed in a buffer solution and coupled to activated Concanavalin A-coated magnetic beads to facilitate cell handling and reduce cell loss during successive washes. Cell membranes are then permeabilized with antibody buffer which contain digitonin

so that the primary antibody can enter the nuclei and bind to its target *in situ* (Chromatin-associated protein or PTM). At this stage of the protocol, pAG-MNase is added to each sample and pAG domain of the pAG-MNase fusion protein binds to the heavy chain of the primary antibody. This is subsequently directing the enzyme to the desired chromatin region. This incubation is performed in the presence of the digitonin buffer which is free of Calcium Chloride ($CaCl_2$) to inhibit the premature activation of the MNase. The final step is to initiate the DNA digestion by the addition of Calcium ions (Ca^{2+}). This will promote the cleavage and the release of antibody-bound chromatin from genomic DNA, out of the nuclei, into the supernatant, where it can be easily collected and purified using either a DNA spin column or phenol/chloroform extraction. The purified enriched DNA is then quantified using qPCR or used directly for library preparation and sequencing (248, 249).

Similar to ChIP-seq, the success of CUT&RUN is largely depending on the antibody's affinity for its targets and its specificity under the binding conditions. Therefore, the antibodies should be successfully tested for specificity using immunoprecipitation (IP) or immunofluorescence (IF) (249). The primary reported limitation of this technique is that the amount of DNA recovered can be very low. Analysing samples with very low DNA is usually difficult and even with sensitive equipment such as capillary electrophoresis Agilent TapeStation and Qubit, it is hard to detect the cleaved fragment or measure the sample's concentration (254). To address this limitation, it is recommended to increase the PCR amplification cycles to 12-15 cycles in order to generate a library with DNA concentration of 10-30 ng/µl (249).

Despite the fact that different technologies have been developed to profile the transcription factor binding and chromatin states, an end-to-end computational pipeline for analysing such data is still lacking. To fill this gap, a simple and flexible bioinformatics pipeline for CUT&RUN data analysis and visualization was developed. The details of this pipeline are explained in the following section (249).

6.1.3 CUT&RUN analysis pipeline

As with many high-throughput sequencing technologies, CUT&RUN generates massive datasets that require appropriate computational pipelines designed specifically to analyse

these data. CUT&RUNTools has been introduced as a flexible computational pipeline that provides complete data analysis and includes quality assessment, trimming of low quality reads, sequence alignment, pre-processing of reads and filtering, peak calling, data aggregation and visualization, motif finding and motif footprinting steps. It uses GNU parallel processing to increase the computing performance in the main steps such as read trimming, mapping and filtering. The tool was implemented using Python, R and Bash script (249).

In general, this module takes raw FASTQ files as inputs and perform an initial trimming with Trimmomatic. Due to the presence of short fragments in CUT&RUN experiment which mainly occur because of fine cutting by pAG-MNase, the trimming is optimized with settings that are most effective for detecting adapter contamination in short reads. Trimmomatic is a fast command line tool that trims and filters poor quality reads and removes adapters. In addition to Trimmomatic, a separate tool called Kseq is also used to filter and trim reads with 6 bp or less from the 3' end of each read. Then, the trimmed reads are aligned to the reference genome using the most commonly used tool, bowtie2. Uniquely mapped reads with high mapping quality are retained for further analysis. MACS2 is mainly applied to call peaks with default parameters. The remaining steps include motif finding using MEME, motif footprinting analysis using CENTIPEDE and determining direct binding sites (249).

CUT&RUNTools reports some quality metrices to evaluate the quality of CUT&RUN datasets. These include library size, adapter content percentage, fragment size distribution, reads duplication reads, alignment percentage, number of enriched peaks and enrichment of expected motifs. The library size, determined by the number of reads in the sample library, should be at least 10 million reads, and ideally at least 15-20 million. The percentage of reads with adapter, or percentage of reads kept after read trimming, should be less than 10-15% in a good-quality dataset. The fragment size distribution should be within the expected range (≤ 120 bp) but this is mainly applicable for transcription factor binding analysis because reads with < 120 bp are likely to contain TF binding sites. The percentage of read duplication should be low (10% –a maximum of 15%). The fraction of reads that map concordantly to the reference genome is used to compute the alignment percentage. A good dataset should have a high alignment percentage, > 90%. Users are encouraged to make their own decision on these metrices as there is no single score that captures the overall quality (249, 255).

In order to achieve Aim 1 (Objective 4), I decided to investigate whether CUT&RUN could be used to study protein-DNA interactions and histone modification locations in GBM. I performed the assay for H3K4me3, H3K27me3 and JARID2 on GBM63 patient derived cell lines and fresh frozen patient tumours. I present, in this chapter, the optimization of this assay on the above-mentioned samples and I describe the main challenges that I faced, especially in the library preparation. Also, I show that the ChIP-seq analysis pipeline that I developed in chapter 4 is compatible to use for CUT&RUN assay. In general, a complete analysis of these samples from quality assessment of raw data to peak calling is described in detail.

6.2 Results

To investigate protein–DNA interactions and histone modification locations in GBM samples as per Aim 1 (Objective 4), I initially performed and optimized CUT&RUN assays on 2 biological replicates of patient derived cell lines (GBM63) using antibodies against histone modifications H3K4me3 and H3K27me3 and JARID2, following the manufacturer's protocol. The two main criteria to consider when optimizing this assay are what sonication conditions to use for the input sample to generate the optimal DNA fragment size of 100-600 bp and what is the most suitable control DNA sample for downstream analysis.

6.2.1 Sonication condition was successfully optimized for CUT&RUN experiment on patient derived cell lines (GBM63)

The sonication condition for the input DNA sample was optimized at high power setting at 4^{o}C for 15,20,25 and 30 cycles of 30 sec 'On' and 30 sec 'Off'. The fragment length at each tested condition was checked by running an aliquot of each sample in 1% agarose gel. An intense stained part of smear DNA was found between 100 and 200 bp at 20, 25 and 30 cycles (**Figure 6.2**) which would suggest that the majority of the sonicated DNA is at that fragment size at these sonication conditions. No band was observed at the sonication condition of 15 cycle and this might be due to insufficient amount of loaded DNA. In view of these results, and since 20, 25 and 30 cycles show similar outcomes, I decided to use 25 cycles rather 30 in order to save some time during my following experiments. In general, the selected sonication condition will allow me to obtain the highest amount of chromatin fragments with an optimal fragment length between 100-200 bp region as recommended.

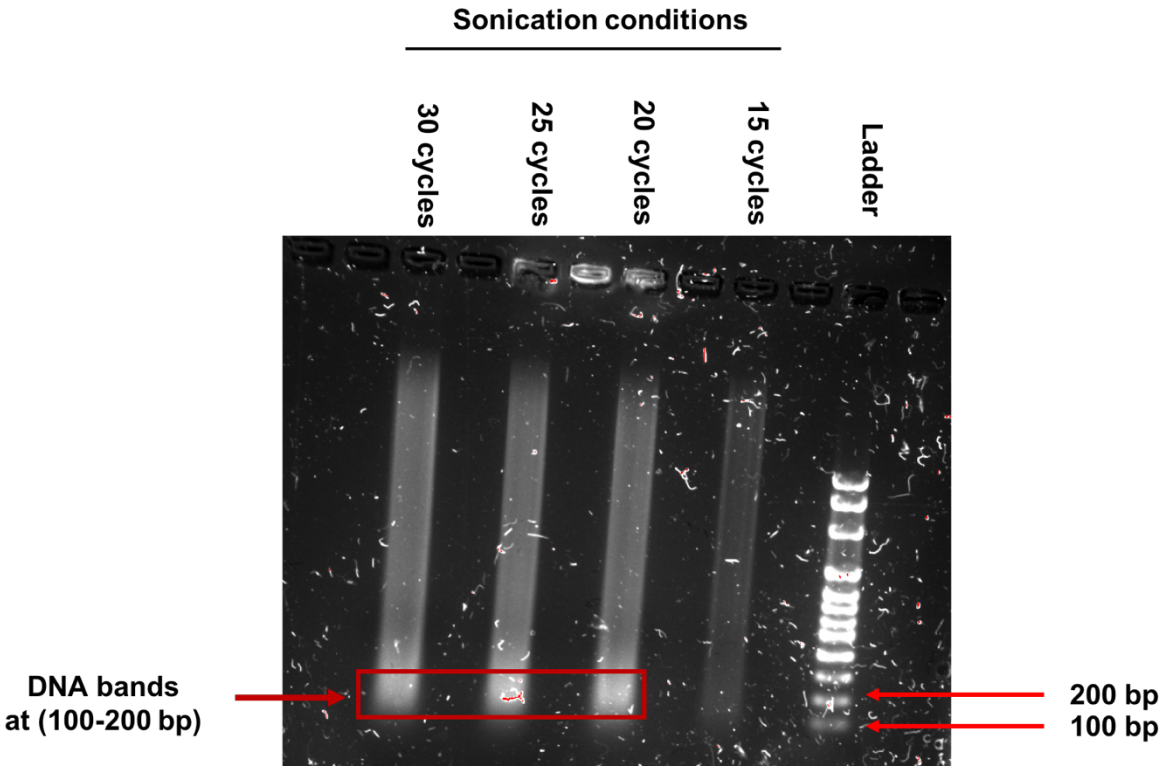


Figure 6-2: Agarose gel analysis of the length of input DNA fragmented by sonication at 15, 20, 25 and 30 cycles.

6.2.2 Quantitative real-time PCR (qPCR) analysis revealed a successful amplification of H3K4me3 in both replicates and unsuccessful amplifications of H3K27me3

An initial check of the H3K4me3 and H3K27me3 amplifications were evaluated using qPCR to check for enrichment in the reactions of both GBM63 replicates. A CUT&RUN validated antibody for H3K4me3 is provided with CUT&RUN kit, as positive control. A PCR primer set of the human RPL30 gene is provided also with CUT&RUN kits to be used in conjunction with H3K4me3 antibody. This gene locus is universally activated, therefore, always contains the H3K4me3 histone mark, and can be used to verify that the CUT&RUN experiment has worked. I examine the literature and identified SAT2 as a gene which is commonly present in heterochromatin, and has been used as a positive control for the presence of H3K27me3

(116). Due to the absence of positive control for JARID2, the enrichment of JARID2 target-sites was not assessed via qPCR.

Quantitative PCR was performed on the enriched H3K4me3 reaction, H3K27me3 reaction, IgG reaction and sonicated input DNA using SYBR-green. The reason for using two control samples (i.e. input DNA and IgG) was to determine which one is better to be used as control to normalize the signal of CUT&RUN reaction and assesses the significance of enrichment, so that regions with high level of background binding (i.e. artifacts) can be ignored during analysis.

The amount of DNA released by an immune-linked enzyme of each sample is represented as a signal in relation to the overall amount of chromatin input, which is equal to 1 (Percent input for 100,000 cells) as shown in **Figure 6.3a**. The data are normalized using spike-in DNA added to each reaction. The qPCR analysis revealed a successful enrichment of H3K4me3 in both replicates, suggesting that the protocol had been completed correctly (**Figure 6.2a and b**). H3K27me3 enrichment was assessed only in the second GBM63 replicate due to the delay in receiving the positive control for this mark. A slight enrichment of H3K27me3 was observed in comparison with input sample as shown in **Figure 6.3b**. The lack of better enrichment for the H3K27me3 experiment may be due to: 1) improper binding of H3K27me3 antibody, 2) Lack of universality of the H3K27me3 mark within the SAT2 gene.

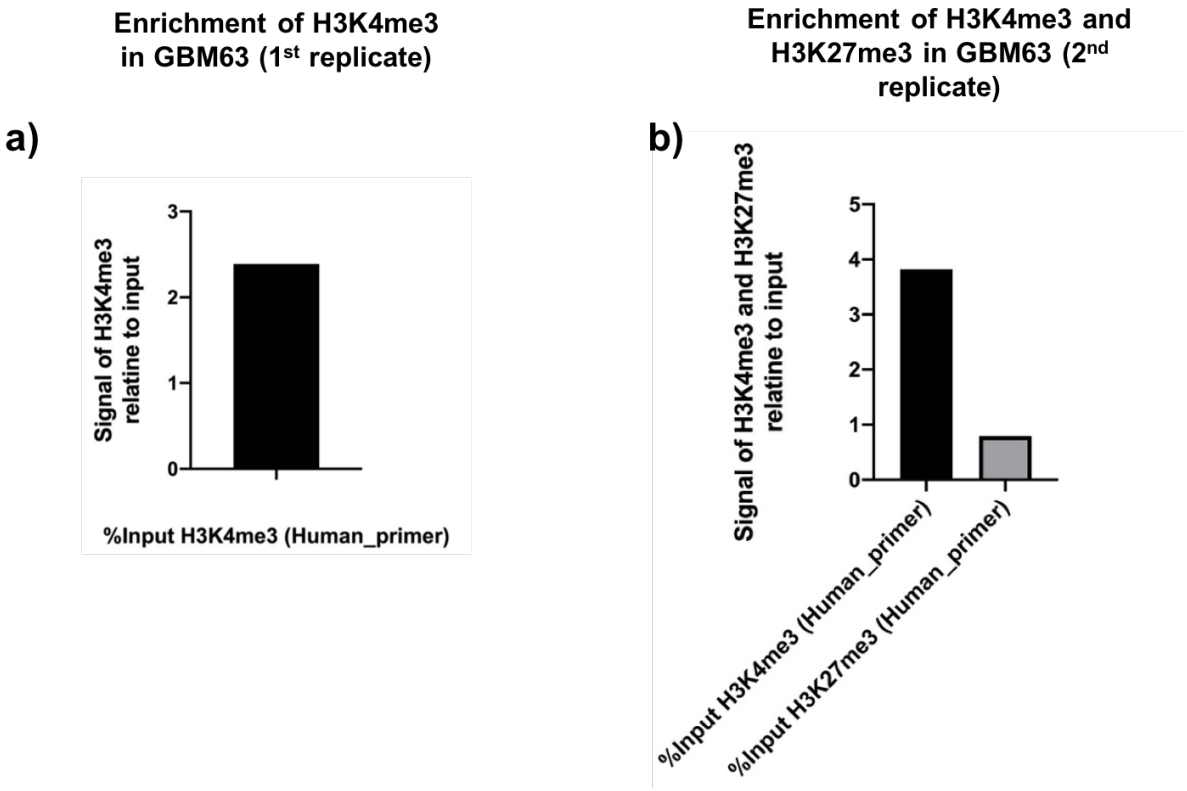


Figure 6-3: Enrichment of H3K4me3 and H3K27me3 relative to the total amount of input chromatin in GBM63 replicates using qPCR. Enriched chromatin sample and input DNA from CUT&RUN assay was amplified and quantified using qPCR with SYBR-green. A successful amplification of H3K4me3 was observed in both replicates **(a&b).** A lower amplification was observed for H3K27me3 in the second replicate **(b)**. A primer set of RPL30 and SAT2 were used as a positive control for H3K4me3 and H3K27me3 respectively. The amount of immunoprecipitated DNA in each sample is represented as signal relative to the total amount of input chromatin, which is equivalent to 1 (Percent input for 100,000 cells). The data are normalized using spike-in DNA added to each CUT&RUN reaction.

6.2.3 DNA sequencing and CUT&RUN data analysis

Prior to Library preparation, all purified DNA samples of H3K4me3, H3K27me3, JARID2 and input DNA were quantified initially using Quant-iT kit, however, none of the samples had any readable DNA. The samples were then re-quantified on Qubit HS kit which is highly selective

for double stranded DNA (dsDNA) over single stranded DNA. (ssDNA), protein, RNA and free nucleotides. Only the input samples seem to have a reading and the remaining samples had a yield which is too low to read on Qubit kit (See Appendix F). On investigation, I realised that this is considered normal as a successful CUT&RUN experiment can often yield less than 5ng of DNA from 100,000 starting materials. As mentioned above, the recommendation based on CUT&RUN kit in this case is to increase the number of PCR amplification cycles to 12-15 cycles in order to generate a library with DNA concentration of 10-30 ng/µl. So, I followed this recommendation and I used 5ng of the samples that had a readable concentration, diluting down the samples that had higher concentrations and used all of the volume of the remaining samples that have no or very low ng/ul readings for library preparation. Additionally, I used 12 cycles for PCR amplification during library preparation as recommended. Bioanalyzer traces were used to assess the size of the DNA fragments and only the input samples had an average size of fragments around 250bp (See Appendix F) as the remaining samples has low or no readable concentrations as described above. Therefore, the library was prepared by assuming that all samples had an average fragment size of 250 bp. The final libraries were cleaned up twice and then checked again on Bioanalyzer. The results indicated the presence of adapter dimer contamination peaks in 6 samples. The samples were pooled at 5x the concentration and the sequencing performed on the NextSeq 550 MO platform. The input data was formatted in FASTQ format and paired-end sequencing reads with 76 bp were generated for each sample.

The data was analysed according to the developed ChIP-seq pipeline as described in **Chapter 2, section 2.2-8**. In brief, an initial quality assessment of the FASTQ files was performed using FASTQC to check for poor quality reads and adapter content. All samples had good quality scores with a highest peak observed at 35 which means that Q35 have more read number than other quality scores. The samples were then quality and adapter trimmed using cutadapt and then checked again for the adapter content. No adapter content was found after the trimming step. The trimmed data was then aligned to human reference genome using bwa-mem with default parameters. I then employed samtools to compute the alignment statistics as shown in **Table 6.2**. The alignment percentages for all samples ranged between 83% - 99%, except for one sample which has an alignment percentage < 80%.

Sample name	Total number of reads	Total number of mapped reads	Alignment percentages
GBM63_P17 (1st replicate)			
GBM63_Input	7498080	7406085	98.77%
GBM63_IgG	10731027	7847117	73.13%
GBM63_H3K4me3	8781117	8188563	93.25%
GBM63_H3K27me3	32225150	31504004	97.76%
GBM63_JARID2	9358396	8410477	89.87%
GBM63_P18 (2nd replicate)			
GBM63_Input	11826068	11716488	99.07%
GBM63_IgG	10846359	9405655	86.72%
GBM63_H3K4me3	9513035	8817327	92.69%
GBM63_H3K27me3	42247651	40259617	95.29%
GBM63_JARID2	9752249	8363500	85.76%

Table 6-2: Mapping statistics of the analysed replicated samples of GBM63 cell lines.
Table summarizes the mapping statistics of GBM63_P17 and GBM63_P18 samples.

