# Exploring the pattern of DNA double strand breaks in glioma cancer stem cells before and after radiation

Sarah Jane Derby

MBChB, MRCP, FRCR

Submitted in fulfilment of the requirements for the Degree of PhD

School of Cancer Sciences
College of Medical, Veterinary and Life Sciences
University of Glasgow

June 2025

# Abstract

**Introduction:** Glioblastoma (GBM) is an incurable brain malignancy with a median survival of 15 months. Glioma stem cells (GSCs) are a radioresistant cell subpopulation promoting recurrence with aberrant DNA damage response (DDR). Radiotherapy prolongs survival and induces DNA double strand breaks (DSBs), however the genome-wide distribution of endogenous and ionising radiation (IR) induced DSBs remain unexplored.

**Aims:** This will be the first study to define the DSB distribution ('breakome') of GSCs and their differentiated progeny before and after radiation.

**Methods:** Breaks ligation in-situ sequencing (BLISS) was performed in GSCs and matched differentiated progeny before and after IR exposure and correlated with RNA-seq, ATAC-seq data and public neural cell and cancer cell BLISS datasets. High DSB density locations, long genes, transcription start and termination sites (TSS & TTS) as well as euchromatin-rich regions were mapped for DSBs. Data were validated using an alternative DSB mapping technique (INDUCE-seq).

**Results:** Few differences in DSB distribution between GSC and differentiated progeny were observed, however DSB frequency was higher in GSCs. *MALAT1* gene had the highest DSB density across all GBM lines. Long genes showed higher absolute DSBs but not higher DSB density. GSCs showed a significantly higher accumulation of DSBs at TTS compared to neural cell lines. Highly transcribed genes showed an increase in DSB density. Both BLISS and INDUCE-seq data displayed higher DSB frequency following IR in differentiated GBM R10 cells but lower DSB frequency following IR in R10 GSCs.

**Conclusions:** DSBs in GSCs and differentiated progeny are not random but show distinct endogenous break patterns. Additionally, DSB frequency following IR is increased in radiosensitive R10 differentiated cells but decreased in radioresistant R10 GSCs relative to unirradiated GSCs. This is likely due to changes in cellular activities including global transcriptional downregulation and improved DDR in GSCs resulting in radioresistance.

# Table of Contents

## Contents

# List of Tables

# List of Figures

# List of Accompanying Material

Supplementary files and figures are available in the appendix. The UMI pattern files that accompany the BLISS commandline files are attached as .txt files. In addition, RStudio scripts are provided per chapter as accompanying material as .R files.

# Acknowledgement

There are so many people to thank and acknowledge as part of this thesis. At times, this has felt like both a marathon and a sprint. I would like to thank Dr Ross Carruthers as my first original supervisor who was hugely instrumental in supporting me to apply for an academic clinical post and in pursuing this PhD. Thank you for hanging in there with me even when you went back to the clinic and taking the time to keep reading and advising on the project. Thank you to Dr Emily Clough who was my M33 lab PhD pal and was always available for support, commiseration and tea in our little original lab group.

Thank you to Prof. Anthony Chalmers for both taking on the role of primary supervisor in my last year and for giving me some hugely beneficial opportunities in both the clinic and academic writing such as taking on the junior principle investigator role for the PARADIGM 2 trial, getting the opportunity to write a review article and to write up the PARADIGM clinical trial paper which was published and presented in EANO 2023. Thank you very much to everyone in the M12 Chalmers lab group: Anthony, Mark, Karen, Karin, Sandy, Katrina and also Dr Conchita Vens for making me feel so welcome when I changed supervisors.

Thank you to Dr Colin Semple, Dr Tracy Ballinger and Dr Stuart Aitken in the Semple lab for the expertise provided in the processing of the BLISS data and for the very generous extended access to the Edinburgh high performance computing cluster.

I would like to give a huge thanks to Dr Mark Jackson who has at times truly felt like a 4$^{th}$ supervisor. He has spent so much of his personal time in reading, critiquing and problem-solving with me for parts of this project. I am certain that I would not have gotten anywhere near as far as I have without his time, interest and patience. Thank you Mark!

I would really like to also thank Karen Strathdee who prepared a number of the sequencing samples for ATAC-seq and BLISS and latterly was instrumental in troubleshooting as I got skilled up for the BLISS protocol. Thank you also to Katrina Stevenson and Karen Strathdee for all of the time taken on the mouse

window model (which never made it into the thesis but was truly a feat in itself!).

Thank you so much to Dr Jo Birch and the M45 lab for adopting me as an honorary lab member with Louise and Kathy: I've loved being in our little lab cubby and enjoying excellent cat chat over a good breakfast roll.

I would like to also thank the Reed lab including Kierney O'Dare and Patrick Van Eijk for facilitating the INDUCE-seq sample troubleshooting, processing and sequencing which became a hugely helpful dataset for the final parts of this project.

Finally, I would like to thank my husband Josh who has easily seen the worst and the best of me in this PhD. Not only that but, having been pregnant during write-up he has truly done so much to keep me sane, calm, well fed and watered. He has been my biggest supporter and most importantly my advocate. Thank you Josh for always caring and pointing me to Jesus. I feel like much of my thesis has been written on prayer (my own, Josh's, my mum's and my church pals), I know that without my faith in Jesus I would have fallen flat on my face at the first hurdle!

And finally, finally I would like to thank our new addition to the family, baby John. You pipped me to the post in arriving first before the thesis did and what a joy you have been. You have put what has felt at times an insurmountable task into perspective. I am realising that this new stage was a test that I could never study for but what a great adventure it is to discover the answers with you and your dad along the way!

To everyone above (and everyone else who didn't get a mention), thank you.

# Author's Declaration

I declare that I am the sole author of this thesis. The work recorded in this thesis is my own and the research described herein was carried out by me unless otherwise stated or acknowledged. No part of this thesis has been submitted for consideration for another degree at the University of Glasgow or any other institution.

*Sarah Jane Derby*

June 2025

# Abbreviations

| | |
|---|---|
| 53BP1 | p53-binding protein 1 |
| a-EJ | Alternative-end joining |
| ANOVA | Analysis of variance |
| ATAC-seq | Assay for transposase-accessible chromatin sequencing |
| ATM | Ataxia telangiectasia mutated |
| ATR | Ataxia telangiectasia and Rad3-related |
| ATRIP | ATR-interacting protein |
| BHP | Benjamini-Hochberg procedure |
| BLESS | Break Labelling, Enrichment on Streptavidin, and Sequencing |
| BLISS | breaks ligation in-situ sequencing |
| bp | Base pairs |
| BRCA1 | Breast cancer gene 1 |
| BRCA2 | Breast cancer gene 2 |
| BWA | Burrow-Wheeler Aligner |
| CFS | Chromosomal fragile site |
| ChIP-seq | Chromatin immunoprecipitation sequencing |
| Chk1 | Checkpoint kinase 1 |
| Chk2 | Checkpoint kinase 2 |
| chr | chromosome |
| CN | Copy number |
| CNV | Copy number variation |
| CSR | Class switch recombination |
| EGF | Epidermal growth factor |
| DAPI | 4', 6-Diamidino-2-Phenylindole, Dihydrochloride |
| DDW | Deionised distilled water |
| DDR | DNA damage response |
| DMEM | Dulbecco's Modified Eagle Medium |
| DMSO | Dimethylsulfoxide |
| DNA | Deoxyribonucleic acid |
| DNA-PKcs | DNA-dependent protein kinase |
| DNase-seq | DNase I Hypersensitive Sites Sequencing |
| dsDNA | Double stranded DNA |
| DSB | Double strand break |
| EDTA | Ethylenediaminetetraacetic acid |
| EGF | Epidermal growth factor |
| EGFR | Epidermal growth factor receptor |
| FCS | Foetal calf serum |
| FDR | False discovery rate |
| FGF | Basic fibroblast growth factor |
| FISH | Fluorescence in-situ hybridisation |
| g | Gravity |
| G4 | G-quadruplexes |
| GBM | Glioblastoma |
| GSC | Glioma cancer stem cells |
| Gy | Gray |
| HCl | Hydrochloric acid |
| HDAC | histone deacetylase |
| Hg19 | Genome Reference Consortium Human Build 19 |
| Hg38 | Genome Reference Consortium Human Build 38 |

| | |
|---|---|
| HR | Homologous recombination |
| Hsp40 | Heat shock protein 40kD |
| HTGTS | high-throughput genome-wide translocation mapping |
| iBLESS | immobilised BLESS |
| IF | Immunofluorescence |
| IGV | Integrative Genome Viewer |
| iHTGTS | Improved HTGTS |
| IR | Ionising radiation |
| IRIF | Ionising radiation-induced foci |
| IVT | In vitro transcription |
| kbp | Kilobase pair |
| kV | kilovolts |
| LAM-HTGTS | linear amplification mediated HTGTS |
| LET | Linear energy transfer |
| Lig III | DNA ligase III |
| Lig IV | DNA ligase IV |
| lincRNA | Long intergenic non-coding RNA |
| lncRNA | Long non-coding RNA |
| LOESS | Locally estimated scatterplot smoothing |
| mA | milliamps |
| MALAT1 | Metastasis associated lung adenocarcinoma transcript 1 |
| MGMT | O6-methylguanine-DNA methyltransferase |
| ml | millilitres |
| M | Moles |
| mm | millimetres |
| NaCl | Sodium Chloride |
| NES | Neuroepithelial stem cells |
| NEU | Post mitotic neural cells. |
| NHEJ | Non homologous end-joining |
| ng | nanogram |
| NPC | Neural progenitor cells |
| NSC | Neural stem cells |
| NuRD | Nucleosome Remodelling and Deacetylase |
| ORA | Over-representation analysis |
| PARP | Poly [ADP-ribose] polymerase |
| PCBP1 | Poly(RC) Binding Protein 1 |
| PCA | Principal component analysis |
| PCI | Phenol/chloroform/isoamyl-alcohol |
| PCR | Polymerase chain reaction |
| PDL | Poly-D-lysine |
| PFA | Paraformaldehyde |
| Polθ | DNA polymerase theta |
| PFGE | pulsed-field gel electrophoresis |
| QC | Quality control |
| qDSBseq | quantitative DSB sequencing |
| RGN | RNA-guided nucleases |
| R-Loop | RNA:DNA hybrid |
| RNA | Ribonucleic acid |
| RNASEH1 | Ribonuclease H1 |
| ROS | Reactive oxygen species |
| rpm | Revolutions per minute |
| RPA | Replication protein A |

| | |
|---|---|
| RS | Replication stress |
| RSS | Recombination signal sequences |
| RT | Room temperature |
| RTP | Reverse transcription primer |
| sBLISS | In suspension BLISS |
| scaRNA | Small Cajal body associated RNA |
| SD | Standard deviation |
| SDS | Sodium dodecyl sulphate |
| snoRNA | Small nucleolar RNA |
| SNPs | Single nucleotide polymorphisms |
| SOX2 | SRY-box 2 |
| SSA | Single strand annealing |
| SSB | Single strand break |
| ssDNA | Single stranded DNA |
| SV | Structural variants |
| TMZ | Temozolomide |
| TopI | Topoisomerase I |
| TopII | Topoisomerase II |
| TopIIß | Topoisomerase II beta |
| TopIIIα | Topoisomerase III alpha |
| TopBP1 | TOPBP1 DNA topoisomerase II binding protein 1 |
| TPM | Transcripts per million |
| TSS | Transcription start site |
| TTS | Transcription termination site |
| U | Units |
| UMI | Unique molecular identifier |
| UTR | Untranslated region |
| VEGF | Vascular endothelial growth factor |
| WGS | Whole genome sequencing |
| WSL | Windows Subsystem for Linux |
| XLF | XRCC4-like factor |
| XPG | Xeroderma pigmentosum group G |
| XRCC1 | X-ray repair cross-complementing protein 1 |
| XRCC4 | X-ray repair cross-complementing protein 4 |
| µl | microlitres |
| γH2AX | γH2A histone family, member X |

# Chapter 1 Introduction

## 1.1 Introduction

Glioblastoma (GBM) remains a highly aggressive and devastating diagnosis facing patients. With a median survival of 14.6 months despite maximum treatment with optimum surgical debulking, concurrent chemoradiotherapy and adjuvant temozolomide (TMZ), patients will inevitably experience recurrent disease (Stupp et al., 2005). GBM remains a cancer of unmet need, with an urgent requirement for improvement in treatment efficacy. Few significant scientific advances have been made to improve outcomes in the past two decades with the most significant being the addition of concurrent and adjuvant TMZ to radiotherapy treatment published by Stupp et al in 2005 (Stupp et al., 2005).

Radiotherapy is an important mainstay in treatment of GBM and primarily acts to kill cancer cells through the induction of double strand breaks (DSB) in deoxyribonucleic acid (DNA) resulting in apoptosis or mitotic catastrophe (Chadwick and Leenhouts, 1973, Glücksmann and Spear, 1939). The majority of DSBs are induced via reactive oxygen species (ROS), though direct ionising radiation (IR) DSBs can also occur (Petkau, 1987). The induction of single strand breaks (SSBs) is far more common following IR; however, SSBs are usually repaired efficiently. SSBs can also be converted into DSBs when two SSBs occur close together (Ma and Dai, 2018).

Despite exposure to a cumulative dose of 60 Gray (Gy) of fractionated radiotherapy, GBM tumours repopulate, usually within the treated volume of the tumour. Higher doses of radiotherapy are not practical due to the significant impact on normal brain function (Dropcho, 1991). GBM tumours appear to repopulate from a subset of cells known as glioma cancer stem cells (GSCs), so-named for their ability to self-renew and produce progeny of multiple lineages. In addition, these GSCs have demonstrated aberrant upregulated repair pathways, specifically in the DNA damage response (DDR) to DSBs (Bao et al., 2006).

Many pre-clinical therapies have been taken forward in to clinical trials which have unfortunately been unsuccessful including use of tyrosine kinases such as

EGFR inhibitors, angiogenesis inhibition with anti-VEGF therapies, targeting DDR pathways and anti-migratory therapies (Cruz Da Silva et al., 2021). Unfortunately, given the heterogeneity of the disease and the nature of GBM changing over time in response to therapy, long-lasting effective treatments demonstrating sustained response are yet to be found (Nathanson et al., 2014). A better understanding of the underlying mechanisms causing GBM treatment resistance is required to develop better treatment strategies to improve survival in this highly aggressive cancer.

## 1.2 Glioblastoma

The cause of GBM remains ambiguous. A small number of GBMs have been linked to heritable syndromes such as neurofibromatosis, Li-Fraumeni and Turcot syndrome or brain irradiation at an early age (Neglia et al., 2006, Louis et al., 2021, Bougeard et al., 2015, Smith et al., 2024). However, in the majority of GBM tumours, the cause of the origin remains enigmatic. GSCs have been postulated as the potential originators of the tumour (Watanabe et al., 2024, Lee et al., 2018). These cells have also been hypothesised to arise from mutated neural stem cells (NSCs) in the subventricular zone (de Almeida Sassi et al., 2012, Lee et al., 2012). Whilst there has been demonstration that NSCs appear capable of developing the oncogenic potential to propagate tumours, there are disagreements on whether NSCs are truly the originators (Lee et al., 2012). For example, pre-clinical investigations have also demonstrated that differentiated neural cells such as neurons and astrocytes are capable of reverting back to a dedifferentiated form which may imply that GSCs could arise from dedifferentiated cells as opposed to NSCs (Friedmann-Morvinski et al., 2012).