Mapped reads were post-processed using picard tools to remove unmapped, non- primary alignments, reads with low mapping quality (MAPQ > 25) and duplicated reads. The quality metrics in terms of fragment size distribution, adapter content percentage, library size, read duplication rate were assessed. With regards to adapter content, the percentage of adapter content in each sample is less than 5%. The library size for each sample is > 20 million and the read duplication rate is < 10% and this is ideal for CUT&RUN data. In general, these samples fulfil almost all metrices except the alignment percentages which is lower than the suggested guidelines (**See section 6.1.3**).

6.2.4 Peak detection is increased when CUT&RUN reaction and Input DNA sample as control are combined

Enriched regions (peaks) for H3K27me3, H3K4me3 and JARID2 were identified using MACS2 pairing each CUT&RUN reaction with its input DNA sample as controls. The peaks were called first using the input DNA sample as control because the pipeline was optimized and validated for the use of input DNA sample as control. In addition, input DNA sample was recommended by ENCODE and was cited many times to be used as control to identify the significant peak regions and filtering out false-positive signals. JARID2 and H3K27me3 peaks were called using the default parameters for broad peaks (i.e. broad peak cut-off value of 0.1) as described in **Chapter 2, Section 2.2.8** and the total number of significant peaks was reported (**Table 6.3**). H3K4me3 peaks were identified using the default parameters for narrow peak with a q-value of 0.01. The results revealed the presence of a lower number of H3K27me3 and JARID2 peaks and a higher number of H3K4me3 peaks.

Sample	**H3K4me3**	**H3K27me3**	**JARID2**
GBM63_1st replicate	18047	1613	4
GBM63_2nd replicate	15095	3619	10

Table 6-3: Summary of the enriched peaks of the GBM63 cell lines (replicates) called by MACS2.
Table includes the number of H3K4me3, H3K27me3 and JARID2 peaks for GBM63 replicates resulted from MACS2

Due to the lower number of obtained H3K27me3 and JARID2 peaks, I tried to optimize MACS2 parameters and I used only H3K4me3 samples for this purpose because of the success of CUT&RUN experiment. I re-called the peaks using the following conditions: IgG as control and a q-value of 0.1, IgG as control and a q-value of 0.05 and the input as control and a q-value of 0.1 to see if I can get a higher number of peaks and all these tested parameters resulted in higher number of peaks in comparison to the default parameters (**Table 6.4**).

Sample	H3K4me3 peaks
Number of peaks with IgG as control and q-value of 0.1	
GBM63_1st replicate	20563
GBM63_2nd replicate	17146
Number of peaks with IgG as control and q-value of 0.05	
GBM63_1st replicate	20358
GBM63_2nd replicate	16615
Number of peaks with input as control and q-value of 0.1	
GBM63_1st replicate	20664
GBM63_2nd replicate	17496

Table 6-4: Comparison of MACS2 peaks for H3K4me3 samples using different q-value cut-off and controls
Table summarizes the number of H3K4me3 peaks for GBM63 replicates with different q-value and control samples

In view of these results, I decided to use the input as control and a q-value of 0.1 to call the peaks for H3K4me3. I then attempted to use SEACR as an alternative tool to call peaks and enriched regions, to verify whether it gave improved results, but it resulted in many fewer peaks (**Table 6.4**). Stringent thresholding, and input sample as control, were used to call peaks. This finding proves the capability of MACS2 to call peaks with high resolution and evaluate the significance of the enriched region. Additionally, MACS2 can accurately capture local biases in the genome sequence enabling more sensitive and robust prediction.

Sample	H3K4me3	H3K27me3	JARID2
GBM63_1st replicate	2	12	6
GBM63_2nd replicate	0	395370	10

Table 6-5: Summary of the enriched peaks of the GBM63 cell lines (replicates) using SEACR
Table includes H3K4me3, H3K27me3 and JARID2 peaks for GBM63 replicates resulted from SEACR

Taking the qPCR and peak calling results together, it seems that CUT&RUN experiment worked successfully for H3K4me3 but not for H3K27me3 or JARID2 and this might be due to the non-specificity of the antibody used in this experiment which affects their ability to bind to chromatin associated protein or the MNase digestion as described above.

6.2.5 Consistent CUT&RUN enrichments of H3K4me3 was observed across biological replicates using IDR measures

The significance of the resulting H3K4me3 peaks across the two biological replicates was further assessed using the irreproducible discovery rate (IDR) as recommended by ENCODE (The description of this test can be found in **Chapter 2, Section 2.2.8**). Using the input sample as control, a q-value of 0.1, and IDR threshold of 0.05, a high reproducibility between H3K4me3 replicates were observed (**Figure 6.4**). The upper left figure plots the peak ranks in replicate 1 against the peak ranks in the second replicate, showing that a high proportion of the called peaks are replicated i.e., that they pass the specified IDR threshold (points colored in black). Peaks that do not pass the IDR threshold are represented by points colored in red. The upper right plot is analogous to the upper left one but with the log10 score plotted for peaks in both replicates. The bottom plots show the peak rank versus IDR scores for each replicate separately, with boxplots showing the IDR distribution in each 5% quantile.

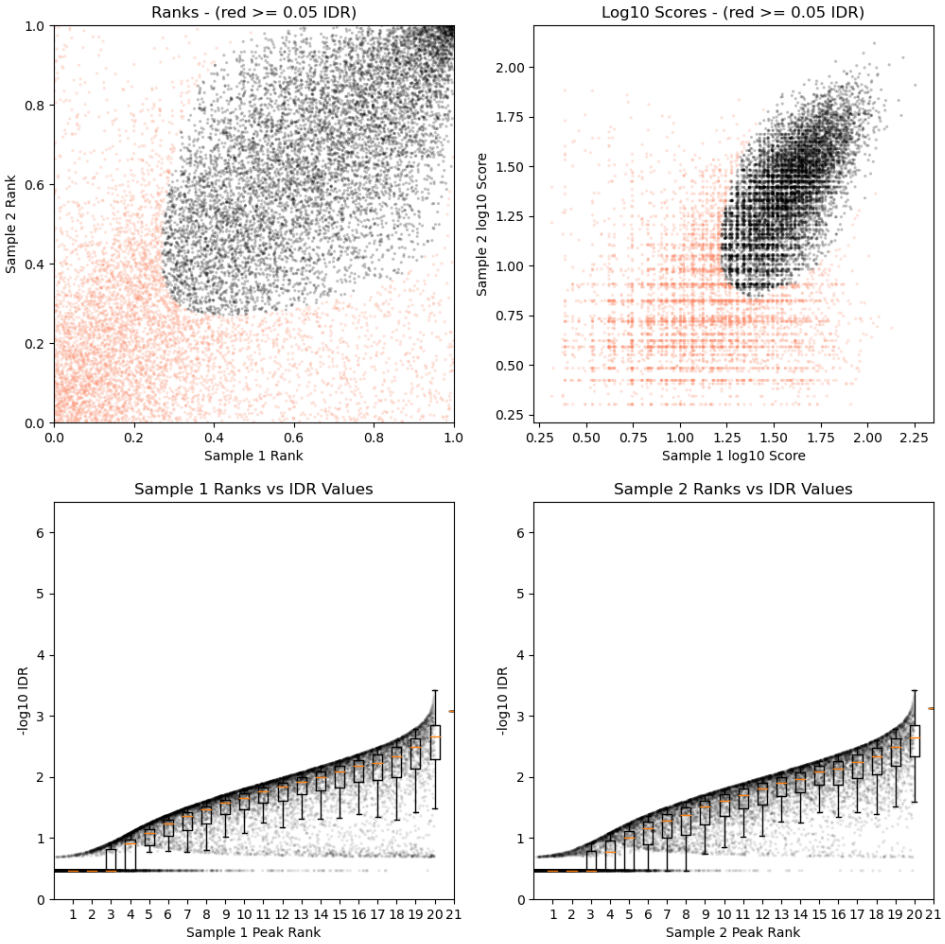


Figure 6-4: An IDR plot of called peaks in GBM63 replicates. High reproducibility was observed for these two replicates using MACS2 parameter of q-value of 0.1 and input sample as control. Peak ranks for the 1st replicate versus the 2nd replicates (Upper left) and log10 peak ranks of replicate 1 versus replicate 2 (Upper right). Peak versus IDR scores was also plotted (Bottom plots). An Idr threshold of 0.05 was used.

Due to the lower enrichment of H3K27me3 and JARID2, these samples were not processed for any further analysis such as promoter calling pipeline. In general, the optimization and the implementation of CUT&RUN was successful for H3K4me3 in the selected cell lines. These results indicated that the success of CUT&RUN experiment is mainly depends on the quality of the antibody used. It also showed that the quality of the antibodies in terms of sensitivity and specificity remains still a major issue.

6.3 Failure of CUT&RUN experiment on fresh frozen patient tumors due to the limited size of the tissue sample

In parallel with the cell lines, I tried to optimize CUT&RUN assay to generate a genome-wide mapping of chromatin states from fresh frozen patient tumors. I started with two independent samples as described in **Chapter 2, Section 2.2.7.1**. The tissues were harvested and the cells were prepared for counting to collect 100,000 cells for each reaction and input sample, however, I was unable to see any cells and only fibers and tissue appeared on the hemocytometer. Therefore, I divided the harvested suspension equally for each reaction and the input. I proceeded with CUT&RUN assay and library preparation as outlined above and as I described in **Chapter 2, Section 2.2.7.8**. Purified DNA of the anti-H3K4me3 reaction applied to the first tissue sample (NB17/39) was quantified by qPCR and the results showed a significant enrichment of the RPL30 gene in comparison with the input sample, suggesting that the experiment had worked (**Figure 6.5**).

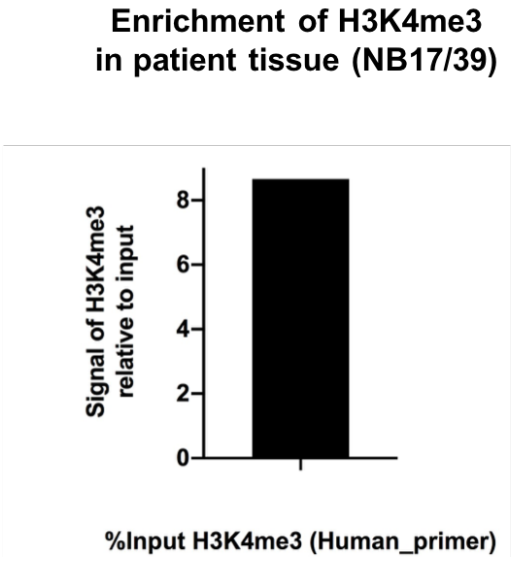


Figure 6-5: Amplification of H3K4me3 relative to the total amount of input chromatin in fresh frozen patient tumours (NB17/39) using qPCR. Enriched H3K4me3 sample and input DNA from CUT&RUN assay was amplified and quantified using qPCR with SYBR-green. A successful amplification of H3K4me3 was observed. A primer set of RPL30 was used as a positive control for H3K4me3. The amount of immunoprecipitated DNA in each sample is represented as signal relative to the total amount of input chromatin, which is equivalent to

1 (Percent input for 100,000 cells). The data are normalized using spike-in DNA added to each reaction.

A qPCR for the second tissue was not performed due to the COVID-19 pandemic and I decided to proceed with library preparation for the sake of time. The quality of the purified DNA was checked as described above, using Qubit HS kit and Bioanalyzer HS chip (See Appendix F). The results indicated the presence of DNA for the input sample, but the CUT&RUN reactions had very low or no readable DNA. Final libraries were assessed on the Bioanalyzer to check for presence of adapter dimer peaks or poor-quality libraries. The results revealed the presence of adapter dimer peaks. The samples were pooled at 5x the concentration and it were sequenced on the NextSeq 550 MO platform. 76bp paired-end reads in FASTQ format were generated for each sample.

The data was analyzed and processed according to the proposed ChIP-seq and CUT&RUN pipeline (**See Chapter 2, Section 2.2.8.1 and 2.2.8.2.3 for the detail description of the proposed pipelines**). FASTQ files was passed through FASTQC to perform an initial check on the quality of the sequencing data. The samples passed all the quality checks except the adapter content parameter. FASTQ files were quality and adapter trimmed using cutadapt and then was checked again using FASTQC, to ensure the removal of adapters. The samples were aligned to the human reference genome and the alignment statistics were evaluated. An alignment percentage of more than 90% was reported for all samples except: the IgG samples for both tissues; and H3K4me3, H3K27me3 and JARID2 for the second tissue. The alignment percentages for these samples ranged between 45% - 70% (**Table 6.5**).

Sample name	Total number of reads	Total number of mapped reads	Alignment percentages
NB17/39_Input	10770403	10664358	99.02%
NB17/39_IgG	6633664	5564686	83.89%
NB17/39_H3K4me3	8186191	7523515	91.90%
NB17/39_H3K27me3	38330933	36864035	96.17%
NB17/39_JARID2	4768810	4526159	94.91%

NB169/12_Input	9100242	8964059	98.50%
NB169/12_IgG	12069552	6452046	53.46%
NB169/12_H3K4me3	12757653	7126473	55.86%
NB169/12_H3K27me3	54174393	38420467	70.92%
NB169/12_JARID2	14719974	6671350	45.32%

Table 6-6: Mapping statistics of the analysed fresh frozen patient tumours.
Table includes the mapping statistics of NB17/39 and NB169/12 samples in terms of total number of reads, total number of mapped reads and the alignment percentages.

The samples were post-processed as described in **Chapter 2, Section 2.2.8** and the quality metrices were assessed according to CUT&RUN assay guidelines. The samples passed all parameters except the alignment percentage. Peaks were then called for the two histone marks and JARID2 and the results indicated the presence of very low number of peaks for all samples (**Table 6.6**).

Sample	**H3K4me3**	**H3K27me3**	**JARID2**
NB17/39	9544	2	9
NB169/12	1	10	1

Table 6-7: Summary of the enriched peaks of fresh frozen patient tumours by MACS2
Table includes the number of H3K4me3, H3K27me3 and JARID2 peaks for NB17/39 and NB169/12 samples resulted from MACS2

This finding indicated the failure of CUT&RUN experiment on these two tissues. This might be due to the fact that tissues are often quite limited in terms of biological materials and sample size, or that the cells were not intact. Alternatively, the protocol may not have worked on isolated nuclei (though these may also not be intact). It was impossible to proceed with the

promoter calling pipeline and therefore, I was unable to profile the chromatin landscape for the two histone marks in these tissues.

6.4 Discussion

Numerous biological processes such as regulation of gene expression, DNA replication, transcription, packaging of chromosomal DNA and disease states are governed by protein-DNA interactions. This epigenetic information is complementary to genotype and expression analysis (235). Tremendous efforts have been made to understand these interactions, and as a result, genome wide mapping of transcription factor binding sites, chromatin-associated complex and chromatin states including histone variants and post-translational modifications has become a major focus of research (75). For over 30 years, ChIP-seq has been the powerful and predominant tool of mapping protein-DNA interactions. However, the requirement of high amounts of starting material, the high background signal, which limits sensitivity, artifacts resulting from cross linking and solubilization, the GC bias in the fragments, poor resolution and high sequencing cost remain major limitations of this technique (131, 133, 134).

Efforts have been made by researchers to overcome these issues, which led to the development of CUT&RUN. Since its introduction, the advantages of CUT&RUN over ChIP-seq has facilitated rapid profiling of protein-DNA complexes with high resolution (248). Here, I decided to try and take advantage of this technique. I performed CUT&RUN on 2 biological replicates of a patient derived GBM cell line (GBM63) and two fresh frozen GBM patient tumors to profile the histone modifications and define the chromatin states that drive the occurrence of the disease. Peak analysis demonstrated the failure of the CUT&RUN experiment for both the cell lines and the tissue samples. The assay worked well in cell lines for the histone mark H3K4me3, for which the antibody is provided with the kit as a positive control. The number of the obtained H3K27me3 and JARID2 peaks were substantially lower compared to those that have been seen in glioma as reported in the literature, though these reported results are all from studies using ChIP-seq as there is a lack of availability of CUT&RUN information about the number of H3K27me3 and JARID2 peaks (116, 256, 257). In one study, the number of H3K27me3 ChIP-seq peaks that was found in the primary and K27M mutant Pediatric high-grade glioma were 21,217 and 15,853 peaks respectively (257). In

another study, the number of ChIP-seq peaks of H3K27me3 in SN186 glioma stem cell line was 5965 peaks (116). With regards JARID2: 3644 and 4916 peaks were found in a liver carcinoma cell lines (HepG2) and a normal liver cell line (THLE-2), respectively, using MACS version 1.4 (256).

CUT&RUN has been used widely in a number of research projects to profile transcription factor binding sites and the histone modifications. In humans, it was used to identify the HGATAA GATA1 recognition motif using GATA1 antibodies. GATA1 is related to a master regulator in erythroid lineage cells. The results prove the ability of CUT&RUN to identify this motif correctly (249). In addition, it was applied on K562 cells using anti-MAX and ant-MYC, which was performed previously with ChIP-seq experiment, and the results were comparable with ChIP-seq findings (258). Furthermore, as expected, CUT&RUN provided a high resolution in identifying these motif sites. This technique was applied recently on budding yeast and the group was able to profile histone co-occupancies genome-wide with high efficiency and resolution (259). Moreover, genome-wide mapping of histone modifications in *Plasmodium falciparum* was successfully generated using CUT&RUN.

I concluded that the failure of CUT&RUN experiments in cell lines was mainly due to the quality and non-specificity of the selected H3K27me3 and JARID2 antibodies for CUT&RUN. With regards to tissues, the failure is mainly due to the fact that that tissues are often quite limited in terms of biological materials and size. In general, and as with ChIP-seq, CUT&RUN efficiency is mainly depending on the amount and quality of the starting materials and on the quality of the antibody used in the experiment.

Chapter 7

Final thesis discussion

7.1 Summary of key findings

Glioblastoma is the most aggressive and the most common brain tumour in adults with a median survival time of 14-20 months from initial diagnosis. The current standard treatment strategies consist of maximal surgical debulking followed by radiotherapy and chemotherapy (35, 36). The use of radiotherapy in combination with chemotherapy have relatively little effect on survival with a median overall survival increase of only approximately 2.5 months (260). This poor prognosis could be explained by the presence of subpopulations of cells that infiltrate into the surrounding brain parenchyma and serve as an origin of recurrence. These cells are typically resistant to the initial therapy because they exhibit tremendous cellular and molecular heterogeneity and have further evolved from the primary tumour to be even more treatment resistant (3, 4).

Numerous studies have demonstrated that tumour heterogeneity which encompasses complicated genetic alterations, epigenetic abnormalities, growth rate, protein modification, and apoptosis is the primary factor in tumour recurrence and progression (10, 66). Like other cancers, GBM harbours several genetic mutations that disrupt pathways related to cancer (68). Recent advances in molecular and genetics profiling and characterization of tumours have led to the identifications of new targetable approaches that targets several molecular changes and pathways alterations such as EGFR mutations, TP53 mutations and PTEN mutations. However, these scientific developments have not yet be proven to be successful in preventing the recurrence of GBM, thus failed to improve the prognosis of GBM patients (89, 261, 262). Deeper understanding of the mechanisms that drive GBM resistance is the key to improving GBM treatment and is essential in designing more effective therapeutic strategies (262).

GBM recurrence and gliomagenesis have been extensively studied in an effort to identify and understand the mechanisms that potentiate GBM's aggressiveness (68). The Cancer Genome Atlas (TCGA) research network carried out whole genome sequencing of GBM tumours, and discovered a connection between epigenetic phenomena and GBM progression (20).

Epigenetic modifications have been recently found to be involved in the tumorigenesis of almost all cancers, including GBM (81, 101). Numerous studies suggested that these alterations are essential for controlling DNA accessibility and chromatin structure, which regulates gene expression. Since chromatin is thought to be the major transcriptional impediment, it will be crucial to comprehend chromatin shape and how epigenetic changes modify it in order to target epigenetic pathways linked to tumour resistance (76, 78). The two most prevalent epigenetic processes associated with all malignancies are DNA methylation and histone alterations. The most extensively researched epigenetic alteration in cancer, and notably in GBM, is DNA methylation. In contrast, little is known about how histone changes affect chromatin state and the regulation of gene expression in cancer (69). In order to acquire a greater understanding of the intricate interaction of various epigenetic changes in cellular processes, recent research has been increasingly focused on the function of post-translational modifications (PTMs) of histones in cancer and, especially, in GBM (98). Despite the growing volume of knowledge regarding GBM, little is understood about how the epigenome aids in the GBM progression, and the precise epigenetic pathways behind therapeutic resistance in GBM still need to be clarified. Because there is limited data on this field, there is a growing interest in creating a genome-wide histone modification map for gliomas (10, 89).

Recent developments in high-throughput technologies have enabled researchers to precisely identify locations (and the associated, affected genes) of histone modifications, and the coordinators thereof, which coordinate gene expression. ChIP-seq is one of the powerful tools for mapping and identifying global genome-wide patterns of these modifications (117, 123). The present study aimed to understand the epigenetic mechanisms involved in GBM resistance and recurrence, as recent work conducted by our group has shown that transcriptional changes occur dynamically after treatment in GBM. I hypothesized that histone remodelling is driving the changes in the gene expression observed in GBM through treatment. To investigate this, a genome-wide profiling of H3K4me3, H3K27me3 and EZH2 by ChIP-seq for matched primary and recurrent GBM samples was generated. My findings suggest that bivalency is key to enabling GBM tumour cells to adapt to treatment. Bivalency refers to the regions that marked with both the repressive mark H3K27me3 and the active mark H3K4me3, keeping gene expression repressed but poised for transcription (263).

Bivalent regions were first found in the developmental gene promoters of embryonic stem cells (ESCs), but they have also been observed in cancer cells that have stem cell-like characteristics including glioma stem cells (GSCs) (116, 241). Recent researches pointed out toward the role of bivalent genes in the heterogeneity and plasticity of different types of cancers including gliomas (241, 264). Consistent with this, I found that JARID2 binding genes stay bivalent through treatment or acquire the bivalency through state transition and this bivalency promote tumour recurrence in GBM patients. This finding highlights the role of JARID2 genes in GBM recurrence and chemo-resistance.

Collectively, bivalent genes and their underlying processes will likely be more understood in the future, opening up new prospects for the development of patient-specific and selective treatment methods based on personalized and precision medicine, as well as for the identification of novel biomarkers for the diagnosis and progression of illness. In addition, JARID2 can be used as a novel therapeutic target for the treatment of GBM patients.

7.2 Future work and directions in GBM

This study can be further expanded in a number of different ways. Firstly, there is a need to validate the specificity and the sensitivity of JARID2 antibodies for the intended application which is here, CUT&RUN. One way to further validate the selected antibodies is through immunoprecipitation of the target protein followed by mass spectrometry (MS). This approach is regarded as the gold standard for identifying and measuring a specific set of proteins in a sample. Mass spectrometry is the only validation technique that can specifically identify the antibody target(s), isoforms, post-translational modifications, and target-associated proteins that are present in a sample. Only MS is capable of characterizing antibodies with this degree of specificity and depth (265, 266).

Second, ChIP-seq or CUT&RUN experiments should be performed to profile the binding of JARID2 in GBM in order to closely examine its role in treatment resistance mechanisms in GBM patients. I had planned to profile the binding of JARID2 but the limiting factor was the amount of tissue available so I was able only to profile the two histone marks (i.e. H3K4me3 and H3K27me3) along with EZH2 binding. These were prioritised because the company we

used (owing to the COVID pandemic meaning our labs were shut) had in-house validated antibodies for these protein (modifications) but not for JARID2.

Third, to further study the link between histone modifications and changes in gene expression, there is a need to look into histone modifiers and to show that when histone methylation is stopped, it stops the gene expression changes, in order to provide a causal link.

Bibliography

1. Ramirez YP, Weatherbee JL, Wheelhouse RT, Ross AH. Glioblastoma multiforme therapy and mechanisms of resistance. Pharmaceuticals (Basel, Switzerland). 2013;6(12):1475-506.
2. Paw I, Carpenter RC, Watabe K, Debinski W, Lo HW. Mechanisms regulating glioma invasion. Cancer letters. 2015;362(1):1-7.
3. Manini I, Caponnetto F, Bartolini A, Ius T, Mariuzzi L, Di Loreto C, et al. Role of Microenvironment in Glioma Invasion: What We Learned from In Vitro Models. International journal of molecular sciences. 2018;19(1).
4. Hatoum A, Mohammed R, Zakieh O. The unique invasiveness of glioblastoma and possible drug targets on extracellular matrix. Cancer management and research. 2019;11:1843-55.
5. Bergmann N, Delbridge C, Gempt J, Feuchtinger A, Walch A, Schirmer L, et al. The Intratumoral Heterogeneity Reflects the Intertumoral Subtypes of Glioblastoma Multiforme: A Regional Immunohistochemistry Analysis. Frontiers in oncology. 2020;10:494.
6. Dymova MA, Kuligina EV, Richter VA. Molecular Mechanisms of Drug Resistance in Glioblastoma. International journal of molecular sciences. 2021;22(12).
7. Wen J, Chen W, Zhu Y, Zhang P. Clinical features associated with the efficacy of chemotherapy in patients with glioblastoma (GBM): a surveillance, epidemiology, and end results (SEER) analysis. BMC cancer. 2021;21(1):81.
8. Grochans S, Cybulska AM, Simińska D, Korbecki J, Kojder K, Chlubek D, et al. Epidemiology of Glioblastoma Multiforme-Literature Review. Cancers. 2022;14(10).
9. Nizamutdinov D, Stock EM, Dandashi JA, Vasquez EA, Mao Y, Dayawansa S, et al. Prognostication of Survival Outcomes in Patients Diagnosed with Glioblastoma. World neurosurgery. 2018;109:e67-e74.
10. Hanif F, Muzaffar K, Perveen K, Malhi SM, Simjee Sh U. Glioblastoma Multiforme: A Review of its Epidemiology and Pathogenesis through Clinical Presentation and Treatment. Asian Pacific journal of cancer prevention : APJCP. 2017;18(1):3-9.
11. Ohgaki H, Kleihues P. Genetic pathways to primary and secondary glioblastoma. The American journal of pathology. 2007;170(5):1445-53.
12. Crespo I, Vital AL, Gonzalez-Tablas M, Patino Mdel C, Otero A, Lopes MC, et al. Molecular and Genomic Alterations in Glioblastoma Multiforme. The American journal of pathology. 2015;185(7):1820-33.
13. Hasanzadeh N, Niknejad A. Cerebral Glioblastoma: A Review on Genetic Alterations, Signaling Pathways, and Clinical Managements. Jentashapir J Cell Mol Biol. 2021;12(4):e119223.
14. Khabibov M, Garifullin A, Boumber Y, Khaddour K, Fernandez M, Khamitov F, et al. Signaling pathways and therapeutic approaches in glioblastoma multiforme (Review). International journal of oncology. 2022;60(6).
15. Pan PC, Magge RS. Mechanisms of EGFR Resistance in Glioblastoma. International journal of molecular sciences. 2020;21(22).
16. Li X-P, Guo Z-Q, Wang B-F, Zhao M. EGFR alterations in glioblastoma play a role in antitumor immunity regulation. Frontiers in oncology. 2023;13.
17. Xu H, Zong H, Ma C, Ming X, Shang M, Li K, et al. Epidermal growth factor receptor in glioblastoma. Oncology letters. 2017;14(1):512-6.
18. Liu X, Chen X, Shi L, Shan Q, Cao Q, Yue C, et al. The third-generation EGFR inhibitor AZD9291 overcomes primary resistance by continuously blocking ERK signaling in glioblastoma. Journal of Experimental & Clinical Cancer Research. 2019;38(1):219.

19. Zhang Y, Dube C, Gibert M, Jr., Cruickshanks N, Wang B, Coughlan M, et al. The p53 Pathway in Glioblastoma. Cancers. 2018;10(9).
20. Clarke J, Penas C, Pastori C, Komotar RJ, Bregy A, Shah AH, et al. Epigenetic pathways and glioblastoma treatment. Epigenetics. 2013;8(8):785-95.
21. Pearson JRD, Regad T. Targeting cellular pathways in glioblastoma multiforme. Signal Transduction and Targeted Therapy. 2017;2(1):17040.
22. Kato S, Ross JS, Gay L, Dayyani F, Roszik J, Subbiah V, et al. Analysis of MDM2 Amplification: Next-Generation Sequencing of Patients With Diverse Malignancies. JCO Precision Oncology. 2018(2):1-14.
23. Chen J. The Cell-Cycle Arrest and Apoptotic Functions of p53 in Tumor Initiation and Progression. Cold Spring Harbor perspectives in medicine. 2016;6(3):a026104.
24. Nishikawa S, Iwakuma T. Drugs Targeting p53 Mutations with FDA Approval and in Clinical Trials. Cancers. 2023;15(2).
25. Pellot Ortiz KI, Rechberger JS, Nonnenbroich LF, Daniels DJ, Sarkaria JN. MDM2 Inhibition in the Treatment of Glioblastoma: From Concept to Clinical Investigation. Biomedicines. 2023;11(7).
26. Knudsen ES, Wang JY. Targeting the RB-pathway in cancer therapy. Clinical cancer research : an official journal of the American Association for Cancer Research. 2010;16(4):1094-9.
27. Biasoli D, Kahn SA, Cornélio TA, Furtado M, Campanati L, Chneiweiss H, et al. Retinoblastoma protein regulates the crosstalk between autophagy and apoptosis, and favors glioblastoma resistance to etoposide. Cell Death & Disease. 2013;4(8):e767-e.
28. Chkheidze R, Raisanen J, Gagan J, Richardson TE, Pinho MC, Raj K, et al. Alterations in the RB Pathway With Inactivation of RB1 Characterize Glioblastomas With a Primitive Neuronal Component. Journal of Neuropathology & Experimental Neurology. 2021;80(12):1092-8.
29. Goldhoff P, Clarke J, Smirnov I, Berger MS, Prados MD, James CD, et al. Clinical stratification of glioblastoma based on alterations in retinoblastoma tumor suppressor protein (RB1) and association with the proneural subtype. Journal of neuropathology and experimental neurology. 2012;71(1):83-9.
30. Paternot S, Colleoni B, Bisteau X, Roger PP. The CDK4/CDK6 inhibitor PD0332991 paradoxically stabilizes activated cyclin D3-CDK4/6 complexes. Cell cycle (Georgetown, Tex). 2014;13(18):2879-88.
31. Cadoo KA, Gucalp A, Traina TA. Palbociclib: an evidence-based review of its potential in the treatment of breast cancer. Breast cancer (Dove Medical Press). 2014;6:123-33.
32. Guan R, Zhang X, Guo M. Glioblastoma stem cells and Wnt signaling pathway: molecular mechanisms and therapeutic targets. Chinese Neurosurgical Journal. 2020;6(1):25.
33. Lee Y, Lee J-K, Ahn SH, Lee J, Nam D-H. WNT signaling in glioblastoma and therapeutic opportunities. Laboratory Investigation. 2016;96(2):137-50.
34. Tan AC, Ashley DM, López GY, Malinzak M, Friedman HS, Khasraw M. Management of glioblastoma: State of the art and future directions. CA: A Cancer Journal for Clinicians. 2020;70(4):299-312.
35. Noch EK, Ramakrishna R, Magge R. Challenges in the Treatment of Glioblastoma: Multisystem Mechanisms of Therapeutic Resistance. World neurosurgery. 2018;116:505-17.
36. Osuka S, Van Meir EG. Overcoming therapeutic resistance in glioblastoma: the way forward. The Journal of clinical investigation. 2017;127(2):415-26.
37. Aum DJ, Kim DH, Beaumont TL, Leuthardt EC, Dunn GP, Kim AH. Molecular and cellular heterogeneity: the hallmark of glioblastoma. Neurosurgical focus. 2014;37(6):E11.

38. Parker NR, Khong P, Parkinson JF, Howell VM, Wheeler HR. Molecular heterogeneity in glioblastoma: potential clinical implications. Frontiers in oncology. 2015;5:55.
39. Fisher R, Pusztai L, Swanton C. Cancer heterogeneity: implications for targeted therapeutics. British journal of cancer. 2013;108(3):479-85.
40. Becker AP, Sells BE, Haque SJ, Chakravarti A. Tumor Heterogeneity in Glioblastomas: From Light Microscopy to Molecular Pathology. Cancers. 2021;13(4):761.
41. Wang Y, Zhang J, Li W, Jiang T, Qi S, Chen Z, et al. Guideline conformity to the Stupp regimen in patients with newly diagnosed glioblastoma multiforme in China. Future Oncology. 2021;17(33):4571-82.
42. Bjorland LS, Fluge O, Gilje B, Mahesparan R, Farbu E. Treatment approach and survival from glioblastoma: results from a population-based retrospective cohort study from Western Norway. BMJ open. 2021;11(3):e043208.
43. Lakomy R, Kazda T, Selingerova I, Poprach A, Pospisil P, Belanova R, et al. Real-World Evidence in Glioblastoma: Stupp's Regimen After a Decade. Frontiers in oncology. 2020;10:840.
44. Singh N, Miner A, Hennis L, Mittal S. Mechanisms of temozolomide resistance in glioblastoma - a comprehensive review. Cancer drug resistance (Alhambra, Calif). 2021;4(1):17-43.
45. Arora A, Somasundaram K. Glioblastoma vs temozolomide: can the red queen race be won? Cancer biology & therapy. 2019;20(8):1083-90.
46. Zhang J, Stevens MF, Bradshaw TD. Temozolomide: mechanisms of action, repair and resistance. Current molecular pharmacology. 2012;5(1):102-14.
47. Nagasaka T, Goel A, Notohara K, Takahata T, Sasamoto H, Uchida T, et al. Methylation pattern of the O6-methylguanine-DNA methyltransferase gene in colon during progressive colorectal tumorigenesis. International journal of cancer. 2008;122(11):2429-36.
48. Lee SY. Temozolomide resistance in glioblastoma multiforme. Genes & diseases. 2016;3(3):198-210.
49. Kitange GJ, Carlson BL, Schroeder MA, Grogan PT, Lamont JD, Decker PA, et al. Induction of MGMT expression is associated with temozolomide resistance in glioblastoma xenografts. Neuro-oncology. 2009;11(3):281-91.
50. Yu W, Zhang L, Wei Q, Shao A. O(6)-Methylguanine-DNA Methyltransferase (MGMT): Challenges and New Opportunities in Glioma Chemotherapy. Frontiers in oncology. 2019;9:1547.
51. Brandt B, Németh M, Berta G, Szünstein M, Heffer M, Rauch TA, et al. A Promising Way to Overcome Temozolomide Resistance through Inhibition of Protein Neddylation in Glioblastoma Cell Lines. International journal of molecular sciences. 2023;24(9).
52. Oldrini B, Vaquero-Siguero N, Mu Q, Kroon P, Zhang Y, Galán-Ganga M, et al. MGMT genomic rearrangements contribute to chemotherapy resistance in gliomas. Nature communications. 2020;11(1):3883.
53. Uno M, Oba-Shinjo SM, Camargo AA, Moura RP, Aguiar PH, Cabrera HN, et al. Correlation of MGMT promoter methylation status with gene and protein expression levels in glioblastoma. Clinics (Sao Paulo, Brazil). 2011;66(10):1747-55.
54. Rivera AL, Pelloski CE, Gilbert MR, Colman H, De La Cruz C, Sulman EP, et al. MGMT promoter methylation is predictive of response to radiotherapy and prognostic in the absence of adjuvant alkylating chemotherapy for glioblastoma. Neuro-oncology. 2010;12(2):116-21.
55. Choi HJ, Choi SH, You SH, Yoo RE, Kang KM, Yun TJ, et al. MGMT Promoter Methylation Status in Initial and Recurrent Glioblastoma: Correlation Study with DWI and DSC PWI Features. AJNR American journal of neuroradiology. 2021;42(5):853-60.