The risk of GBM increases with age with a peak of incidence at 75-84 years, however it can affect patients of any age and remains a devastating diagnosis (Ostrom et al., 2017). GBM more commonly affects men and GBMs can occur across all regions of the brain, though most commonly affects the frontal, temporal and parietal lobes (Wirsching et al., 2016). An important molecular characteristic relevant to survival in GBM includes the GBM methylation status. MGMT (O6-methylguanine-DNA methyltransferase) methylation status refers to the activity of MGMT on removing the lethal O6-alkylguanine lesions caused by TMZ chemotherapy. Tumours that are found to be MGMT methylated are more

sensitive to TMZ treatment and as a result have an improved median overall survival of over 20 months (Hegi et al., 2005). Around 45% of patients will exhibit MGMT methylated tumours.

### 1.2.1 Glioma cancer stem cells

A key underlying problem in the treatment challenges faced in the management of GBM is the aggressive subset of cells known as GSCs. GSCs show ability to self-renew, produce progeny of multiple lineages and are capable of initiating tumours in xenograft models (Safa et al., 2015). Importantly, in addition to the ability to establish tumours and self-renew, these cells are in possession of a number of methods that characterise the aggressive nature of GBM tumours including radiation resistance, promoting invasive phenotypes through upregulation of migration and resisting apoptotic signals (Birch et al., 2018, Ivanov and Hei, 2014, Carruthers et al., 2018).

Cancer stem cells have been widely described in several cancers and provide an explanation for the challenging heterogeneity seen in many cancers resulting in treatment resistance and tumour recurrence (Shackleton et al., 2009, Lathia et al., 2015). Whilst cancer stem cells have been an important subject of study to GBM, cancer stem cells have been described in a number of tumour types including breast, stomach and colorectal cancer to name a few (Humphries et al., 2013, Balic et al., 2006, Takaishi et al., 2009). Interestingly, whilst several cancers appear to conform to a cancer stem cell model, there remain some cancers that do not adhere to the cancer stem cell originator theory. Melanoma has been researched with regards to cancer stem cells, however cancer stem cells are not characteristic of melanoma tumours. Rather than a specific subset of cells being capable of tumour initiation, Quintana et al. (Quintana et al., 2008) observed that upwards of 25% of unselected melanoma cells demonstrated tumour propagating capability. Melanoma cells that express stem cell markers also do not show phenotypic differences when it comes to tumour propagating potential.

Whilst the presence of cancer stem cells has been established in many cancer types, quantifying and isolating these cell subgroups can be controversial. In GBM, the surface marker CD133+ has been widely associated as a cancer stem

cell marker and is indicative of a cell's capability to form neurospheres as well as form tumours in xenografts (Safa et al., 2015). However, despite CD133 positivity being widely agreed as an indication of stemness in GSCs, CD133 positivity is not universal in GSCs. Some populations of cells have also been shown to possess the same properties of cancer stem cells despite no CD133 expression (Tirino et al., 2013, Beier et al., 2007).The CD133 expression on cancer cells has also been shown to be influenced by the microenvironment in which CD133 loss and re-expression can occur depending on culture conditions (Yang et al., 2012). There are a number of other markers that have also been discussed in the context of GSCs associated with cancer stem cell-like properties such as nestin, SOX2, CD44, OLIG2, integrin-α6 and L1CAM (Pietras et al., 2014, Lathia et al., 2010, Bulstrode et al., 2017, Singh et al., 2017, Bao et al., 2008). However, despite the preponderance of these markers to associate with stem-like characteristics, they are not specific to GSCs and may have varying levels of expression even in established GSCs. This presents a challenge in the consistent identification of GSC populations as there is no universal GSC marker. Additionally, GSCs have an association with NSCs, though the relationship between GSCs and NSCs can be controversial. NSCs share a number of markers with GSCs, including SOX2, OLIG2, CD144 and CD133 (Brescia et al., 2012). NSCs have been postulated as originators of GSCs, resulting in the flexible phenotype and abilities to propagate tumours (Goffart et al., 2013, Safa et al., 2015). There is also some debate however that, rather than originating from NSCs, GSCs are cells that have undergone de-differentiation to a more stem-like state (Friedmann-Morvinski et al., 2012). Regardless of origin state, the parallels to NSCs remain important for consideration in understanding GBM tumour development.

### 1.2.2 Modelling GSCs *in vitro*

In order to model GSCs *in vitro,* cells need to be maintained under strict conditions to optimise an undifferentiated stem-like state. For this, GSCs must be cultured in serum-free media and supplemented with growth factors (epidermal growth factor: EGF and basic fibroblast growth factor: FGF) (Lee et al., 2006). Patient derived primary GSC lines were gifted by the Watts' laboratory in Cambridge (Fael Al-Mayhani et al., 2009). From these, Dr Ross Carruthers established matched differentiated progeny cells for comparison with

primary GSC lines (Carruthers et al., 2018). For these comparator cell lines, cells were differentiated from primary GSC lines by supplementing with foetal calf serum (FCS). These cells were investigated for a number of the above markers described. GSCs demonstrated significantly higher levels of nestin and SOX2 on western blot and significantly higher CD133+ staining cells than differentiated cell populations (Carruthers et al., 2018). Clonogenic assays of GSCs and differentiated cells demonstrated that GSCs were significantly more radioresistant compared to differentiated cells. In addition, CD133+ GSCs demonstrated a significantly greater ability to form neurospheres compared to CD133- cells (From the thesis of Dr Carruthers "Response to ionising radiation of glioblastoma stem-like cells"). Furthermore, when comparing GSC and differentiated populations with regards to the DDR, GSCs demonstrated significantly elevated phosphorylation of key repair proteins Ataxia telangiectasia mutated (ATM), Ataxia telangiectasia and Rad3-related (ATR) and Checkpoint kinase 2 (Chk2) following exposure to 5 Gy IR.

This GSC versus differentiated progeny model has the advantage of providing an internal control for radiosensitivity for GSC patient derived lines. This allows for a useful comparator when looking to investigate populations of contrasting radiosensitivity. However, this model has limitations in terms of reflecting the GSC and non-GSC populations seen in GBM tumours. Whilst GSC populations have been shown to be significantly more enriched for CD133+ and for other GSC markers, it is known that these GSC populations will harbour cells that are not expressing CD133. Similarly, whilst differentiated populations demonstrate a significant reduction in CD133 expression, some expression remains indicating that a small GSC population may be persistent. Additionally, GBM tumours are notoriously hypoxic sites, however these cells have been grown in standard tissue culture conditions with atmospheric oxygen and carbon dioxide, both of which could impact on the relative radiosensitivity of cells. As these cell lines are paired in their growth conditions other than enrichment media, this goes some way to limiting these challenges by way of the allowance for relative comparison across groups. This in itself is also a potentially confounding factor given that the different growth conditions between media also influence comparison. However, phenotypic changes have been demonstrated in GSC populations when sorted for CD133 positivity, giving confidence in the model

(Carruthers et al., 2018)(See also Dr Emily Clough's thesis: "Investigating mechanisms and indicators of sensitivity to replication stress-targeting therapies in glioblastoma").

### 1.2.3 GSCs and the DNA damage response

GSCs have upregulation of the DDR including particular key proteins such as ATM, checkpoint kinase 1 (Chk1), Poly [ADP-ribose] polymerase 1 (PARP-1) and ATR (Ahmed et al., 2015, Bao et al., 2006). This upregulation of DDR proteins allows a rapid response to DNA damage and, importantly, DSBs. Cancer stem cells are able to mediate an efficient response to DNA damage which appears to contribute to the treatment-resistant phenotype (Facchino et al., 2010). This has been demonstrated in CD133+ cells where GSCs demonstrate improved checkpoint inhibition via phosphorylation of Chk1 and Chk2 with a more rapid resolution of DSBs in CD133+ cells compared to CD133- cells (Bao et al., 2006). Interestingly, whilst these CD133+ cells have been demonstrated to have a more rapid repair, research suggests that GSCs have a predominant reliance upon homologous recombination (HR) pathways over non-homologous end joining (NHEJ) (Lim et al., 2014). Lim et al described this preferential repair of DSBs using HR in tandem with a non-functional G1/S checkpoint which should result in slower repair of DSBs. However preferential repair via HR may also act as a more robust repair mechanism, particularly in the context of complex repair which may be required following IR or as a means of resolving replication stress (RS) (Lim et al., 2012). Interestingly, both neural progenitor cells and glioma-initiating cells in this study demonstrated elevated Rad51 expression, a marker for HR. However, it was noted that the neural progenitor cells showed less preference for HR compared to the glioma-initiating cells which, when treated with ATM inhibitor to block HR, demonstrated greater levels of DNA damage following IR compared to neural progenitors.