56. Wang JB, Dong DF, Wang MD, Gao K. IDH1 overexpression induced chemotherapy resistance and IDH1 mutation enhanced chemotherapy sensitivity in Glioma cells in vitro and in vivo. Asian Pacific journal of cancer prevention : APJCP. 2014;15(1):427-32.
57. Han S, Liu Y, Cai SJ, Qian M, Ding J, Larion M, et al. IDH mutation in glioma: molecular mechanisms and potential therapeutic targets. British journal of cancer. 2020;122(11):1580-9.
58. Cohen AL, Holmen SL, Colman H. IDH1 and IDH2 mutations in gliomas. Current neurology and neuroscience reports. 2013;13(5):345.
59. Barresi V, Simbolo M, Mafficini A, Martini M, Calicchia M, Piredda ML, et al. IDH-wild type glioblastomas featuring at least 30% giant cells are characterized by frequent RB1 and NF1 alterations and hypermutation. Acta Neuropathologica Communications. 2021;9(1):200.
60. Qazi MA, Vora P, Venugopal C, Sidhu SS, Moffat J, Swanton C, et al. Intratumoral heterogeneity: pathways to treatment resistance and relapse in human glioblastoma. Annals of Oncology. 2017;28(7):1448-56.
61. Burrell RA, Swanton C. Tumour heterogeneity and the evolution of polyclonal drug resistance. Molecular oncology. 2014;8(6):1095-111.
62. Machnik M, Oleksiewicz U. Dynamic Signatures of the Epigenome: Friend or Foe? Cells. 2020;9(3).
63. Brock MV, Herman JG, Baylin SB. Cancer as a manifestation of aberrant chromatin structure. Cancer journal (Sudbury, Mass). 2007;13(1):3-8.
64. Minata M, Audia A, Shi J, Lu S, Bernstock J, Pavlyukov MS, et al. Phenotypic Plasticity of Invasive Edge Glioma Stem-like Cells in Response to Ionizing Radiation. Cell reports. 2019;26(7):1893-905.e7.
65. McGranahan N, Swanton C. Clonal Heterogeneity and Tumor Evolution: Past, Present, and the Future. Cell. 2017;168(4):613-28.
66. Ramón YCS, Sesé M, Capdevila C, Aasen T, De Mattos-Arruda L, Diaz-Cano SJ, et al. Clinical implications of intratumor heterogeneity: challenges and opportunities. Journal of molecular medicine (Berlin, Germany). 2020;98(2):161-77.
67. Rippaus N, F-Bruns A, Tanner G, Taylor C, Droop A, Care MA, et al. JARID2 facilitates transcriptional reprogramming in glioblastoma in response to standard treatment. bioRxiv. 2019:649400.
68. Romani M, Pistillo MP, Banelli B. Epigenetic Targeting of Glioblastoma. Frontiers in oncology. 2018;8(448).
69. Caren H, Pollard SM, Beck S. The good, the bad and the ugly: epigenetic mechanisms in glioblastoma. Molecular aspects of medicine. 2013;34(4):849-62.
70. Bhattacharjee D, Shenoy S, Bairy KL. DNA Methylation and Chromatin Remodeling: The Blueprint of Cancer Epigenetics. Scientifica. 2016;2016:6072357.
71. Safa AR, Saadatzadeh MR, Cohen-Gadol AA, Pollok KE, Bijangi-Vishehsaraei K. Glioblastoma stem cells (GSCs) epigenetic plasticity and interconversion between differentiated non-GSCs and GSCs. Genes & diseases. 2015;2(2):152-63.
72. Kim S, Kaang BK. Epigenetic regulation and chromatin remodeling in learning and memory. Experimental & molecular medicine. 2017;49(1):e281.
73. Lafon-Hughes L, Di Tomaso MV, Mendez-Acuna L, Martinez-Lopez W. Chromatin-remodelling mechanisms in cancer. Mutation research. 2008;658(3):191-214.
74. Nowacka-Zawisza M, Wiśnik E. DNA methylation and histone modifications as epigenetic regulation in prostate cancer (Review). Oncol Rep. 2017;38(5):2587-96.
75. Maury E, Hashizume R. Epigenetic modification in chromatin machinery and its deregulation in pediatric brain tumors: Insight into epigenetic therapies. Epigenetics. 2017;12(5):353-69.

76. Keenen B, de la Serna IL. Chromatin remodeling in embryonic stem cells: regulating the balance between pluripotency and differentiation. Journal of cellular physiology. 2009;219(1):1-7.
77. Audia JE, Campbell RM. Histone Modifications and Cancer. Cold Spring Harbor perspectives in biology. 2016;8(4):a019521.
78. Quina AS, Buschbeck M, Di Croce L. Chromatin structure and epigenetics. Biochemical pharmacology. 2006;72(11):1563-9.
79. Harikumar A, Meshorer E. Chromatin remodeling and bivalent histone modifications in embryonic stem cells. EMBO reports. 2015;16(12):1609-19.
80. Fyodorov DV, Zhou BR, Skoultchi AI, Bai Y. Emerging roles of linker histones in regulating chromatin structure and function. Nature reviews Molecular cell biology. 2018;19(3):192-206.
81. Hauer MH, Gasser SM. Chromatin and nucleosome dynamics in DNA damage and repair. Genes & development. 2017;31(22):2204-21.
82. McGinty RK, Tan S. Nucleosome structure and function. Chemical reviews. 2015;115(6):2255-73.
83. Klemm SL, Shipony Z, Greenleaf WJ. Chromatin accessibility and the regulatory epigenome. Nature reviews Genetics. 2019;20(4):207-20.
84. Van HT, Santos MA. Histone modifications and the DNA double-strand break response. Cell cycle (Georgetown, Tex). 2018;17(21-22):2399-410.
85. Handy DE, Castro R, Loscalzo J. Epigenetic modifications: basic mechanisms and role in cardiovascular disease. Circulation. 2011;123(19):2145-56.
86. Clapier CR, Iwasa J, Cairns BR, Peterson CL. Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. Nature reviews Molecular cell biology. 2017;18(7):407-22.
87. Dawson MA, Kouzarides T. Cancer epigenetics: from mechanism to therapy. Cell. 2012;150(1):12-27.
88. M JD, Wojtas B. Global DNA Methylation Patterns in Human Gliomas and Their Interplay with Other Epigenetic Modifications. International journal of molecular sciences. 2019;20(14).
89. Uddin MS, Mamun AA, Alghamdi BS, Tewari D, Jeandet P, Sarwar MS, et al. Epigenetics of glioblastoma multiforme: From molecular mechanisms to therapeutic approaches. Seminars in cancer biology. 2022;83:100-20.
90. Nagarajan RP, Costello JF. Epigenetic mechanisms in glioblastoma multiforme. Seminars in cancer biology. 2009;19(3):188-97.
91. Gusyatiner O, Hegi ME. Glioma epigenetics: From subclassification to novel treatment options. Seminars in cancer biology. 2018;51:50-8.
92. Dong Z, Cui H. Epigenetic modulation of metabolism in glioblastoma. Seminars in cancer biology. 2019;57:45-51.
93. Ehrlich M. DNA hypomethylation in cancer cells. Epigenomics. 2009;1(2):239-59.
94. Kan S, Chai S, Chen W, Yu B. DNA methylation profiling identifies potentially significant epigenetically-regulated genes in glioblastoma multiforme. Oncology letters. 2019;18(2):1679-88.
95. de Souza CF, Sabedot TS, Malta TM, Stetson L, Morozova O, Sokolov A, et al. A Distinct DNA Methylation Shift in a Subset of Glioma CpG Island Methylator Phenotypes during Tumor Recurrence. Cell reports. 2018;23(2):637-51.
96. Maleszewska M, Kaminska B. Is Glioblastoma an Epigenetic Malignancy? Cancers [Internet]. 2013; 5(3):[1120-39 pp.].
97. Zhao Z, Shilatifard A. Epigenetic modifications of histones in cancer. Genome Biology. 2019;20(1):245.

98. Noberini R, Osti D, Miccolo C, Richichi C, Lupia M, Corleone G, et al. Extensive and systematic rewiring of histone post-translational modifications in cancer model systems. Nucleic acids research. 2018;46(8):3817-32.
99. Benmelouka AY, Munir M, Sayed A, Attia MS, Ali MM, Negida A, et al. Neural Stem Cell-Based Therapies and Glioblastoma Management: Current Evidence and Clinical Challenges. International journal of molecular sciences. 2021;22(5).
100. Lund AH, van Lohuizen M. Epigenetics and cancer. Genes & development. 2004;18(19):2315-35.
101. Kunadis E, Lakiotaki E, Korkolopoulou P, Piperi C. Targeting post-translational histone modifying enzymes in glioblastoma. Pharmacology & Therapeutics. 2021;220:107721.
102. Kim YZ. Altered histone modifications in gliomas. Brain tumor research and treatment. 2014;2(1):7-21.
103. Wu Q, Berglund AE, Etame AB. The Impact of Epigenetic Modifications on Adaptive Resistance Evolution in Glioblastoma. International journal of molecular sciences. 2021;22(15).
104. Yang Y, Zhang M, Wang Y. The roles of histone modifications in tumorigenesis and associated inhibitors in cancer therapy. Journal of the National Cancer Center. 2022;2(4):277-90.
105. De Majo F, Calore M. Chromatin remodelling and epigenetic state regulation by non-coding RNAs in the diseased heart. Non-coding RNA research. 2018;3(1):20-8.
106. Kumar D, Cinghu S, Oldfield AJ, Yang P, Jothi R. Decoding the function of bivalent chromatin in development and cancer. Genome research. 2021;31(12):2170-84.
107. Kumar D, Jothi R. Bivalent chromatin protects reversibly repressed genes from irreversible silencing2020.
108. Hock H. A complex Polycomb issue: the two faces of EZH2 in cancer. Genes & development. 2012;26(8):751-5.
109. Holoch D, Margueron R. Mechanisms Regulating PRC2 Recruitment and Enzymatic Activity. Trends in Biochemical Sciences. 2017;42(7):531-42.
110. Aranda S, Mas G, Di Croce L. Regulation of gene transcription by Polycomb proteins. Science advances. 2015;1(11):e1500737.
111. Stazi G, Taglieri L, Nicolai A, Romanelli A, Fioravanti R, Morrone S, et al. Dissecting the role of novel EZH2 inhibitors in primary glioblastoma cell cultures: effects on proliferation, epithelial-mesenchymal transition, migration, and on the pro-inflammatory phenotype. Clinical epigenetics. 2019;11(1):173.
112. Al-Raawi D, Jones R, Wijesinghe S, Halsall J, Petric M, Roberts S, et al. A novel form of JARID2 is required for differentiation in lineage-committed cells. The EMBO journal. 2019;38(3).
113. Sanulli S, Justin N, Teissandier A, Ancelin K, Portoso M, Caron M, et al. Jarid2 Methylation via the PRC2 Complex Regulates H3K27me3 Deposition during Cell Differentiation. Molecular cell. 2015;57(5):769-83.
114. Kaneko S, Bonasio R, Saldana-Meyer R, Yoshida T, Son J, Nishino K, et al. Interactions between JARID2 and noncoding RNAs regulate PRC2 recruitment to chromatin. Molecular cell. 2014;53(2):290-300.
115. Li G, Margueron R, Ku M, Chambon P, Bernstein BE, Reinberg D. Jarid2 and PRC2, partners in regulating gene expression. Genes & development. 2010;24(4):368-80.
116. Lin B, Lee H, Yoon JG, Madan A, Wayner E, Tonning S, et al. Global analysis of H3K4me3 and H3K27me3 profiles in glioblastoma stem cells and identification of SLC17A7 as a bivalent tumor suppressor gene. Oncotarget. 2015;6(7):5369-81.

117. van Leeuwen F, van Steensel B. Histone modifications: from genome-wide maps to functional insights. Genome Biol. 2005;6(6):113.
118. Park YJ, Claus R, Weichenhan D, Plass C. Genome-wide epigenetic modifications in cancer. Progress in drug research Fortschritte der Arzneimittelforschung Progres des recherches pharmaceutiques. 2011;67:25-49.
119. Nagarajan RP, Fouse SD, Bell RJ, Costello JF. Methods for cancer epigenome analysis. Advances in experimental medicine and biology. 2013;754:313-38.
120. Ngollo M, Lebert A, Daures M, Judes G, Rifai K, Dubois L, et al. Global analysis of H3K27me3 as an epigenetic marker in prostate cancer progression. BMC cancer. 2017;17(1):261.
121. Xi Y, Shi J, Li W, Tanaka K, Allton KL, Richardson D, et al. Histone modification profiling in breast cancer cell lines highlights commonalities and differences among subtypes. BMC genomics. 2018;19(1):150.
122. Churko JM, Mantalas GL, Snyder MP, Wu JC. Overview of high throughput sequencing technologies to elucidate molecular pathways in cardiovascular diseases. Circulation research. 2013;112(12):1613-23.
123. Reuter JA, Spacek DV, Snyder MP. High-throughput sequencing technologies. Molecular cell. 2015;58(4):586-97.
124. Heather JM, Chain B. The sequence of sequencers: The history of sequencing DNA. Genomics. 2016;107(1):1-8.
125. Qin D. Next-generation sequencing and its clinical application. Cancer biology & medicine. 2019;16(1):4-10.
126. Alekseyev YO, Fazeli R, Yang S, Basran R, Maher T, Miller NS, et al. A Next-Generation Sequencing Primer-How Does It Work and What Can It Do? Academic pathology. 2018;5:2374289518766521.
127. Soon WW, Hariharan M, Snyder MP. High-throughput sequencing for biology and medicine. Molecular Systems Biology. 2013;9(1):640.
128. Nakato R, Shirahige K. Recent advances in ChIP-seq analysis: from quality management to whole-genome annotation. Briefings in bioinformatics. 2017;18(2):279-90.
129. Mardis ER. ChIP-seq: welcome to the new frontier. Nature methods. 2007;4(8):613-4.
130. Robertson G, Hirst M, Bainbridge M, Bilenky M, Zhao Y, Zeng T, et al. Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing. Nature methods. 2007;4(8):651-7.
131. Park PJ. ChIP-seq: advantages and challenges of a maturing technology. Nature reviews Genetics. 2009;10(10):669-80.
132. Brind'Amour J, Liu S, Hudson M, Chen C, Karimi MM, Lorincz MC. An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations. Nature communications. 2015;6(1):6033.
133. Gilfillan GD, Hughes T, Sheng Y, Hjorthaug HS, Straub T, Gervin K, et al. Limitations and possibilities of low cell number ChIP-seq. BMC genomics. 2012;13(1):645.
134. Landt SG, Marinov GK, Kundaje A, Kheradpour P, Pauli F, Batzoglou S, et al. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. Genome research. 2012;22(9):1813-31.
135. Park PJ. ChIP–seq: advantages and challenges of a maturing technology. Nature Reviews Genetics. 2009;10:669.
136. Flensburg C, Kinkel SA, Keniry A, Blewitt ME, Oshlack A. A comparison of control samples for ChIP-seq of histone modifications. Frontiers in genetics. 2014;5:329.
137. Kidder BL, Hu G, Zhao K. ChIP-Seq: technical considerations for obtaining high-quality data. Nature immunology. 2011;12(10):918-22.

138. Bordeaux J, Welsh A, Agarwal S, Killiam E, Baquero M, Hanna J, et al. Antibody validation. BioTechniques. 2010;48(3):197-209.
139. Weller MG. Ten Basic Rules of Antibody Validation. Analytical chemistry insights. 2018;13:1177390118757462.
140. Shin H, Liu T, Duan X, Zhang Y, Liu XS. Computational methodology for ChIP-seq analysis. Quantitative biology (Beijing, China). 2013;1(1):54-70.
141. Mundade R, Ozer HG, Wei H, Prabhu L, Lu T. Role of ChIP-seq in the discovery of transcription factor binding sites, differential gene regulation mechanism, epigenetic marks and beyond. Cell cycle (Georgetown, Tex). 2014;13(18):2847-52.
142. Brown J, Pirrung M, McCue LA. FQC Dashboard: integrates FastQC results into a web-based, interactive, and extensible FASTQ quality control tool. Bioinformatics. 2017;33(19):3137-9.
143. Liao Y, Shi W. Read trimming is not required for mapping and quantification of RNA-seq reads at the gene level. NAR Genomics and Bioinformatics. 2020;2(3):lqaa068.
144. Li H, Durbin R. Fast and accurate short read alignment with Burrows–Wheeler transform. Bioinformatics. 2009;25(14):1754-60.
145. Langmead B, Trapnell C, Pop M, Salzberg SL. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biology. 2009;10(3):R25.
146. Fonseca NA, Rung J, Brazma A, Marioni JC. Tools for mapping high-throughput sequencing data. Bioinformatics. 2012;28(24):3169-77.
147. Li H, Ruan J, Durbin R. Mapping short DNA sequencing reads and calling variants using mapping quality scores. Genome research. 2008;18(11):1851-8.
148. Li R, Yu C, Li Y, Lam TW, Yiu SM, Kristiansen K, et al. SOAP2: an improved ultrafast tool for short read alignment. Bioinformatics. 2009;25(15):1966-7.
149. Zhang H, Song L, Wang X, Cheng H, Wang C, Meyer CA, et al. Fast alignment and preprocessing of chromatin profiles with Chromap. Nature communications. 2021;12(1):6566.
150. Carroll TS, Liang Z, Salama R, Stark R, de Santiago I. Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data. Frontiers in genetics. 2014;5:75.
151. Kolmykov SK, Kondrakhin YV, Yevshin IS, Sharipov RN, Ryabova AS, Kolpakov FA. Population size estimation for quality control of ChIP-Seq datasets. PloS one. 2019;14(8):e0221760.
152. Jeon H, Lee H, Kang B, Jang I, Roh TY. Comparative analysis of commonly used peak calling programs for ChIP-Seq analysis. Genomics & informatics. 2020;18(4):e42.
153. Oh D, Strattan JS, Hur JK, Bento J, Urban AE, Song G, et al. CNN-Peaks: ChIP-Seq peak detection pipeline using convolutional neural networks that imitate human visual inspection. Scientific reports. 2020;10(1):7933.
154. Nakato R, Sakata T. Methods for ChIP-seq analysis: A practical workflow and advanced applications. Methods (San Diego, Calif). 2021;187:44-53.
155. Zhang Y, Liu T, Meyer CA, Eeckhoute J, Johnson DS, Bernstein BE, et al. Model-based Analysis of ChIP-Seq (MACS). Genome Biology. 2008;9(9):R137.
156. Bailey T, Krajewski P, Ladunga I, Lefebvre C, Li Q, Liu T, et al. Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS computational biology. 2013;9(11):e1003326.
157. Ernst J, Kellis M. Chromatin-state discovery and genome annotation with ChromHMM. Nature protocols. 2017;12(12):2478-92.
158. Yoon BJ. Hidden Markov Models and their Applications in Biological Sequence Analysis. Current genomics. 2009;10(6):402-15.
159. Ernst J, Kellis M. ChromHMM: automating chromatin-state discovery and characterization. Nature methods. 2012;9(3):215-6.

160. Mammana A, Chung H-R. Chromatin segmentation based on a probabilistic model for read counts explains a large portion of the epigenome. Genome Biology. 2015;16(1):151.
161. Reece-Hoyes JS, Walhout AJM. Gateway Recombinational Cloning. Cold Spring Harbor protocols. 2018;2018(1):pdb.top094912.
162. Lin JS, Lai EM. Protein-Protein Interactions: Co-Immunoprecipitation. Methods in molecular biology (Clifton, NJ). 2017;1615:211-9.
163. Uhlen M, Bandrowski A, Carr S, Edwards A, Ellenberg J, Lundberg E, et al. A proposal for validation of antibodies. Nature methods. 2016;13(10):823-7.
164. Chiu ML, Goulet DR, Teplyakov A, Gilliland GL. Antibody Structure and Function: The Basis for Engineering Therapeutics. Antibodies (Basel, Switzerland). 2019;8(4).
165. Hoffman W, Lakkis FG, Chalasani G. B Cells, Antibodies, and More. Clinical journal of the American Society of Nephrology : CJASN. 2016;11(1):137-54.
166. Sela-Culang I, Kunik V, Ofran Y. The structural basis of antibody-antigen recognition. Frontiers in immunology. 2013;4:302.
167. Schroeder HW, Jr., Cavacini L. Structure and function of immunoglobulins. The Journal of allergy and clinical immunology. 2010;125(2 Suppl 2):S41-52.
168. Goulet DR, Atkins WM. Considerations for the Design of Antibody-Based Therapeutics. Journal of pharmaceutical sciences. 2020;109(1):74-103.
169. Leenaars M, Hendriksen CFM. Critical Steps in the Production of Polyclonal and Monoclonal Antibodies: Evaluation and Recommendations. ILAR Journal. 2005;46(3):269-79.
170. Lipman NS, Jackson LR, Trudel LJ, Weis-Garcia F. Monoclonal Versus Polyclonal Antibodies: Distinguishing Characteristics, Applications, and Information Resources. ILAR Journal. 2005;46(3):258-68.
171. de St. Groth SF, Scheidegger D. Production of monoclonal antibodies: Strategy and tactics. Journal of Immunological Methods. 1980;35(1):1-21.
172. Liu JK. The history of monoclonal antibody development - Progress, remaining challenges and future innovations. Annals of medicine and surgery (2012). 2014;3(4):113-6.
173. Tabll A, Abbas AT, El-Kafrawy S, Wahid A. Monoclonal antibodies: Principles and applications of immmunodiagnosis and immunotherapy for hepatitis C virus. World journal of hepatology. 2015;7(22):2369-83.
174. Ascoli CA, Aggeler B. Overlooked benefits of using polyclonal antibodies. BioTechniques. 2018;65(3):127-36.
175. Custers R, Steyaert J. Discussions on the quality of antibodies are no reason to ban animal immunization. EMBO reports. 2020;21(12):e51761.
176. Min J, Song EK, Kim H, Kim KT, Park TJ, Kang S. A Recombinant Secondary Antibody Mimic as a Target-specific Signal Amplifier and an Antibody Immobilizer in Immunoassays. Scientific reports. 2016;6:24159.
177. Tie L, Xiao H, Wu D-l, Yang Y, Wang P. A brief guide to good practices in pharmacological experiments: Western blotting. Acta Pharmacologica Sinica. 2021;42(7):1015-7.
178. Abeel T, Saeys Y, Bonnet E, Rouzé P, Van de Peer Y. Generic eukaryotic core promoter prediction using structural features of DNA. Genome research. 2008;18(2):310-23.
179. Magaki S, Hojat SA, Wei B, So A, Yong WH. An Introduction to the Performance of Immunohistochemistry. Methods in molecular biology (Clifton, NJ). 2019;1897:289-98.
180. Voskuil J. Commercial antibodies and their validation. F1000Research. 2014;3:232.
181. Groff K, Allen D, Fiebig M, Cosson P, Casey W, Clippinger AJ. An approach to identifying quality research antibodies. BioTechniques. 2022;73(4):167-70.
182. Weller MG. Quality Issues of Research Antibodies. Analytical chemistry insights. 2016;11:21-7.

183. Pillai-Kastoori L, Heaton S, Shiflett SD, Roberts AC, Solache A, Schutz-Geschwender AR. Antibody validation for Western blot: By the user, for the user. The Journal of biological chemistry. 2020;295(4):926-39.
184. Han H. RNA Interference to Knock Down Gene Expression. Methods in molecular biology (Clifton, NJ). 2018;1706:293-302.
185. Dana H, Chalbatani GM, Mahmoodzadeh H, Karimloo R, Rezaiean O, Moradzadeh A, et al. Molecular Mechanisms and Biological Functions of siRNA. International journal of biomedical science : IJBS. 2017;13(2):48-57.
186. Wu W, Hodges E, Redelius J, Höög C. A novel approach for evaluating the efficiency of siRNAs on protein levels in cultured cells. Nucleic acids research. 2004;32(2):e17.
187. Liang W, Mason AJ, Lam JK. Western blot evaluation of siRNA delivery by pH-responsive peptides. Methods in molecular biology (Clifton, NJ). 2013;986:73-87.
188. Mocellin S, Provenzano M. RNA interference: learning gene knock-down from cell physiology. Journal of translational medicine. 2004;2(1):39.
189. Vanli G, Cuesta-Marban A, Widmann C. Evaluation and validation of commercial antibodies for the detection of Shb. PloS one. 2017;12(12):e0188311.
190. Alon M, Emmanuel R, Qutob N, Bakhman A, Peshti V, Brodezki A, et al. Refinement of the endogenous epitope tagging technology allows the identification of a novel NRAS binding partner in melanoma. Pigment cell & melanoma research. 2018;31(5):641-8.
191. Shuaib M, Parsi KM, Thimma M, Adroub SA, Kawaji H, Seridi L, et al. Nuclear AGO1 Regulates Gene Expression by Affecting Chromatin Architecture in Human Cells. Cell systems. 2019;9(5):446-58.e6.
192. Zhou H, Stein CB, Shafiq TA, Shipkovenska G, Kalocsay M, Paulo JA, et al. Rixosomal RNA degradation contributes to silencing of Polycomb target genes. Nature. 2022;604(7904):167-74.
193. Virk HS, Rekas MZ, Biddle MS, Wright AKA, Sousa J, Weston CA, et al. Validation of antibodies for the specific detection of human TRPA1. Scientific reports. 2019;9(1):18500.
194. Lemos Duarte M, Trimbake NA, Gupta A, Tumanut C, Fan X, Woods C, et al. High-throughput screening and validation of antibodies against synaptic proteins to explore opioid signaling dynamics. Communications biology. 2021;4(1):238.
195. Desai SS, Kharade SS, Parekh VI, Iyer S, Agarwal SK. Pro-oncogenic Roles of HLXB9 Protein in Insulinoma Cells through Interaction with Nono Protein and Down-regulation of the c-Met Inhibitor Cblb (Casitas B-lineage Lymphoma b). The Journal of biological chemistry. 2015;290(42):25595-608.
196. Burckhardt CJ, Minna JD, Danuser G. Co-immunoprecipitation and semi-quantitative immunoblotting for the analysis of protein-protein interactions. STAR Protoc. 2021;2(3):100644.
197. Liu C, Song X, Nisbet R, Götz J. Co-immunoprecipitation with Tau Isoform-specific Antibodies Reveals Distinct Protein Interactions and Highlights a Putative Role for 2N Tau in Disease. The Journal of biological chemistry. 2016;291(15):8173-88.
198. Liau BB, Sievers C, Donohue LK, Gillespie SM, Flavahan WA, Miller TE, et al. Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. Cell stem cell. 2017;20(2):233-46.e7.
199. Dupont C, Armant DR, Brenner CA. Epigenetics: definition, mechanisms and clinical perspective. Seminars in reproductive medicine. 2009;27(5):351-7.
200. Kagohara LT, Stein-O'Brien GL, Kelley D, Flam E, Wick HC, Danilova LV, et al. Epigenetic regulation of gene expression in cancer: techniques, resources and analysis. Briefings in functional genomics. 2018;17(1):49-63.
201. Romani M, Pistillo MP, Banelli B. Epigenetic Targeting of Glioblastoma. Frontiers in oncology. 2018;8:448.

202. Kang JG, Park JS, Ko JH, Kim YS. Regulation of gene expression by altered promoter methylation using a CRISPR/Cas9-mediated epigenetic editing system. Scientific reports. 2019;9(1):11960.
203. Kang JG, Park JS, Ko J-H, Kim Y-S. Regulation of gene expression by altered promoter methylation using a CRISPR/Cas9-mediated epigenetic editing system. Scientific reports. 2019;9(1):11960.
204. Haberle V, Stark A. Eukaryotic core promoters and the functional basis of transcription initiation. Nature reviews Molecular cell biology. 2018;19(10):621-37.
205. Terranova C, Tang M, Orouji E, Maitituoheti M, Raman A, Amin S, et al. An Integrated Platform for Genome-wide Mapping of Chromatin States Using High-throughput ChIP-sequencing in Tumor Tissues. Journal of visualized experiments : JoVE. 2018(134).
206. Mikkelsen TS, Ku M, Jaffe DB, Issac B, Lieberman E, Giannoukos G, et al. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature. 2007;448(7153):553-60.
207. Sugathan A, Waxman DJ. Genome-wide analysis of chromatin states reveals distinct mechanisms of sex-dependent gene regulation in male and female mouse liver. Molecular and cellular biology. 2013;33(18):3594-610.
208. Wilbanks EG, Larsen DJ, Neches RY, Yao AI, Wu CY, Kjolby RA, et al. A workflow for genome-wide mapping of archaeal transcription factors with ChIP-seq. Nucleic acids research. 2012;40(10):e74.
209. Satterlee JS, Chadwick LH, Tyson FL, McAllister K, Beaver J, Birnbaum L, et al. The NIH Common Fund/Roadmap Epigenomics Program: Successes of a comprehensive consortium. Science advances. 2019;5(7):eaaw6507.
210. Sherman MA, Yaari AU, Priebe O, Dietlein F, Loh P-R, Berger B. Genome-wide mapping of somatic mutation rates uncovers drivers of cancer. Nature Biotechnology. 2022.
211. Steinhauser S, Kurzawa N, Eils R, Herrmann C. A comprehensive comparison of tools for differential ChIP-seq analysis. Briefings in bioinformatics. 2016;17(6):953-66.
212. Feng J, Liu T, Qin B, Zhang Y, Liu XS. Identifying ChIP-seq enrichment using MACS. Nature protocols. 2012;7(9):1728-40.
213. Awdeh A, Turcotte M, Perkins TJ. WACS: improving ChIP-seq peak calling by optimally weighting controls. BMC Bioinformatics. 2021;22(1):69.
214. Hon G, Ren B, Wang W. ChromaSig: a probabilistic approach to finding common chromatin signatures in the human genome. PLoS computational biology. 2008;4(10):e1000201.
215. Hoffman MM, Buske OJ, Wang J, Weng Z, Bilmes JA, Noble WS. Unsupervised pattern discovery in human chromatin structure through genomic segmentation. Nature methods. 2012;9(5):473-6.
216. Kundaje A, Meuleman W, Ernst J, Bilenky M, Yen A, Heravi-Moussavi A, et al. Integrative analysis of 111 reference human epigenomes. Nature. 2015;518(7539):317-30.
217. Bjørklund SS, Aure MR, Häkkinen J, Vallon-Christersson J, Kumar S, Evensen KB, et al. Subtype and cell type specific expression of lncRNAs provide insight into breast cancer. Communications biology. 2022;5(1):834.
218. Gopi LK, Kidder BL. Integrative pan cancer analysis reveals epigenomic variation in cancer type and cell specific chromatin domains. Nature communications. 2021;12(1):1419.
219. Srivastava S, Mishra RK, Dhawan J. Regulation of cellular chromatin state: insights from quiescence and differentiation. Organogenesis. 2010;6(1):37-47.
220. Yan H, Liu Y, Zhang K, Song J, Xu W, Su Z. Chromatin State-Based Analysis of Epigenetic H3K4me3 Marks of Arabidopsis in Response to Dark Stress. Frontiers in genetics. 2019;10.

221. Papait R, Cattaneo P, Kunderfranco P, Greco C, Carullo P, Guffanti A, et al. Genome-wide analysis of histone marks identifying an epigenetic signature of promoters and enhancers underlying cardiac hypertrophy. Proceedings of the National Academy of Sciences of the United States of America. 2013;110(50):20164-9.
222. Mikkelsen TS, Xu Z, Zhang X, Wang L, Gimble JM, Lander ES, et al. Comparative epigenomic analysis of murine and human adipogenesis. Cell. 2010;143(1):156-69.
223. Zhu J, Adli M, Zou JY, Verstappen G, Coyne M, Zhang X, et al. Genome-wide chromatin state transitions associated with developmental and environmental cues. Cell. 2013;152(3):642-54.
224. Rheinbay E, Suvà Mario L, Gillespie Shawn M, Wakimoto H, Patel Anoop P, Shahid M, et al. An Aberrant Transcription Factor Network Essential for Wnt Signaling and Stem Cell Maintenance in Glioblastoma. Cell reports. 2013;3(5):1567-79.
225. Palmer PB, O'Connell DG. Regression analysis for prediction: understanding the process. Cardiopulmonary physical therapy journal. 2009;20(3):23-6.
226. Schneider A, Hommel G, Blettner M. Linear regression analysis: part 14 of a series on evaluation of scientific publications. Deutsches Arzteblatt international. 2010;107(44):776-82.
227. Arashi M, Roozbeh M, Hamzah NA, Gasparini M. Ridge regression and its applications in genetic studies. PloS one. 2021;16(4):e0245376.
228. Wallisch C, Bach P, Hafermann L, Klein N, Sauerbrei W, Steyerberg EW, et al. Review of guidance papers on regression modeling in statistical series of medical journals. PloS one. 2022;17(1):e0262918.
229. Eberly LE. Multiple linear regression. Methods in molecular biology (Clifton, NJ). 2007;404:165-87.
230. Alexopoulos EC. Introduction to multivariate regression analysis. Hippokratia. 2010;14(Suppl 1):23-8.
231. Pavlou M, Ambler G, Seaman SR, Guttmann O, Elliott P, King M, et al. How to develop a more accurate risk prediction model when there are few events. BMJ (Clinical research ed). 2015;351:h3868.
232. Chintalapudi N, Angeloni U, Battineni G, di Canio M, Marotta C, Rezza G, et al. LASSO Regression Modeling on Prediction of Medical Terms among Seafarers' Health Documents Using Tidy Text Mining. Bioengineering (Basel, Switzerland). 2022;9(3).
233. Demir-Kavuk O, Kamada M, Akutsu T, Knapp EW. Prediction using step-wise L1, L2 regularization and feature selection for small data sets with large number of features. BMC Bioinformatics. 2011;12:412.
234. de Vlaming R, Groenen PJ. The Current and Future Use of Ridge Regression for Prediction in Quantitative Genetics. Biomed Res Int. 2015;2015:143712.
235. Luscombe NM, Austin SE, Berman HM, Thornton JM. An overview of the structures of protein-DNA complexes. Genome Biology. 2000;1(1):reviews001.1.
236. Enroth S, Rada-Iglesisas A, Andersson R, Wallerman O, Wanders A, Påhlman L, et al. Cancer associated epigenetic transitions identified by genome-wide histone methylation binding profiles in human colorectal cancer samples and paired normal mucosa. BMC cancer. 2011;11:450.
237. Lu Y, Cao Q, Yu Y, Sun Y, Jiang X, Li X. Pan-cancer analysis revealed H3K4me1 at bivalent promoters premarks DNA hypermethylation during tumor development and identified the regulatory role of DNA methylation in relation to histone modifications. BMC genomics. 2023;24(1):235.
238. Dunican DS, Mjoseng HK, Duthie L, Flyamer IM, Bickmore WA, Meehan RR. Bivalent promoter hypermethylation in cancer is linked to the H327me3/H3K4me3 ratio in embryonic stem cells. BMC Biology. 2020;18(1):25.

239. Bardhan K, Liu K. Epigenetics and colorectal cancer pathogenesis. Cancers. 2013;5(2):676-713.
240. Ku M, Koche RP, Rheinbay E, Mendenhall EM, Endoh M, Mikkelsen TS, et al. Genomewide analysis of PRC1 and PRC2 occupancy identifies two classes of bivalent domains. PLoS genetics. 2008;4(10):e1000242.
241. Alarcón T, Sardanyés J, Guillamon A, Menendez JA. Bivalent chromatin as a therapeutic target in cancer: An in silico predictive approach for combining epigenetic drugs. PLoS computational biology. 2021;17(6):e1008408.
242. Zhu XX, Yan YW, Ai CZ, Jiang S, Xu SS, Niu M, et al. Jarid2 is essential for the maintenance of tumor initiating cells in bladder cancer. Oncotarget. 2017;8(15):24483-90.
243. Lei X, Xu JF, Chang RM, Fang F, Zuo CH, Yang LY. JARID2 promotes invasion and metastasis of hepatocellular carcinoma by facilitating epithelial-mesenchymal transition through PTEN/AKT signaling. Oncotarget. 2016;7(26):40266-84.
244. Zhang X, Li J, Yang Q, Wang Y, Li X, Liu Y, et al. Tumor mutation burden and JARID2 gene alteration are associated with short disease-free survival in locally advanced triple-negative breast cancer. Annals of Translational Medicine. 2020;8(17):1052.
245. Cao J, Li H, Liu G, Han S, Xu P. Knockdown of JARID2 inhibits the proliferation and invasion of ovarian cancer through the PI3K/Akt signaling pathway. Molecular medicine reports. 2017;16(3):3600-5.
246. Sreeshma B, Devi A. JARID2 and EZH2, the eminent epigenetic drivers in human cancer. Gene. 2023;879:147584.
247. Pchelintsev NA, Adams PD, Nelson DM. Critical Parameters for Efficient Sonication and Improved Chromatin Immunoprecipitation of High Molecular Weight Proteins. PloS one. 2016;11(1):e0148023.
248. Skene PJ, Henikoff S. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. eLife. 2017;6:e21856.
249. Zhu Q, Liu N, Orkin SH, Yuan G-C. CUT&RUNTools: a flexible pipeline for CUT&RUN processing and footprint analysis. Genome Biology. 2019;20(1):192.
250. Kaya-Okur HS, Janssens DH, Henikoff JG, Ahmad K, Henikoff S. Efficient low-cost chromatin profiling with CUT&Tag. Nature protocols. 2020;15(10):3264-83.
251. Chen Z, Djekidel MN, Zhang Y. Distinct dynamics and functions of H2AK119ub1 and H3K27me3 in mouse preimplantation embryos. Nature Genetics. 2021;53(4):551-63.
252. Sarthy JF, Meers MP, Janssens DH, Henikoff JG, Feldman H, Paddison PJ, et al. Histone deposition pathways determine the chromatin landscapes of H3.1 and H3.3 K27M oncohistones. eLife. 2020;9:e61090.
253. Kong NR, Chai L, Tenen DG, Bassal MA. A modified CUT&RUN protocol and analysis pipeline to identify transcription factor binding sites in human cell lines. STAR Protocols. 2021;2(3):100750.
254. He C, Bonasio R. A cut above. Elife. 2017;6.
255. Boyd J, Rodriguez P, Schjerven H, Frietze S. ssvQC: an integrated CUT&RUN quality control workflow for histone modifications and transcription factors. BMC Research Notes. 2021;14(1):366.
256. Wen Z, He K, Zhan M, Li Y, Liu F, He X, et al. Distinct binding pattern of EZH2 and JARID2 on RNAs and DNAs in hepatocellular carcinoma development. Frontiers in oncology. 2022;12.
257. Bender S, Tang Y, Lindroth Anders M, Hovestadt V, Jones David TW, Kool M, et al. Reduced H3K27me3 and DNA Hypomethylation Are Major Drivers of Gene Expression in K27M Mutant Pediatric High-Grade Gliomas. Cancer Cell. 2013;24(5):660-72.
258. Skene PJ, Henikoff JG, Henikoff S. Targeted in situ genome-wide profiling with high efficiency for low cell numbers. Nature protocols. 2018;13(5):1006-19.