### 1.2.4 GSCs and replication stress

Drivers of the upregulated DDR in GSCs have been investigated including RS (Carruthers et al., 2018, Zhang et al., 2021). RS occurs when the replication machinery pauses, due to obstacle, dysregulated origin firing or failure to supply the necessary components for DNA replication (Zeman and Cimprich, 2014). The

DNA continues to unwind and unzip, leaving open vulnerable stretches of single stranded DNA (ssDNA). Replication protein A (RPA) localises to sites of ssDNA, indicating a stall, signalling to DDR proteins ATR interacting protein (ATRIP), ATR and other DDR proteins to initiate HR to allow for fork restart (Shiotani and Zou, 2009, Frattini et al., 2021). Causes of RS include replication-transcription machinery collisions, deregulation of origin firing, oncogene activation or DNA lesions such as DNA:RNA hybrids (R-Loops) and G-quadruplexes (G4) (Skourti-Stathaki et al., 2014, Eddy et al., 2011, Zhong et al., 2013). Whilst the term RS covers several events as above, it is usually used to describe the slowing or halting of replication fork machinery and subsequent generation of single stranded DNA. HR is the preferred mechanism of resolution of RS (Nickoloff, 2022, Arnaudeau et al., 2001). Stalled replication forks, if left unrepaired, will lead to fork collapse, replication runoff and potentially the formation of DSBs (Strumberg et al., 2000). Replication runoff is where a replication fork comes into contact with an unrepaired SSB, replication subsequently stalls and can result in replication fork collapse and a subsequent single ended DSB (So et al., 2017). Therefore, resolution of these events is highly important to the GSC survival, otherwise cells will risk mitotic catastrophe. Elevated levels of RS trigger the key protein ATR to initiate DSB repair, priming the pathway to rapidly respond to exogenous damage. Resolution of RS occurs via co-localisation of RPA to stall sites, initiating ATRIP and ATR, and HR required for fork repair (Shiotani and Zou, 2009). Interestingly, elevated RS levels have also been identified in neural progenitor cells which may confer further important shared mechanisms between GSCs and neural cells (Wei et al., 2016).

## 1.3 DNA double strand breaks

DNA DSBs are arguably one of the greatest threats to genome stability and have the potential to induce chromosomal instability, cell death and cancer development (Pfeiffer et al., 2000). DSBs are known to be highly deleterious with even a single persistent unrepaired DSB being capable of triggering cell death in normal cells (Featherstone and Jackson, 1999). However, DSBs are not solely pathological and there are cell processes which require the induction of DSBs to be completed. Two such categories are the act of promoting genetic recombination and the modulation of the shape and structure of DNA.

## 1.3.1 Physiologically induced DSBs

When considering physiological DSBs, this is referring to DSBs that have been induced as part of planned cell activity. These are DSBs that are important to the continuation of normal cell processes, though they can be co-opted as part of pathological activity or indeed be mis-repaired resulting in genomic instability. Regarding physiological DSBs these can be broadly identified as pertaining to two categories: DSBs to promote the recombination of genetic material and DSBs to modulate the DNA shape and structure (Cui and Meek, 2007, Neale and Keeney, 2006). These will be described in brief below and for a more detailed summary these processes the reader may refer to the review article "Physiological Roles of DNA Double Strand Breaks" by Khan (Khan and Ali, 2017).

### 1.3.1.1 DSBs promoting genetic recombination

Briefly, physiological DSBs that promote genetic recombination include meiotic recombination, class switch recombination (CSR) and V(D)J recombination (Khan and Ali, 2017, Ramsden and Gellert, 1995, Dunnick et al., 1993).

Meiotic recombination is the process in sexual reproduction where genetic material between two homologous chromosomes is exchanged as a means of diversification of genetic material for future progeny (Ohkura, 2015). Much of this process shares steps with HR and indeed also requires Rad51 for invasion of single stranded DNA into the recombination double stranded DNA. Errors in meiosis will result in mis-segregation and either failure of viability or defective DNA.

With regards to V(D)J recombination, this refers to the act of the adaptive immune response to develop a library of immunoglobulin antigen-receptors and T cell receptors from immunoglobulin chains (Jung and Alt, 2004). These chains have three exon segments that can be recombined: variable (V), diversity (D) and joining (J) segments. Recombination signal sequences (RSS) highlight locations where recombination sites are to occur, allowing DNA hairpins to form and be cleaved, creating DSBs (Ma et al., 2002). This allows the immune system to provide a wide variety of immune responses from the immunoglobulin

information provided. Errors in RSS pairing have been associated with lymphomas as has the aberrant formation of DSBs (Papaemmanuil et al., 2014, Zhang et al., 2011).

Finally, CSR is the process by which DNA is rearranged specifically in B-cells during B-cell maturation to make antibodies depending on cytokines (Chaudhuri and Alt, 2004). This allows for antigen specificity from a generic IgM expression to a more specific immunoglobulin isotype such as IgA, IgG or IgE. For this recombination to occur, DSBs must be created and re-ligated to facilitate the relevant immunoglobulin. Errors in CSR have been associated with the disorder ataxia-telangiectasia and genomic instability from inappropriate DSB induction and re-ligation (Reina-San-Martin et al., 2005, Stavnezer et al., 2008).

#### 1.3.1.2 DNA structure-related DSBs

Regarding physiological DSBs related to DNA structure, these can broadly be described as relating to intervention in DNA replication or transcription, maintaining chromatin architecture and in supervision of mitochondrial DNA.

The use of DSBs to intervene in DNA replication and transcription refers to the need for constant access to DNA which requires unwinding and unzipping, putting the double helix under torsional stress. This topological stress on the DNA is managed by a family of proteins known as topoisomerases which induce SSBs or DSBs in the DNA to relieve the pressure of torsional forces. The stress on the DNA comes from unwinding and unzipping of long stretches of DNA from replication and transcription activity. As the DNA unwinds permitting access to the DNA, this increases the torsion on the DNA behind it resulting in negative supercoiling. This negative supercoiling exerts increased pressure on the base pair bonds which become weakened and are at risk of buckling. To prevent this, topoisomerases (type II) induce DNA DSBs in order to relieve torsional stress, allowing for the continuation of elongation of replicating DNA or transcription elongation (Schoeffler and Berger, 2008). Regions of unreplicated DNA occurring between replication forks also undergo torsional stress. This happens where, two replicating regions are close together and the intervening sequence experiences pressure from positive supercoiling from the torsional forces on both sides. These topoisomerases can also induce DSBs to relieve the pressure from positive

supercoiling (Brill et al., 1987). Topoisomerases will also act to untangle DNA by inducing DSBs and allow for better accessibility to regions of DNA needing accessed (Durand-Dubief et al., 2011).

Topoisomerases also appear to play an important role in the maintenance of chromatin architecture, although the exact role of topoisomerases is yet to be fully described. High levels of Topoisomerase II (TopII) are present in mitosis and associate closely with chromatin loops: these are stretches of DNA present on the same chromosome but which are physically closer to one another than the DNA sequences in between the two sites (Gasser et al., 1986, Valdés et al., 2018). Topoisomerases have also been associated with nucleosome disassembly and appear to facilitate chromatin mobility through the removal or movement of nucleosomes, facilitating nucleosome-free region formation (Durand-Dubief et al., 2010).

Mitochondrial DNA also suffers from the topological stress experienced in replication and transcription. Mitochondrial DNA is known to interact with three topoisomerases: namely TopI, TopIIß and TopIIIα (Khan and Ali, 2017, Sobek and Boege, 2014). It remains uncertain the exact role that DSB induction has on mitochondrial DNA, though it is reasonable to surmise that activity on relieving torsional stress and promoting DNA access occurs (Pommier et al., 2022). Interestingly, there is emerging evidence that mitochondrial topoisomerases may play some role in the ageing process, where it has been noted that enzyme dysfunction can contribute to age-related decline (Tsai et al., 2016).

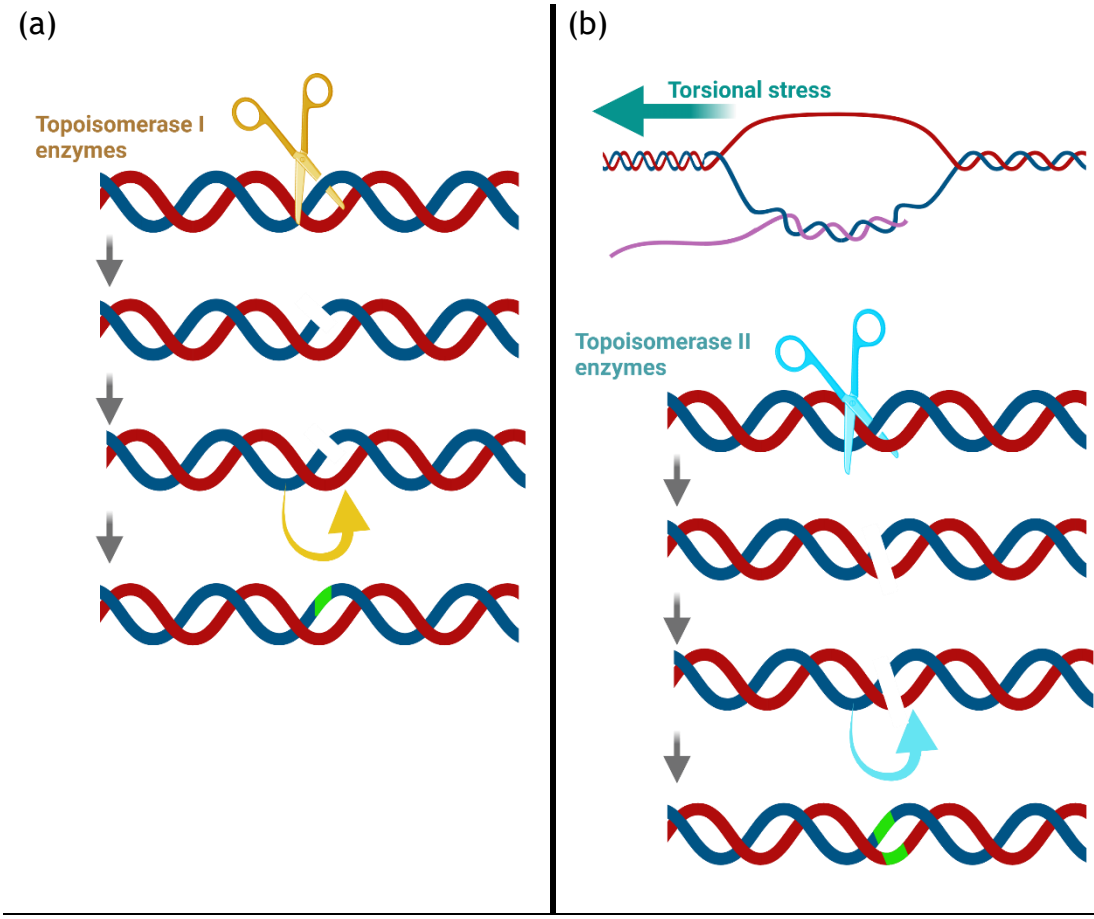



**Figure 1.1 The activity of topoisomerases in induction of DNA breaks**

(a) Topoisomerase I enzymes act to induce SSBs in the DNA to allow mobility of the broken strand to pass around the intact strand and so relieve pressure exerted on the DNA double helix (Type 1a) or to allow controlled rotation (Type 1B/1C). (b) TopII enzymes induce double stranded breaks in the DNA to relieve supercoiling, allow for unlinking of DNA chains or to untangle knots of DNA. Topoisomerases make a double cut to allow DNA relaxation, relief of torsional stress and subsequent ligation. An example of negative supercoiling is shown above at an area of transcriptional elongation. Images created using BioRender.com.

## 1.3.2 DDR to DSBs

As discussed, the DDR plays a vital role in the maintenance of genome integrity and cell survival. It has also been established that GSCs have upregulation of aberrant DDR with a preference toward utilising HR for DSBs. There are two primary pathways by which DSBs are repaired; HR and NHEJ as well as two less commonly utilised pathways; alternative end-joining (a-EJ) and single strand annealing (SSA). For an in-depth review of the DDR to DSBs the reader may refer to the review article by Brandsma and Gent “Pathway choice in DNA double strand break repair: observations of a balancing act” (Brandsma and Gent, 2012).

#### 1.3.2.1 Non homologous end-joining

The NHEJ DSB repair pathway can occur at any point during the cell cycle and provides a more rapid form of repair than that of HR (Figure 1.2). The action of NHEJ is more error-prone due to the lack of a template strand, however, allows for rapid DSB resolution. Briefly, in NHEJ, the DSB triggers the binding of the Ku70/80 heterodimer to the DSB ends, Ku in turn signals for the binding of DNA-dependent protein kinase (DNA-PKcs) (Mimori et al., 1986, Liang and Jasin, 1996). This then signals for the binding of Artemis, which can act to trim DSB ends, removing overhangs that are incompatible with repair (Ma et al., 2002). Following DSB end trimming, the DNA ligase IV (Lig IV) acts to ligate the resected DSB ends together as part of a protein complex with X-ray repair cross-complementing protein 4 (XRCC4) and XRCC4-like factor (XLF), thereby repairing the DSB (Andres et al., 2007). The DDR protein p53-binding protein 1 (53BP1) also has activity in localising to DSB locations to signal for repair and has some protective activity against resection of DSB ends which can result in a favouring of NHEJ (Bothmer et al., 2010).

#### 1.3.2.2 Homologous recombination

The HR repair pathway is limited to late S and the G2 phase of the cell cycle, due to the need of a template strand for reference, however HR provides a more accurate repair mechanism and prevents the loss of DNA sequences that can occur during NHEJ (Figure 1.2). In brief, following the occurrence of a DSB, DNA end resection occurs via the MRN-CtIP complex, resulting in 3' single strand overhangs. These 3' ends signal for RPA to coat the single stranded DNA (ssDNA) whilst the MRN-CtIP complex signals ATM recruitment which will phosphorylate DDR proteins to signal repair including γH2A histone family, member X (γH2AX) (Burma et al., 2001, Limbo et al., 2007). At this point Rad51, facilitated by breast cancer gene 2 (BRCA2), is signalled by RPA where it replaces RPA by associating with the ssDNA. Rad51 promotes access into the sister chromosome where the ssDNA invades and finds the correct complementary sequence, forming a D-loop structure (Jain et al., 1995, Holloman et al., 1975). This can occur on one or both strands from the DSB. The formation on both strands is known as a Holliday junction (Holliday, 1964). With the newly paired single stranded 3' ends in position, the DNA is then replicated from the template

strand to rebuild the lost region of DNA between the two DSB ends. Once this has covered the distance between ends, the single stranded 3' DNA dissociates from the template DNA and the newly synthesised DNA is ligated, restoring the DSB DNA region. More specifically, at stalled replication forks, the DDR response is mediated primarily by ATR-ATRIP following RPA recruitment to the single stranded DNA. ATR acts to signal checkpoint arrest via Chk1 phosphorylation as well as signalling for TOPBP1 DNA topoisomerase II binding protein 1 (TopBP1) along with Rad9-Rad1-HUS1 clamp protein as part of replication fork repair (Brown and Baltimore, 2003, Kumagai et al., 2006, Bermudez et al., 2003).