259. Brahma S, Henikoff S. CUT&RUN Profiling of the Budding Yeast Epigenome. Methods in molecular biology (Clifton, NJ). 2022;2477:129-47.
260. Stupp R, Mason WP, van den Bent MJ, Weller M, Fisher B, Taphoorn MJ, et al. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. The New England journal of medicine. 2005;352(10):987-96.
261. Benedetti V, Banfi F, Zaghi M, Moll-Diaz R, Massimino L, Argelich L, et al. A SOX2-engineered epigenetic silencer factor represses the glioblastoma genetic program and restrains tumor development. Science advances. 2022;8(31):eabn3986.
262. Shergalis A, Bankhead A, 3rd, Luesakul U, Muangsin N, Neamati N. Current Challenges and Opportunities in Treating Glioblastoma. Pharmacological reviews. 2018;70(3):412-45.
263. Markouli M, Strepkos D, Papavassiliou KA, Papavassiliou AG, Piperi C. Bivalent Genes Targeting of Glioma Heterogeneity and Plasticity. International journal of molecular sciences. 2021;22(2).
264. Mabe NW, Garcia NMG, Wolery SE, Newcomb R, Meingasner RC, Vilona BA, et al. G9a Promotes Breast Cancer Recurrence through Repression of a Pro-inflammatory Program. Cell reports. 2020;33(5):108341.
265. Richards AL, Eckhardt M, Krogan NJ. Mass spectrometry-based protein-protein interaction networks for the study of human diseases. Mol Syst Biol. 2021;17(1):e8792.
266. Alfaro JA, Ignatchenko A, Ignatchenko V, Sinha A, Boutros PC, Kislinger T. Detecting protein variants by mass spectrometry: a comprehensive study in cancer cell-lines. Genome Medicine. 2017;9(1):62.

List of Abbreviations

3xflag-JARID2	Jumonji and AT-Rich Interacting Domain 2 plasmid bearing 3xFlag tags
5mC	5-methyl-cytosine
AEBP2	Adipocyte enhancer-binding protein
AIC	5-amino-imidazole-4-carboxamide
AP	Alkaline phosphatase
ARID	AT-rich interaction domain
ATCC	American Type Culture Collection
BAM	Binary alignment map
BER	Base excision repair
bFGF	basic fibroblast growth factor
BH	Benjamini-Hochberg
BLAST	Basic Local Alignment Search Tool
bp	Base pair
BSA	Bovine serum albumin
BWA	Burrows-Wheeler Aligner
C5	Carbon-5 position
Ca2+	Calcium ions
CaCl2	Calcium Chloride
cAMP	Adenosine 3’,5’-cyclic monophosphate
ChIP-seq	Chromatin immunoprecipitation followed by sequencin
Co-IP	Co-immunoprecipitation
CO2	Carbon dioxide
CpG island	5'—C—phosphate—G—3
CSC	Cancer stem cell
CST	Cell Signalling Technology
CUT&RUN	Cleavage under targets & release using nuclease
DAVID	Database for Annotation, visualization and Integrated Discovery
ddNTPs	Dideoxythymidine triphosphate
dH2O	Distilled water
DLX2	Distal-Less Homeobox 2
DMEM	Dulbecco's Modified Eagle Medium
DMG	Diffuse midline glioma
DNA	Deoxyribonucleic acid
DNMT	DNA methyltransferase
dNTP	Deoxythymidine triphosphate
DPF3	Member of the D4 protein family
dsDNA	Double stranded DNA
E.coli	Escherichia coli
EDTA	Ethylenediaminetetraacetic acid
EED1-4	Embryonic ectoderm development 1 to 4
EGF	Epidermal growth factor
EGFR	Epidermal growth factor tyrosine kinase receptor
ELAND	Efficient Large-Scale Alignment of Nucleotide Database
ELISA	The enzyme-linked immunosorbent assay
EMT	Epithelial and mesenchymal transition

ENCODE	ENCyclopedia Of DNA Elements
ESC	Embryonic stem cell
EZH1	Enhancer of Zeste 2 Polycomb Repressive Complex 1 Subunit
EZH2	Enhancer of Zeste 2 Polycomb Repressive Complex 2 Subunit
FABP7	Fatty Acid Binding Protein 7
FBS	Fetal bovine serum
Fc	Crystallizable fragment
FDR	Fold discovery rate
FITC	Fluorescein isothiocyanate
FPKM	Fragments Per Kilobase of transcript per Million mapped reads
FRiP	Fragment of reads in peak
GAPDH	Glyceraldehyde 3-phosphate dehydrogenase
GBM	Glioblastoma
GC	Guanine-Cytocine
GLASS	Glioma Longitudinal AnalysSiS
GRB2	Growth factor receptor bound protein 2
GSC	Glioma stem cell
GSC8	Glioblastoma stem-like cell
GSEA	Gene set enrichment analysis
GTF	Gene transfer format
H3K27me3	Trimethylation of lysine 27 of histone 3
H3K36me3	Trimethylation of lysine 36 of histone 3
H3K4me3	Trimethylation of lysine 4 of histone 3
H3K9me3	Trimethylation of lysine 9 of histone 3
HCC	Hepatocellular carcinoma
HEK293T	A human embryonic kidney cell line
HEY1	Hes related family bHLH transcription factor with YRPW motif 1
HiBiv	High H3K27me3:H3K4me3 ratio
HMM	Hidden Markov Model
HMT	Histone methyltransferase
HRP	Horseradish peroxidase
HTS	High throughput sequencing
IDE	Integrated development environment
IDH	Isocitrate dehydrogenases
IDH1	Isocitrate dehydrogenase 1
IDR	Irreproducibility discovery rate
IF	immunofluorescence
IgG	Immunoglobulin, G
IGV	Integrative Genomics Viewer
IP,	Immunoprecipitation
JARID2	Jumonji and AT-Rich Interacting Domain 2
JBSgenes	JARID2 binding site genes
JmjC	Jumonji C
JmjN	Jumonji N
KG	Ketoglutarate
KMT	Lysine methyltransferases
LASSO	Least absolute shrinkage and selection operator

LB	Luria Bertani
LE	Leading edge
lncRNA	Long non-coding Ribonucleic acid
LoBiv	Low H3K27me3:H3K4me3 ratio
log2FC	Logarithm of fold change
LOH	Loss of heterozygosity
M-fold	High confidence fold enrichment
mAbs	Monoclonal antibodies
MACS	Model based analysis of ChIP-seq
MAPQ	Mapping quality
MAPK	Mitogen-activated protein kinase
MDM2	Mouse double minute 2
MGMT	O-6-methylguanine-DNA methyltransferase
MMR	Mismatch repair pathway
mRNA	Messanger Ribonucleic acid
MS	Mass spectrometry
MTIC	Metabolite 5-(3-methyltriazen-1-yl)-imidazole-4-carboxamide
mTOR	Mammalian target of rapamycin
N3-MA	N3-MA
N7-Meg	N7-methylguanine
N1ICD	Notch 1 intercellular domain-associated genes
NB	Neurobasal media
NCBI	National centre for biotechnology information
NEAA,	Non-essential amin acid
NGS	Next generation sequencing
NOTCH	Neurogenic locus notch homolog protein
NRF	Nonredundant Fraction
NRF	Non-redundant Fraction
NSC	Normalized strand coefficient
NSC	Normalized strand cross-correlation coefficient
NTRK2	Neurotrophic tyrosine receptor kinase
O6-Meg	O6-methylguanine
pAbs	Polyclonal antibodies
pAG-MNase	Protein A/G micrococcal nuclease
PAGE	Polyacrylamide Gel Electrophoresis
PARP	Poly (ADP)-ribose polymerase
PBC1	PCR bottlenecking coefficients 1
PBC2	PCR bottlenecking coefficients 2
PBS	Dulbecco’s phosphate buffered saline
PBST	Phosphate Buffered Saline Tween-20
PC1	Principle component 1
PCR	Polymerase chain reaction
PRC2	Polycomb Repressive Complex 2
PRMT	Arginine methyltransferases
PI3K	Phosphatidylinositol-3-kinase
PTEN	Phosphatase and tensin homolog
PTM	Post-translational modification

PVDF	Polyvinylidene difluoride
QC	Quality control
qPCR	Quantitative polymerase chain reactio
RBBP4/7	Retinoblastoma binding protein 4 and 7
RB	Retinoblastoma
RIPA	Radio-immunoprecipitation assay
RNA	Ribonucleic acid
RNA,	Ribonucleic acid
RNAseq	Ribonucleic acid sequencing
RPL30	Ribosomal protein L30
RR	Ridge regression
RSC	Relative strand correlation
RSC	Relative strand cross-correlation coefficient
RT	Room temperature
SALL2	Spalt Like Transcription Factor 2
SAM	Sequence alignment map
SAT2	Spermidine/Spermine N1-Acetyltransferase Family Member
SDS	Sodium Dodecyl Sulfate
SEACR	Sparse Enrichment Analysis for CUT&RUN
SGK4	Serine/threonine protein kinase 4
shRNA	Short hairpin Ribonucleic acid
siRNA	Small interfering ribonucleic acid
SOS1	Son of sevenless 1
SOX2	Sex Determining Region Y-box 2
ssDNA	Single stranded DNA
SUZ12	Suppressor of zeste 12
TAE	Tris-acetate-Ethylenediaminetetraacetic acid
TC	Tissue culture
TE	Tris Ethylenediaminetetraacetic acid
TF	Transcription factor
TMZ	Temozolomide
TP53	Tumour protein 53
TSS	Transcription start site
UV	Ultraviolet
VEGF	Vascular endothelial growth factor
w	Window of fixed size
WB	Western blot
WCE	Whole cell extract
WHO	World Health Organization
β-actin	Beta-actin
ΔN-JARID2	Cleaved product of full-length JARID2
λ	Lambda

Appendices

Appendix A

A.1 List of tools and software used to develop the ChIP-seq analysis pipeline

Tool/software	Version	Availability
FASTQC	0.11.9	https://www.bioinformatics.babraham.ac.uk/projects/fastqc/
GATK	4.2.5	https://github.com/broadinstitute/gatk/releases
BWA	0.7.17	https://github.com/lh3/bwa
Samtools	1.11	https://github.com/samtools/samtools
Bedtools	2.30.0	https://github.com/arq5x/bedtools2
Picard	2.21.2	https://broadinstitute.github.io/picard/
MACS3	2.2.7.1	https://github.com/macs3-project/MACS
ChromHMM	1.23	https://github.com/jernst98/ChromHMM
R	4.0.3	https://www.r-project.org
Netbeans	8.2	https://netbeans.apache.org
Python	3.8.5	
qsubsec	3.0a28	https://github.com/alastair-droop/qsubsec
Cutadapt	3.6	https://github.com/marcelm/cutadapt https://cutadapt.readthedocs.io/en/stable/
Deeptools	3.5.1	https://github.com/deeptools/deepTools
Miniconda3	4.11	https://docs.conda.io/en/latest/miniconda.html

Appendix B

B.1 List of Fastq files (single-end reads) for two cell lines (GSC8 and GSC8per), which each underwent ChIP-seq to detect the location of both H3K27me3 and H3K4me3 marks, compared to input DNA controls.

Sample name	SRR number	Histone mark
Input_GSC8	SRR4420628	Null
Input_GSC8per	SRR4420631	Null
H3K4me3_GSC8	SRR4420639	Narrow peak
H3K4me3_GSC8per	SRR4420644	Narrow peak
H3K27me3_GSC8	SRR4420649	Broad peak
H3K27me3_GSC8per	SRR4420654	Broad peak

Appendix C

C.1 A script for generating the genomic read count

```
# A function to write a given message to stderr:
log.message <- function(..., verbose=NA){
  if(is.na(verbose)){
    verb <- get0('.verbose', ifnotfound=TRUE)
  } else {
    verb <- verbose
  }
  if(identical(verb, TRUE)){
    message(sprintf(...))
    flush(stderr())
  }
}

# A function to exit with a given error:
error <- function(..., exit.code=1, verbose=TRUE){
  m <- sprintf('ERROR: %s', sprintf(...))
  if(identical(interactive(), TRUE)){
    stop(m)
  } else {
    log.message(m, verbose=verbose)
    quit(save='no', status=exit.code)
  }
}

# A function to quietly load a vector of libraries from character strings:
loadLibrary <- function(x, verbose=NA){
  log.message('loading library "%s"', x, verbose=verbose)
  if(identical(interactive(), TRUE)){
    res <- require(x, character.only=TRUE, quietly=TRUE)
  } else {
    res <- suppressWarnings(suppressPackageStartupMessages(require(x, character.only=TRUE,
quietly=TRUE)))
  }
  if(!identical(res, TRUE)) error('failed to load package "%s"', x)
}

loadGenome <- function(filename){
  genome <- read.table(filename, sep='\t', colClasses=c('character', 'numeric', 'logical', 'character'),
col.names=c('chr', 'length', 'isCircular', 'genome'))
  return(Seqinfo(genome$chr, seqlengths=genome$length, isCircular=genome$isCircular,
genome=genome$genome))
}
```

```
# Attempt to load argparse and construct the input arguments:
if(!identical(interactive(), TRUE)){
  loadLibrary('argparse', verbose=FALSE)
  parser <- ArgumentParser(description='Generate window read counts from a given BAM file')
  parser$add_argument('-v', '--verbose', dest='verbose', default=FALSE, action='store_true', help='provide
verbose output')
  parser$add_argument('-e', '--display-empty', dest='empty', action='store_true', help='return all windows,
even empty ones')
  parser$add_argument('-b', '--blacklist', dest='blacklist', default=NULL, metavar='BED', help='genome
blacklist')
  parser$add_argument(dest='genome', metavar='FILE', help='Genome definition file')
  parser$add_argument(dest='windows', metavar='BED|n', help='Input window BED file or window size')
  parser$add_argument(dest='bam', metavar='BAM', help='Input BAM file to process')
  args <- parser$parse_args()
} else {
  args <- list(
    'verbose' = TRUE,
    'empty' = FALSE,
    'blacklist' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/metadata/blacklist/ENCFF356LFX.bed',
    'genome' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/metadata/genome/GRCh38-genome.txt',
    'windows' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/metadata/promoters/promoters-1k.bed',
    # 'windows' = '2000',
    'bam' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/input/bam/control_P.bam'
  )
}

# Set the global verbosity:
.verbose <- args$verbose

# Load the necessary libraries:
for(l in c('GenomicAlignments', 'Rsamtools')){
  loadLibrary(l)
}

# Read in the genome:
log.message('reading genome file "%s"...', args$genome)
genome <- loadGenome(args$genome)

# Process the input windows, either from a BED file or from a specified window size:
suppressWarnings(res <- as.numeric(args$windows))
if(is.na(res)){
  log.message('generating windows from "%s"', args$windows)
```

```
   tryCatch({
      window.coltypes <- sapply(read.table(args$windows, sep='\t', header=FALSE, nrows=1), class)
      windows <- read.table(args$windows, sep='\t', header=FALSE, colClasses=window.coltypes,
comment.char='#')
      if(ncol(windows) < 3) stop()
      windows <- windows[, 1:3]
      colnames(windows) <- c('chr', 'start', 'end')
      windows <- makeGRangesFromDataFrame(windows)
      log.message('%s windows loaded from file', format(length(windows), big.mark=','))
   }, error= function(e){error('failed to load input regions from "%s"', args$windows)})
} else {
   suppressWarnings(args$windows <- as.integer(args$windows))
   if(is.na(args$windows) || (args$windows < 1)) error('invalid window size')
   log.message('generating length %d genomic windows ', args$windows)
   windows <- tileGenome(genome, tilewidth=args$windows, cut.last.tile.in.chrom=TRUE)
   log.message('%s windows across the genome', format(length(windows), big.mark=','))
}

# Read in the reads:
tryCatch({
   log.message('loading reads from from "%s"...', args$bam)
   reads <- GRanges(readGAlignments(args$bam))
   log.message('%s reads loaded from file', format(length(reads), big.mark=','))
}, error= function(e){error('failed to load reads from "%s"', args$bam)})

# Narrow the reads to their central nucleotide:
reads <- resize(reads, width=1, fix='center')

# Read in the blacklist, if provided:
if(!is.null(args$blacklist)){
   tryCatch({
      log.message('loading blacklist from "%s"...', args$blacklist)
      blacklist <- read.table(args$blacklist, sep='\t', header=FALSE, colClasses=c('character', rep('numeric', 2)),
col.names=c('chr', 'start', 'end'))
      blacklist <- makeGRangesFromDataFrame(blacklist, seqinfo=genome)
      blacklist <- reduce(blacklist)
   }, error= function(e){error('failed to load blacklist from "%s"', args$blacklist)})
   # Mark blacklisted regions:
   log.message('marking blacklisted regions...')
   windows$blacklist <- windows %over% blacklist
}

# Count the overlapping reads:
log.message('mapping reads to windows...')
windows$counts <- countOverlaps(windows, reads)
```

```
# Convert the window counts to a data.frame:
log.message('writing output...')
res <- data.frame(
   'chr' = seqnames(windows),
   'start' = start(windows),
   'end' = end(windows)
)
if('blacklist' %in% colnames(mcols(windows))){res$blacklist <- c('N', 'Y')[as.numeric(windows$blacklist) + 1]}
res$count <- windows$counts

# Trim empty counts, if requested:
if(!identical(args$empty, TRUE)){
   res <- res[res$count > 0, ]
}

# Write the output to stdout:
write.table(res, file=stdout(), sep='\t', quote=FALSE, row.names=FALSE, col.names=TRUE)
```