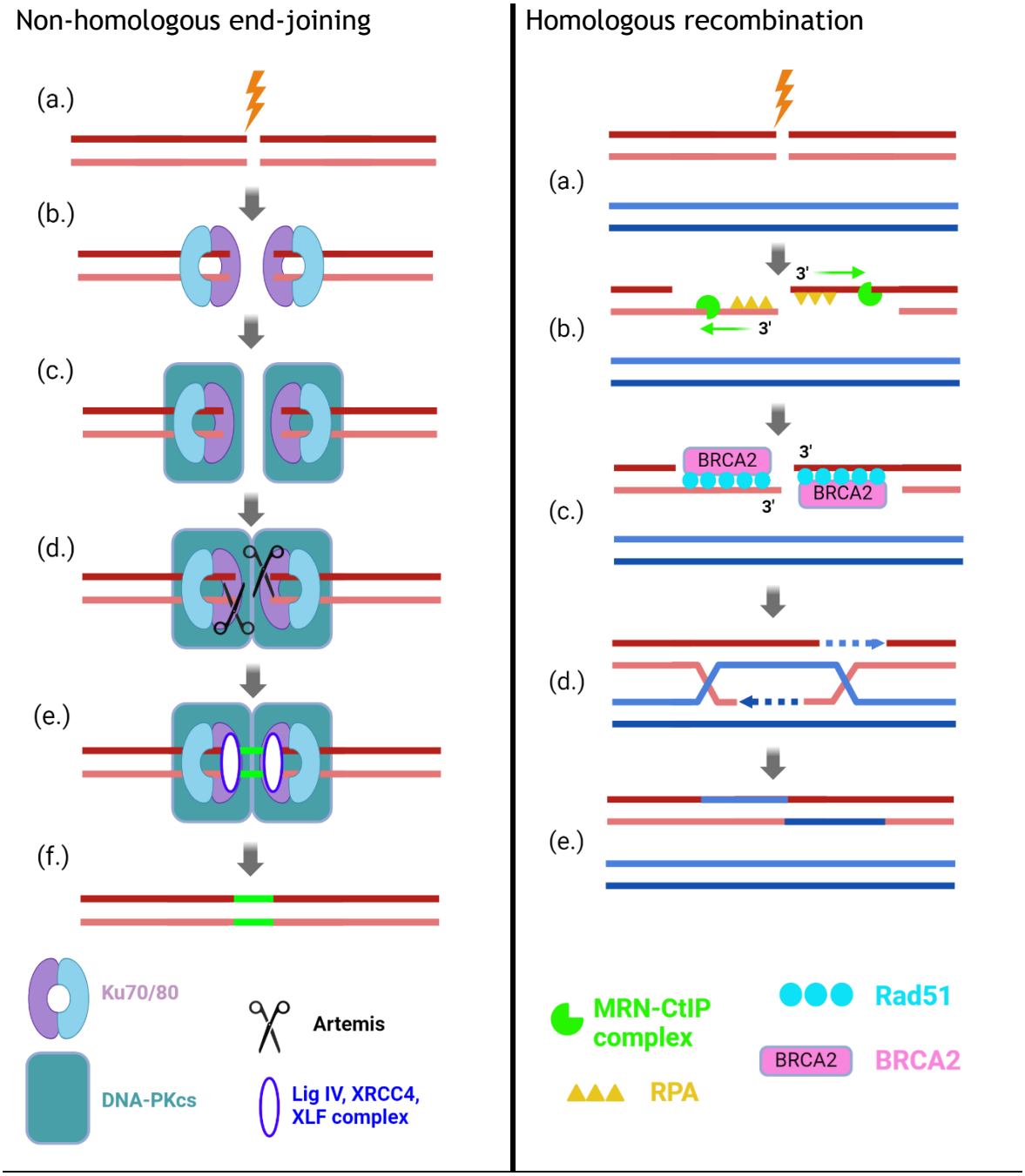



**Figure 1.2 Non homologous end-joining and homologous recombination DSB repair schematics**

Non homologous repair pathway: (a) Induction of DSB not requiring sister chromatid. (b) Ku70/80 heterodimer binds to DSB ends. (c) DNA-PKcs localises to Ku70/80. (d) Artemis trims DSB ends. (e) Lig IV ligates resected DSB ends alongside XRCC4 and XLF.

Homologous recombination pathway: (a) DSB in the presence of sister chromatid. (b) DSB end resection from 3' ends via the MRN-CtIP complex, RPA coats 3' ssDNA overhangs and signals for DNA repair proteins including BRCA2 to localise. (c) Rad51, facilitated by BRCA2 acts to promote invasion of 3' overhang into sister chromatid. (d) Holliday junction is formed and DNA from sister chromatid is used as a template for DNA synthesis from invaded 3' ends. (e) DNA is re-ligated and DNA is restored.

Images created using BioRender.com.

### 1.3.2.3 Alternative end-joining and single strand annealing

Other pathways for DSB repair also include a-EJ, also known as microhomology-mediated end joining, and SSA. The a-EJ pathway acts primarily in S-phase and is predominantly considered a backup repair process, though interestingly has recently been seen to have an association with management of IR-induced damage (Liang et al., 2005, Dutta et al., 2017). It also appears that a-EJ is more prevalent in HR-deficient cells and is much more error-prone. The a-EJ pathway requires recruitment of XRCC1 and DNA ligase III (Lig III) to the DSB location followed by recruitment of PARP and Polθ (El-Khamisy et al., 2003). In contrast to HR, a-EJ uses 3' single stranded DNA of very short microhomologies of between 5 and 25 base pairs (bp) which are then matched on both sides. This process can result in huge loss of DNA and generate genomic instability . Regarding SSA, this pathway is also similar to a-EJ, however, rather than requiring PARP and Polθ, relies on recruitment of Rad52 (Bhargava et al., 2016). The SSA also participates in in 3' end resection like a-EJ and HR, however SSA resections are considered to be much longer than a-EJ regions, with much longer annealed regions compared to a-EJ of around 100 bp. An increase in reliance on SSA has been associated with an increased risk of breast and ovarian cancer, likely related to mutations in BRCA2 and low expression of Rad51 (Tutt et al., 2001).