C.2 A script for calculating the promoter signal

```
#!/usr/bin/env Rscript
# This script runs the promoter tests
# A function to write a given message to stderr:
log.message <- function(..., verbose=NA){
   if(is.na(verbose)){
      verb <- get0('.verbose', ifnotfound=TRUE)
   } else {
      verb <- verbose
   }
   if(identical(verb, TRUE)){
      message(sprintf(...))
      flush(stderr())
   }
}

# A function to exit with a given error:
error <- function(..., exit.code=1, verbose=TRUE){
   m <- sprintf('ERROR: %s', sprintf(...))
   if(identical(interactive(), TRUE)){
      stop(m)
   } else {
      log.message(m, verbose=verbose)
      quit(save='no', status=exit.code)
   }
}

# A function to quietly load a vector of libraries from character strings:
loadLibrary <- function(x, verbose=NA){
   log.message('loading library "%s"', x, verbose=verbose)
   if(identical(interactive(), TRUE)){
      res <- require(x, character.only=TRUE, quietly=TRUE)
   } else {
      res <- suppressWarnings(suppressPackageStartupMessages(require(x, character.only=TRUE,
quietly=TRUE)))
   }
   if(!identical(res, TRUE)) error('failed to load package "%s"', x)
}

# A function to load a genome file:
loadGenome <- function(filename){
   genome <- read.table(filename, sep='\t', colClasses=c('character', 'numeric', 'logical', 'character'),
col.names=c('chr', 'length', 'isCircular', 'genome'))
   return(Seqinfo(genome$chr, seqlengths=genome$length, isCircular=genome$isCircular,
genome=genome$genome))}
```

```
# A function to read a single window file:
readWindowFile <- function(filename, seqinfo){
   res <- read.table(filename, sep='\t', header=TRUE, colClasses=c('character', rep('integer', 2), 'character',
'integer'))
   res$blacklist <- c('N'=FALSE, 'Y'='TRUE')[res$blacklist]
   res <- res[res$chr %in% seqnames(seqinfo), ]
   return(makeGRangesFromDataFrame(res, keep.extra.columns=TRUE, seqinfo=seqinfo))
}

# A function to load a pair (control, experiment) of window files:
readWindowFilePair <- function(control.filename, experiment.filename, seqinfo, filter.blacklist=FALSE){
   control <- readWindowFile(control.filename, seqinfo=seqinfo)
   experiment <- readWindowFile(experiment.filename, seqinfo=seqinfo)
   stopifnot('control and experiment genome window mismatch'=all(control == experiment))
   res <- GRanges(
      seqnames=seqnames(control),
      ranges=ranges(control),
      strand=strand(control),
      seqinfo=seqinfo(control)
   )
   res$blacklist <- control$blacklist
   res$control <- control$count
   res$experiment <- experiment$count
   if(identical(filter.blacklist, TRUE)){
      res <- res[res$blacklist == FALSE]
      mcols(res)$blacklist <- NULL
   }
   return(res)
}

# A function to select a set of counts from a window object:
selectCounts <- function(x, source=c('experiment', 'control'), lower.quantile=NA, upper.quantile=NA,
rm.zero=TRUE){
   source <- match.arg(source)
   x <- mcols(x)[[source]]
   if(identical(rm.zero, TRUE)){
      x <- x[x != 0]
   }
   if(!is.na(lower.quantile)){
      lower.threshold <- quantile(x, lower.quantile)
      x <- x[x >= lower.threshold]
   }
   if(!is.na(upper.quantile)){
      upper.threshold <- quantile(x, upper.quantile)
      x <- x[x < upper.threshold]
   }
```

```
  return(x)}

# A function to calculate the probability of expression for each promoter region:
calculateSignalProbabilities <- function(x, l.exp, l.con, adjust.method='fdr'){
  # Calculate the moderated lambdas:
  epsilon <- x$control / l.con
  epsilon.mod <- epsilon
  epsilon.mod[epsilon.mod < 1] <- 1
  mcols(x)$lambda.mod <- l.exp * epsilon.mod
  # Calculate p-values:
  mcols(x)$p.value <- ppois(x$experiment, lambda=x$lambda.mod, lower.tail=FALSE)
  mcols(x)$p.adj<- p.adjust(mcols(x)$p.value, method=adjust.method)
  # Return the data:
  return(x)
}

# Attempt to load argparse and construct the input arguments:
if(!identical(interactive(), TRUE)){
  loadLibrary('argparse', verbose=FALSE)
  parser <- ArgumentParser(description='Test promoter p-values')
  parser$add_argument('-v', '--verbose', dest='verbose', default=FALSE, action='store_true', help='provide
verbose output')
  parser$add_argument('--lambda-gcontrol', dest='lambda.gcontrol', metavar='<l>', type='double',  default=1,
help='control genome lambda (default 1)')
  parser$add_argument('--lambda-gexp', dest='lambda.gexp', metavar='<l>', type='double',  default=1,
help='experiment genome lambda (default 1)')
  parser$add_argument('--alpha', dest='threshold.alpha', metavar='<a>', type='double',  default=0.05,
help='FDR significance threshold (default 0.05)')
  parser$add_argument(dest='genome', metavar='genome', help='Genome definition file')
  parser$add_argument(dest='control.promoters', metavar='control', help='control promoter window file')
  parser$add_argument(dest='experiment.promoters', metavar='experiment', help='experiment promoter
window file')
  args <- parser$parse_args()
} else {
  args <- list(
    'verbose' = TRUE,
    'lambda.gcontrol' = 19,
    'lambda.gexp' = 17,
    'threshold.alpha' = 0.05,
    'genome' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/metadata/genome/GRCh38-genome.txt',
    'control.promoters' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/pipelines/original/windows/promoters/control_P-promoters.windows',
    'experiment.promoters' = '/Users/alastair/Documents/Work/ad-bioinformatics/projects/LS2020-
GBMProm/pipelines/original/windows/promoters/EZH2_P-promoters.windows'
  )}
```

```
# Set the global verbosity:
.verbose <- args$verbose

# Load the necessary libraries:
for(l in c('GenomicRanges')){
   loadLibrary(l)
}
log.message('control genomic lambda is %d', args$lambda.gcontrol)
log.message('experiment genomic lambda is %d', args$lambda.gexp)
log.message('calculating significance at %0.6f%%', args$threshold.alpha * 100)

# Log the file locations:
log.message('input files:')
log.message('  genome file      : %s', args$genome)
log.message('  control windows   : %s', args$control.promoters)
log.message('  experiment windows: %s', args$experiment.promoters)

# Load the genome seqinfo object:
genome <- loadGenome(args$genome)

# Load the promoter window counts:
log.message('reading promoter window counts...')
promoter.windows <- readWindowFilePair(control=args$control.promoter,
experiment=args$experiment.promoter, seqinfo=genome, filter.blacklist=TRUE)
log.message('%d promoter locations across genome', length(promoter.windows))

# Calculate the normalised counts ratio:
promoter.windows$logFC <- log2(promoter.windows$experiment / promoter.windows$control)

# Calculate the per-promoter modification factors:
mcols(promoter.windows)$modification.factor <- promoter.windows$control / args$lambda.gcontrol
mcols(promoter.windows)[promoter.windows$modification.factor < 1, 'modification.factor'] <- 1
mcols(promoter.windows)$test.lambda <- args$lambda.gexp * promoter.windows$modification.factor

# Calculate the lambdas:
calculate_promoter_pois_pvalue <- function(test.stat, test.lambda){
   res <- poisson.test(test.stat, test.lambda, alternative='greater')
   return(res$p.value)}

# Run the poisson tests:
mcols(promoter.windows)$pvalue <- apply(mcols(promoter.windows), 1, function(i){
   calculate_promoter_pois_pvalue(i['experiment'], i['test.lambda'])
})
```

```
# Calculate the adjusted p-values:
mcols(promoter.windows)$padj <- p.adjust(mcols(promoter.windows)$pvalue, method='fdr')

# Write significance:
promoter.windows$significant = as.factor(c('ns', 'sig')[as.numeric(promoter.windows$padj <=
args$threshold.alpha) + 1])
n.sig <- table(promoter.windows$significant)['sig']
log.message('%d/%d (%0.2f%%) promoters significant at %0.6f%% threshold', n.sig,
length(promoter.windows), (n.sig/length(promoter.windows)) * 100, args$threshold.alpha * 100)

# Write the output:
output <- as.data.frame(promoter.windows)
colnames(output) <- c('chr', 'start', 'end', 'width', 'strand', 'control_counts', 'experiment_counts', 'logFC',
'mod_fac', 'lambda_test', 'p_value', 'p_adj', 'significant')
output$strand <- NULL
write.table(output, file=stdout(), sep='\t', quote=FALSE, row.names=FALSE)
```

C.3 A script for scoring the enrichment of each promoter region based on the corrected p-values.

This script takes the output of window read count of the promoter regions and calculate all possible promoter states based on the corrected p-value with a threshold of P < 0.00001.

```
# Load the necessary libraries:
library(GenomicRanges)
library(jsonlite)
library(ggplot2)

# Set the analysis arguments:
args <- list(
   'comparisons' = file.path('.', 'analysis.json'),
   'input.dir' = file.path('..', '..', 'results'),
   'expression' = file.path('..', '..', '..', '..', 'input', 'expression', 'promoter-expression.txt'),
   'genome' = file.path('..', '..', '..', '..', 'metadata', 'genome', 'GRCh38-genome.txt'),
   'result.dir' = file.path('..', 'results'),
   'threshold' = 0.00001)

# Load the genome:
message(sprintf('loading genome data from "%s"...', args$genome))
genome <- read.table(args$genome, sep='\t', colClasses=c('character', 'numeric', 'logical', 'character'),
col.names=c('chr', 'length', 'isCircular', 'genome'))
genome <- Seqinfo(genome$chr, seqlengths=genome$length, isCircular=genome$isCircular,
genome=genome$genome)

# Load the expression data:
message(sprintf('loading expression data from "%s"...', args$expression))
expression <- read.table(args$expression, sep='\t', header=TRUE, col.names=c('chr', 'pos', 'logFC', 'P', 'R'))
expression$start <- expression$pos
expression$end = expression$start
expression$pos <- NULL
expression <- expression[expression$chr %in% seqlevels(genome), ]
expression <- makeGRangesFromDataFrame(expression, keep.extra.columns=TRUE, seqinfo=genome)
expression <- sort(expression)

# Load the sample file:
message(sprintf('loading sample data from "%s"...', args$comparisons))
samples <- jsonlite::fromJSON(args$comparisons)
sample.labels <- names(samples)
names(sample.labels) <- sapply(samples, '[[', 'label')
samples <- as.data.frame(sapply(sort(setdiff(unique(unlist(lapply(samples, names))), 'label')), function(ctype){
   sapply(samples, '[[', ctype)}))
rownames(samples) <- names(sample.labels)
```

```
# Define the valid datasets:
message('building input file list...')
valid.datasets <- list()
for(sample.id in rownames(samples)){
   sample.label <- sample.labels[sample.id]
   for(ctype in colnames(samples)[which(samples[sample.id,]==TRUE)]){
      dataset.label <- sprintf('%s%s', sample.id, ctype)
      dataset.filename <- file.path(args$input.dir, sprintf('%s_%s', sample.label, ctype), sprintf('%s_%s-
signal.txt', sample.label, ctype))
      if(!file.exists(dataset.filename)){
         stop(sprintf('%s signal file "%s" missing', dataset.label, dataset.filename))
      }
      valid.datasets[[dataset.label]] <- dataset.filename
   }
}

# Load the raw data:
message('loading target data...')
d.raw <- sapply(valid.datasets, function(filename){
   output <- read.table(filename, sep='\t', header=TRUE, colClasses=c('character', rep('numeric', 10), 'factor'))
   return(makeGRangesFromDataFrame(output, keep.extra.columns=TRUE, seqinfo=genome))
}, simplify=FALSE)

# Build the data:
d <- GRanges(
   seqnames = seqnames(d.raw[[1]]),
   ranges = ranges(d.raw[[1]]),
   seqinfo=genome
)

# Merge with expression data:
message('calculating expression data overlaps...')
o <- findOverlaps(query=resize(d, width=1, fix='center'), subject=expression)
d$expression_logFC <- NA
mcols(d)[queryHits(o), 'expression_logFC'] <- expression[subjectHits(o)]$logFC
d$expression_P <- NA
mcols(d)[queryHits(o), 'expression_P'] <- expression[subjectHits(o)]$P
d$expression_R <- NA
mcols(d)[queryHits(o), 'expression_R'] <- expression[subjectHits(o)]$R

# Merge the binding data:
message('merging sample data...')
for(dataset in names(d.raw)){
   # Add both value and p-value to output here!
   mcols(d)[[sprintf('%s_lambda_mod', dataset)]] <- d.raw[[dataset]]$lambda_mod
   mcols(d)[[sprintf('%s_pvalue', dataset)]] <- d.raw[[dataset]]$p_value
```

```
  mcols(d)[[sprintf('%s_padj', dataset)]] <- d.raw[[dataset]]$p_adj
  mcols(d)[[dataset]] <- as.numeric(d.raw[[dataset]]$p_adj <= args$threshold)}

# Calculate all possible promoter statuses:
status.label <- paste(rownames(samples), collapse=':')

# Get the ctype status factors:
for(ctype in colnames(samples)){
  status.levels <- expand.grid(lapply(1:nrow(samples), function(i){c('0', '1')}))
  for(i in which(!samples[[ctype]])){
    status.levels[, i] <- rep('0', nrow(status.levels))
  }
  status.levels <- unique(apply(status.levels, 1, paste, collapse=''))
  res <- rep('', length(d))
  for(sample.id in rownames(samples)){
    col <- sprintf('%s%s', sample.id, ctype)
    if(col %in% names(valid.datasets)){
      v <- mcols(d)[, col]
    } else {
      v <- rep('0', length(d))
    }
    res <- sprintf('%s%s', res, as.character(v))
  }
  mcols(d)[[ctype]] <- factor(res, levels=status.levels)
}

# Write the data to the resuts file:
res.filename <- file.path(args$result.dir, 'original-results.txt')
message(sprintf('writing combined data to "%s"...', res.filename))
write.table(d, file=res.filename, sep='\t', quote=FALSE, row.names=FALSE)

# A function to plot p-value histograms:
plot.pHist <- function(x, col, filename){
  g <- ggplot(x, aes_string(col, fill='label'))
  g <- g + geom_histogram(bins=100, show.legend=FALSE)
  g <- g + scale_y_continuous(trans=scales::pseudo_log_trans())
  g <- g + theme_classic()
  g <- g + facet_grid(rows=vars(label))
  g <- g + labs(x='p-value', y='Count')
  ggsave(g, filename=filename, width=12, height=16, units='in')}

# Extract the p-values for each dataset:
d.pvalues <- data.frame(
  'label' = factor(rep(names(d.raw), each=length(d.raw[[1]]))),
  'pvalue' = unlist(sapply(d.raw, function(i){mcols(i)$p_value}, simplify=FALSE)),
  'padj' = unlist(sapply(d.raw, function(i){mcols(i)$p_adj}, simplify=FALSE)))
```