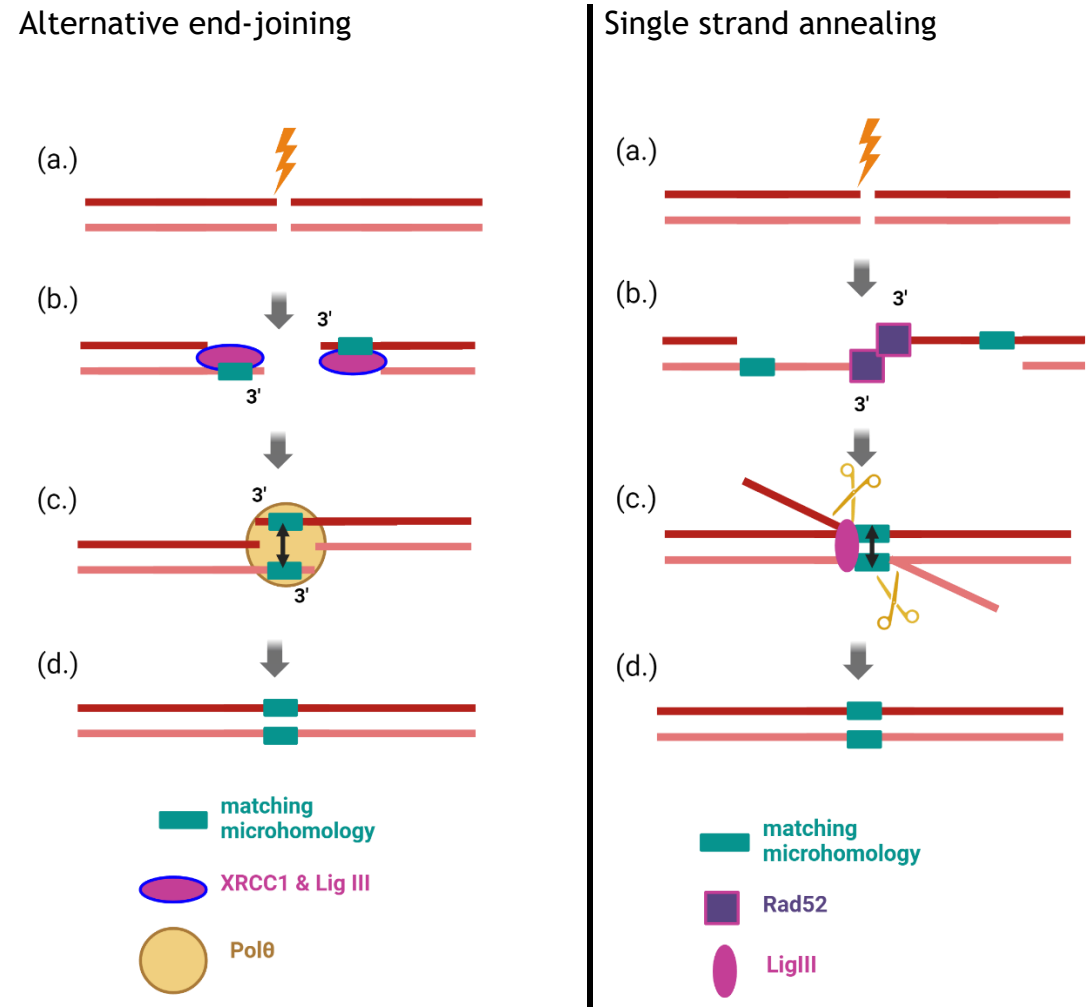



**Figure 1.3 Alternative end-joining and single strand annealing DSB repair schematics**

Alternative end-joining (a) Induction of DSB. (b) Short microhomology 5-25 bp in length found on matching strands via XRCC1 and Lig III. (c) PARP and Polθ recruitment to microhomology for matching and ligation of DSB. (d) DSB repaired with loss of intervening DNA.

Single strand annealing: (a) Induction of DSB. (b) Rad52 recruited in place of Polθ with longer end resection of 3' overhangs for long homologous sequence ~100 bp length. (c) Annealing of homologous sequence via Lig III and trimming of intervening DNA sequences. (d) Loss of large DNA regions resulting in DNA loss and genomic instability.

Images created using BioRender.com

## 1.3.3 DSBs in GSCs

Dysregulation of the normal DDR is a hallmark of cancer and an established feature of GBM and GSCs (Bartkova et al., 2010, Ahmed et al., 2015). For GSCs, DSBs are a double-edged sword. Whilst DSBs risk cell death, DSBs also promote genomic instability which give potential avenues to increase invasive potential and treatment resistance (Valdiglesias et al., 2013). Location of DSBs have been examined across normal tissue and cancer cells, however the genomic locations

of DSB clusters in GSCs has not been investigated (Ballarino et al., 2022, Ballinger et al., 2019, Yan et al., 2017). Clearly, DSB location could have significant impacts on cell phenotypes and survival and is therefore of key potential interest in understanding how DSBs contribute to the GSCs treatment resistance. Interestingly, when compared to normal astrocytes, glioma cells appear to display slower repair kinetics with delayed resolution of DSB repair marker yH2AX (Short et al., 2007). GSCs also exhibit a preference for the HR pathway and display relatively high levels of Rad51 expression compared to astrocytes (Short et al., 2007, King et al., 2017). This combination of elevated DSB repair proteins and concurrent slower repair kinetics poses an interesting dilemma where it appears that GSCs are both prepared for, but slower to respond to, DNA damage than normal cells. As previously mentioned, the replication fork and DSB repair protein ATR is also elevated in GSCs, once again apparently contributing to readiness for DSB repair, whilst at the same time tolerating higher levels of RS which have the potential to lead to replication fork collapse and DSBs (Carruthers et al., 2018). Whilst not investigated in this thesis, an important consideration in GBM and DSBs is the relatively recent finding that 50% or more of GBMs have signs of having undergone catastrophic genomic rearrangements via chromothripsis (Cortés-Ciriano et al., 2020). Chromothripsis can be described as genome shattering, where an apparent single event leads to huge rearrangements in the genome. The underlying driver mechanism of GBM is unclear, however given the rapidity of GBM development, chromothripsis has been postulated as a potential underlying cause (Ah-Pine et al., 2021). Whilst these multi-site events may be rare, the surviving cells post-genome rearrangement may also be important contributors to underlying genomic instability and heterogeneity of the GBM landscape, allowing for adaptation and cell survival.

### 1.3.4 DSBs in ionising radiation

IR is well known to result in both SSBs and DSBs. It is widely accepted that the primary cause of cell death from IR is from DSBs rather than SSB lesions. Following 1 Gy of IR, thousands of SSBs will be induced with a comparatively low number of approximately 40 DSBs induced per cell (Nickoloff et al., 2020, Rothkamm and Löbrich, 2003). These DSB lesions can be caused either as a direct effect of IR on the DNA or, more commonly, as an effect of ROS, on the

DNA (Borrego-Soto et al., 2015). In clinical practice, the intention of treatment of IR is to cause sufficient damage to rapidly replicating cancer cells in order to cause cell death, whilst allowing healthy cells to repair adequately. This balance of cancer cell kill and healthy cell preservation can be a challenging one to strike and is described as the "therapeutic window". This is why aberrant upregulated DNA DSB repair mechanisms in cancer can be such a challenge in radioresistance. Whilst cancer cells will generally have poorer repair capacity than healthy cells, GSCs have better than average repair capacity compared to the majority of GBM cells. With the repair capacity of GSCs being closer to that of normal tissue, IR alone becomes a less effective means of targeting these radioresistant cells. Conversely, whilst upregulated DSB repair and, in particular, HR infers radioresistance, cancer mutations impacting on DSB repair can also result in radiosensitivity such as in ATM, XRCC3 and Rad51 (Vral et al., 2011).

In IR-induced DSBs, the majority of these are isolated sites (~80%), however IR is also known to be capable of inducing highly complex or clustered lesions (~12%) (Ward, 2000, Georgakilas et al., 2013). These complex lesions pose a greater challenge for repair compared to isolated or simple DSBs. These complex sites can be described as more than one lesion occurring within 10-20 base pairs of each other and result in greater lethality and mutagenicity (Nickoloff et al., 2020, Blaisdell et al., 2001). Importantly this is in the context of photon radiation; low linear energy transfer (LET) IR such as radiotherapy. However high LET IR such as proton and carbon ion therapy is increasing in use in clinical practice and so can also be considered as important in terms of IR-induced DSB damage. High LET IR has been proposed as inducing greater levels of clustered DNA damage and interestingly appear to form larger foci of the DSB DDR marker 53BP1 high LET versus low LET IR (Penninckx et al., 2019). Additionally, as with complex IR-induced lesions, DSB damage from high LET IR shifts DDR pathways towards preferential HR (Allen et al., 2017).

## 1.4 Cellular features contributing to DSBs

In addition to the physiologically induced DSBs and DSBs caused by IR, features of cell activity, DNA structure and cell cycle can also affect the risk of developing DSBs. Human cells can experience up to 50 endogenous DSBs per cell per day, hence the need for such robust repair processes (Vilenchik and

Knudson, 2003). Important processes such as replication and transcription can risk DSBs, and additionally the DNA sequence itself can also cause or contribute to the DSB risk.

### 1.4.1 Non-canonical structures and at-risk DNA regions

Non-canonical structures and at-risk DNA regions have been described as potential contributors to DSBs. Structures that have been associated with DSBs include “R-loops”: which are triple stranded RNA:DNA hybrids primarily associated with sites of active transcription and the associated negative supercoiling that occurs behind the transcription machinery (Roy and Lieber, 2009, Roberts and Crothers, 1992). R-loops are also known to be associated with increased RS and DNA damage and can form at transcription-replication collision sites at chromosomal fragile sites (CFSs) on ‘very long genes’(Helmrich et al., 2011). Chromatin is also hindered from correctly condensing around CFSs containing R-loops. This could both be considered a problem in preventing normal condensing but has also been suggested as providing some physiological benefit that allow better access to these regions (Boque-Sastre et al., 2015). Whilst R-loops may have historically been seen as primarily a roadblock in DNA there may yet be a wider role of R-loops in their interaction with DNA (Skourti-Stathaki et al., 2014, Girasol et al., 2023). R-loops have been hypothesised to play a role in transcription-mediated Ribonucleic acid (RNA) template repair through promoting repair via recruitment of proteins such as Rad52 and Xeroderma pigmentosum group G (XPG) (Yasuhara et al., 2018). R-loops have been shown to form at both transcription start and termination sites of highly transcribed genes, which may suggest a greater biological relevance of R-loops rather than as mere problematic DNA lesions (Sanz and Chédin, 2019).