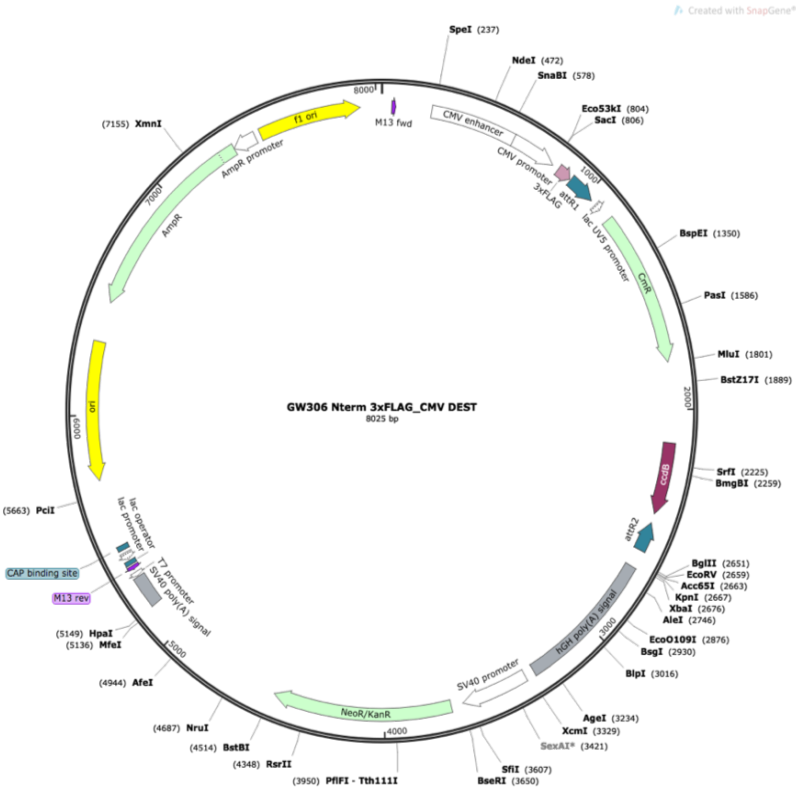
```
# Plot the p-value histograms:
plot.pHist(d.pvalues, 'pvalue', file.path(args$result.dir, 'pvalue-density.pdf'))
plot.pHist(d.pvalues, 'padj', file.path(args$result.dir, 'padj-density.pdf'))

# Extract the expression by status:
d.expression <- rbind.data.frame(
   data.frame(expression=d$expression_P, status=d$P, ctype=factor('P', levels=c('P', 'R'))),
   data.frame(expression=d$expression_R, status=d$R, ctype=factor('R', levels=c('P', 'R')))
)
d.expression <- d.expression[complete.cases(d.expression),]
d.expression <- d.expression[d.expression$expression > 0,]

# A function to plot the expression data boxplots:
plot.expression <- function(x, filename, notch=TRUE){
   g <- ggplot(x, aes(x=status, y=expression, fill=ctype))
   g <- g + geom_boxplot(outlier.shape = NA, notch=notch)
   g <- g + theme_classic()
   g <- g + coord_cartesian(ylim=c(0, quantile(x$expression, 0.9)))
   g <- g + theme(axis.ticks=element_blank(), panel.grid=element_blank(), panel.background=element_blank(),
panel.border=element_blank())
   g <- g + theme(axis.title.x=element_blank(), axis.text.x=element_text(angle=0, vjust=0.5, size=10,
family='mono'), plot.subtitle=element_text(colour='grey25'), legend.title=element_blank())
   g <- g + labs(y='Expression')
   ggsave(g, filename=filename, width=16, height=8, units='in')}

# Plot the expression data:
plot.expression(d.expression[d.expression$ctype=='R',], file.path(args$result.dir, 'expression-R.pdf'),
notch=TRUE)
plot.expression(d.expression, file.path(args$result.dir, 'expression-PR.pdf'), notch=TRUE)

# A function to plot logFC by status change:
plot.dStatus.logFC <- function(x, filename){
   g <- ggplot(x, aes(x=status, y=logFC))
   g <- g + geom_boxplot(outlier.shape = NA, fill='grey95')
   g <- g + theme_classic()
   # g <- g + coord_cartesian(ylim=c(0, quantile(x$logFC, 0.9)))
   g <- g + theme(axis.ticks=element_blank(), panel.grid=element_blank(), panel.background=element_blank(),
panel.border=element_blank())
   g <- g + theme(axis.title.x=element_blank(), axis.text.x=element_text(angle=90, vjust=0.5, size=4,
family='mono'), plot.subtitle=element_text(colour='grey25'), legend.title=element_blank())
   g <- g + labs(x=sprintf('Change in Promoter Status P -> R', status.label), y='LogFC')
   ggsave(g, filename=filename, width=16, height=8, units='in')}

# # Plot the expression by promoter status change:
d.status <- data.frame(
   'status' = factor(sprintf('%s->%s', d$P, d$R)), 'logFC' = d$expression_logFC)
```

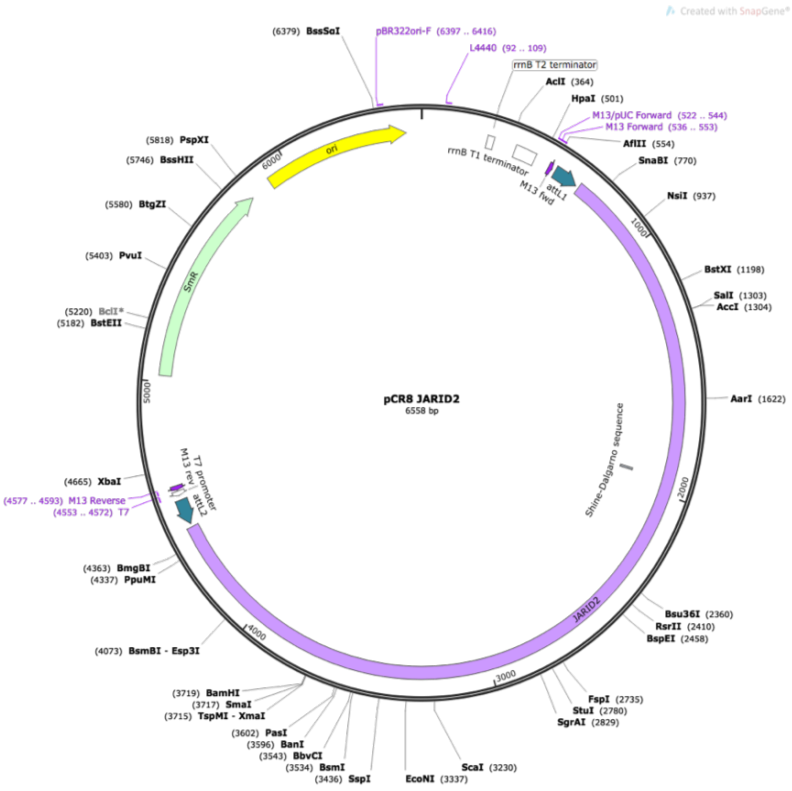
```
d.status <- d.status[complete.cases(d.status), ]
plot.dStatus.logFC(d.status, file.path(args$result.dir, 'logFC-P_R.pdf'))
```

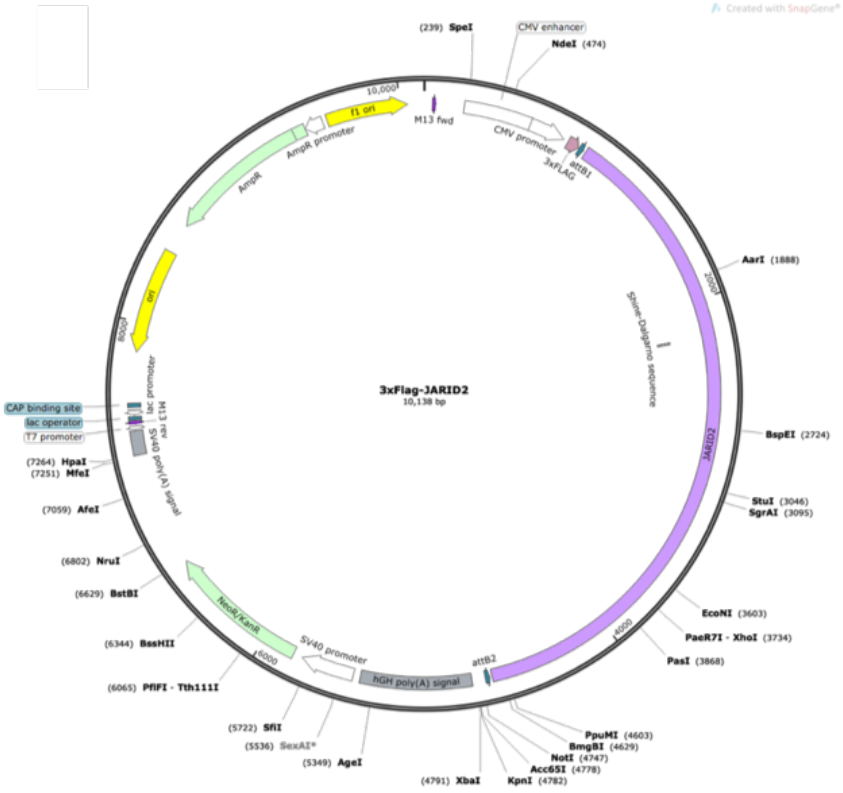
Appendix D

D.1 Plasmids used for LR cloning gateway

The following maps represents the entry plasmid (pCR8 JARID2), the destination plasmid containing 3x Flag tags (GW306 Nterm pDEST 3xFlag) and the resulted full length JARID2 expression plasmid (3xflag-JARID2) obtained from https://www.addgene.org.




Created with SnapGene®
pCR8 JARID2
6558 bp
(6379) BssSαI
pBR322ori-F (6397 .. 6416)
L4440 (92 .. 109)
rrnB T2 terminator
AclI (364)
HpaI (501)
M13/pUC Forward (522 .. 544)
M13 Forward (536 .. 553)
AflII (554)
SnaBI (770)
NsiI (937)
BstXI (1198)
SalI (1303)
AccI (1304)
AarI (1622)
Shine-Dalgarno sequence
JARID2
attL2
attL1
M13 fwd
rrnB T1 terminator
ori
SmR
(5818) PspXI
(5746) BssHII
(5580) BtgZI
(5403) PvuI
(5220) BclI*
(5182) BstEII
(4665) XbaI
(4577 .. 4593) M13 Reverse
(4553 .. 4572) T7
T7 promoter
M13 rev
(4363) BmgBI
(4337) PpuMI
(4073) BsmBI - Esp3I
(3719) BamHI
(3717) SmaI
(3715) TspMI - XmaI
(3602) PasI
(3596) BanI
(3543) BbvCI
(3534) BsmI
(3436) SspI
EcoNI (3337)
ScaI (3230)
FspI (2735)
StuI (2780)
SgrAI (2829)
Bsu36I (2360)
RsrII (2410)
BspEI (2458)


Created with SnapGene®
3xFlag-JARID2
10,138 bp
(239) SpeI
CMV enhancer
NdeI (474)
f1 ori
M13 fwd
CMV promoter
3xFLAG
attB1
AmpR promoter
AmpR
ori
lac promoter
M13 rev
SV40 poly(A) signal
CAP binding site
lac operator
T7 promoter
(7264) HpaI
(7251) MfeI
(7059) AfeI
(6802) NruI
(6629) BstBI
(6344) BssHII
(6065) PflFI - Tth111I
(5722) SfiI
(5536) SexAI*
(5349) AgeI
(4791) XbaI
KpnI (4782)
Acc65I (4778)
NotI (4747)
BmgBI (4629)
PpuMI (4603)
EcoNI (3603)
PaeR7I - XhoI (3734)
PasI (3868)
attB2
hGH poly(A) signal
SV40 promoter
NeoR/KanR
Shine-Dalgarno sequence
JARID2
AarI (1888)
BspEI (2724)
StuI (3046)
SgrAI (3095)

Appendix E

E.1 Emission probability of each state from ChromHMM tool for an external dataset

ChromHMM state (Emission order)	H3K27me3	H3K4me3	State annotation
1	0.00206885	2.20E-04	Null
2	0.179155112	1.11E-04	Repressive
3	0.207145188	0.881164871	Bivalent
4	4.35E-04	0.93974177	Active

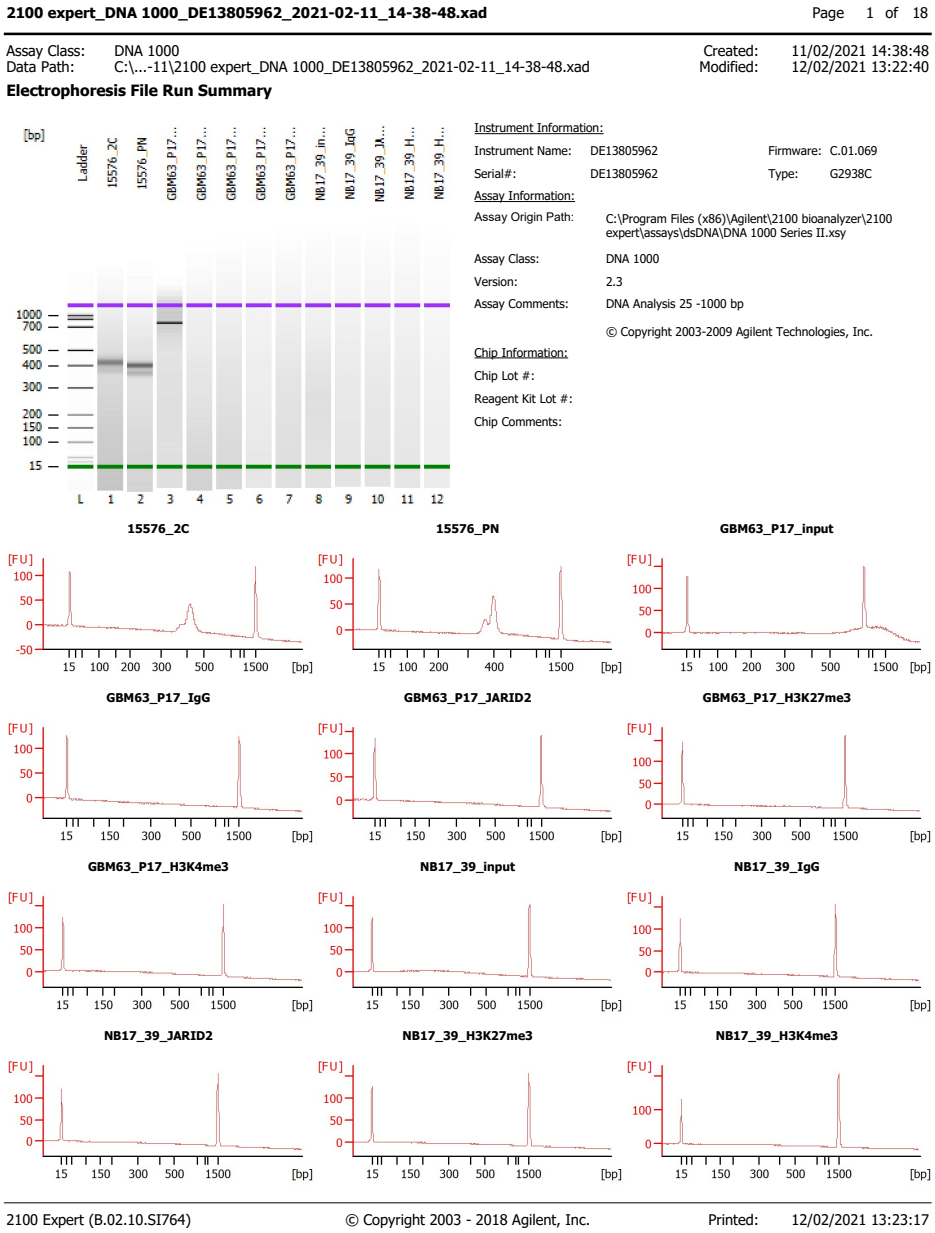
E.2 Emission probability of each state from ChromHMM tool for an in-house dataset

ChromHMM state (Emission order)	EZH2	H3K4me3	H3K27me3	State annotation
1	0.001716529	5.18E-06	0.002019955	Null
2	0.003861848	1.94E-04	0.035216851	Weakly repressed
3	0.011549223	0.023353844	0.820056696	Repressed
4	0.042815413	0.873189627	0.952889896	Bivalent-R
5	0.045083269	0.998411301	0.888182423	Bivalent-A
6	0.013410972	0.949221719	0.023634553	Active
7	0.517699958	0.998867682	0.335493269	Bivalent A+ EZH2
8	0.039354017	0.056239687	0.024649069	Weak bivalent + EZH2

Appendix F

F.1 Agilent 2100 Bioanalyzer DNA 1000 assay

The following traces outline a typical bioanalyzer DNA 1000 assay trace, evaluating the DNA concentrations after DNA purification step of CUT&RUN assay.



F.2 Agilent 2100 Bioanalyzer High Sensitivity DNA assay

The following traces outline a typical bioanalyzer DNA 1000 assay trace, evaluating the DNA concentrations after DNA purification step of CUT&RUN assay.

2100 expert_High Sensitivity DNA Assay_DE13805962_2021-02-12_12-27-53.xad Page 1 of 17

Assay Class: High Sensitivity DNA Assay
Data Path: C:\...gh Sensitivity DNA Assay_DE13805962_2021-02-12_12-27-53.xad

Created: 12/02/2021 12:27:53
Modified: 12/02/2021 13:20:02

Electrophoresis File Run Summary

Instrument Information:

Instrument Name: DE13805962 Firmware: C.01.069
Serial#: DE13805962 Type: G2938C

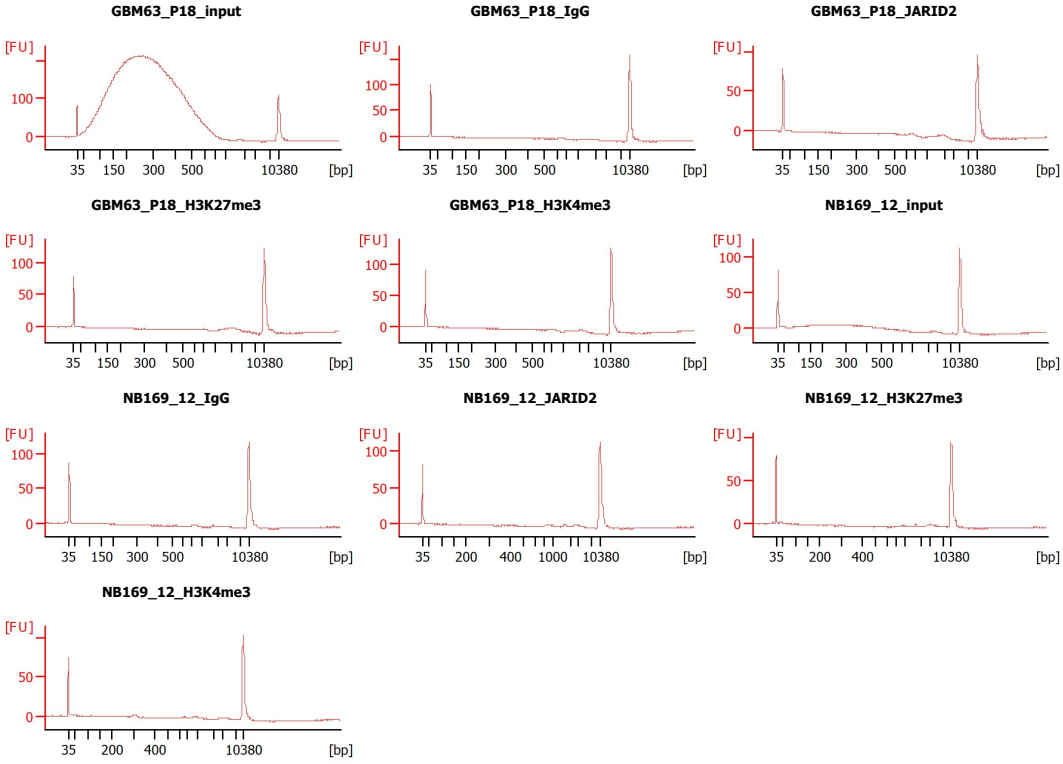
Assay Information:

Assay Origin Path: C:\Program Files (x86)\Agilent\2100 bioanalyzer\2100 expert\assays\dsDNA\High Sensitivity DNA.xsy

Assay Class: High Sensitivity DNA Assay
Version: 1.03
Assay Comments: Copyright © 2003-2010 Agilent Technologies

Chip Information:

Chip Lot #:
Reagent Kit Lot #:
Chip Comments:



2100 Expert (B.02.10.SI764) © Copyright 2003 - 2018 Agilent, Inc. Printed: 12/02/2021 13:21:45