G4s are hydrogen-bound guanine-rich structures, which have also been recognized as associates in promoting DNA damage and genomic instability. The key aspect of G4s is the presence of 4 guanine bases connected by “Hoogsteen” bonds in a square or “g-quartet” (Gellert et al., 1962). These square bonded guanines can stack on top of subsequent g-quartets and so on (Spiegel et al., 2020). As with R-loops, G4s are capable of obstructing replication fork progression and promoting RS, leading to genetic instability (Lopes et al., 2011, Papadopoulou et al., 2015). Physiological roles of G4s remain a matter of

debate, however links to transcription have been observed with enrichment of G4 motifs noted at transcription start sites within paused genes (Eddy et al., 2011). As well as potential physiological roles, G4s have also been implicated in cancer, with cancer cell lines often displaying a greater abundance of G4s than healthy cells (Chambers et al., 2015, Biffi et al., 2014). Interestingly, G4s may also play some role in radioresistance, where the G4 structure may act as a shield against free radicals, reducing IR-induced damage compared to the canonical DNA structure sites (Kumari et al., 2019). As with R-loops, DSBs associated with G4s can be challenging to repair, due to their need for extended resection which promotes a reliance on HR over other repair pathways and therefore can lead to DSB repair delay and contribute to instability (Camarillo et al., 2021).

In addition to R-loops and G4s, other important DNA features that are considered at increased risk of DSBs are CFS regions and very long genes. CFSs vary in length and 125 regions have been identified across the human genome by Kumar et al (Kumar et al., 2019) comprising common fragile sites; identified under replication stressors such as aphidicolin and rare fragile sites which have been classed as folate-sensitive sites; induced following exposure to bromodeoxyuridine (Zlotorynski et al., 2003). As mentioned previously, RS is commonly associated with DSB susceptibility and CFSs are regions that have been identified following the induction of RS identified by Fluorescence in-situ hybridisation (FISH) on metaphase chromosomes, exhibiting gaps indicating incomplete DNA synthesis (Boteva et al., 2020). These CFS regions show incomplete compacting of chromatin and delays in DNA replication (Boteva et al., 2020). This requires extension of the G2 phase to allow for resolution of these late replicating or incompletely replicated regions. Additionally, CFSs specific to cell lines can be associated with sites of copy number variation (CNV), particularly in cancer at deleted regions (Zack et al., 2013). CFS sites have also been associated with repetitive sequence regions, also known to contribute to RS (Ozeri-Galai et al., 2011, Durkin et al., 2008). Very long genes have been associated with CFS and identified as contributors to instability and fragility at associated CFS regions (Helmrich et al., 2011). The term ‘very long gene’ has been used to describe genes of greater than 800 kilobase pairs (kbp) in length. These very long genes can take many hours to transcribe, requiring more

than one cell cycle to complete. Additionally, very long genes also tend to be replicated late in S phase. In addition to late replication and prolongation of transcription, Helmrich et al (Helmrich et al., 2011) also identified that CFS-related instability could only be exhibited when these genes were actively being transcribed. This led to replication-transcription conflicts and the subsequent formation of R-loops. Importantly, many neural genes are included within the set of very long genes and have been associated with increased expression and DSBs in NSCs (Wei et al., 2016). Finally, highly repetitive regions have been identified as hotspots for DNA damage, such as telomeres and centromeres. Centromeres have been identified to harbour clusters of SSBs and DSBs even in quiescence, requiring Rad51-mediated HR for repair (Saayman et al., 2023).

### 1.4.2 Replication-transcription conflicts

As has already been discussed, non-canonical DNA structures, CFSs and very long genes have been associated with replication-transcription conflicts, resulting in RS and DSB formation. Regions that require prolonged transcription, highly transcribed regions and late replicating regions are at increased risk of interacting adversely with replication machinery. Replication-transcription conflicts can be described as ‘head on collisions’; where the replication and transcription machinery are moving towards each other or ‘co-directional’; where both collide moving in the same direction. Hence, circumstances resulting in an overlap of transcription and replication can result in these conflicts. The timing of replication origin firing and the licensing of replication origin sites has important implications on the likelihood of potential replication-transcription conflicts also. With regards to replication machinery and replication firing, replication origin sites vary across genic regions. Both transcription start and termination sites have been shown to have enrichment for replication origins but conversely gene bodies are broadly deplete of origin locations (Goehring et al., 2023, Azvolinsky et al., 2009, Cadoret et al., 2008). Transcription start and termination regions have both been described as sites of replication-transcription collisions in cancer cell lines (Koyanagi et al., 2022). Termination locations of replication have also been shown to occur frequently at the termination sites of highly transcribed genes, though it is unclear whether this is caused by pauses occurring at these sites because of replication-transcription conflicts (Chen et al., 2019b). Replication-transcription conflicts can also occur

because of aberrant origin firing secondary to RS. Nucleotide pool depletion has been shown by Chen et al (Chen et al., 2019b, Chen et al., 2015) to lead to firing of previously dormant regions adjacent to transcription termination sites.

### 1.4.3 Gene expression

In addition to conflicts with replication, transcription alone has also been highlighted as a process that may promote DSBs. As described, transcription, particularly of longer genes can result in torsional stress, requiring resolution via topoisomerases (Ju et al., 2006). The process of transcriptional elongation also appears to lead to DSB formation which may not always be as a direct result of topoisomerase intervention and rather the pressure exerted on the DNA (González-Barrera et al., 2002, Bunch et al., 2015).

Additionally, the presence of transcription-associated DSBs has been seen in neural progenitor cells where DSBs not related to RS but secondary to transcription have been identified (Michel et al., 2022). Furthermore, in the absence of functional p53, neural progenitor cells accumulate DSBs at transcription start sites. This accumulation of transcription start site DSBs has been associated with neuropsychiatric disorders including intellectual disability, neuroinflammation and hyperproliferative autistic spectrum disorder (Wang et al., 2020, Michel et al., 2022). At the same time, highly transcribed genes have also been shown to be prioritised during repair, with DSBs at these sites being preferentially repaired over lower transcribed genes (Chaurasia et al., 2012).

Transcription-associated DSBs are preferentially repaired via HR rather than the relatively more error-prone NHEJ pathway (Yasuhara et al., 2018). Preference of the high accuracy of HR compared to NHEJ in active genes may indicate association with non-canonical structures that require extended resection for removal, limiting repair to late S or G2 phase. It has also been observed that there is clustering of DSBs at transcriptionally active genes in G1 phase which may reflect intentionally delayed repair and avoidance of the NHEJ pathway (Aymard et al., 2017). Additionally, DSBs occurring in G1 have been demonstrated to cluster spatially within the nucleus, requiring modulation of chromatin to facilitate mobilisation of DSB sites (Ismail et al., 2010, Ginjala et al., 2011).

### 1.4.4 Chromatin conformation

The conformation of chromatin has also been discussed extensively in the context of DSB formation and repair. As with replication-transcription conflicts, much overlap occurs with other potential influencers of DSB induction including non-canonical structures, CFSs and replication-transcription conflicts leading to RS. Dynamic modulation of chromatin is required for cells to rapidly access DNA that needs transcribed. Whilst chromatin modulation per-se does not necessarily directly cause DSBs, the remodelling around DSBs and other cellular features has important implications for the repair and recovery from DSBs.

Chromatin modulation in DSB repair is largely dictated by protein signals from the DDR in determining the repair pathways of choice. For example, 53BP1 is a DSB repair protein which acts to oppose the 5' end resection preventing HR, thereby promoting NHEJ repair, whereas breast cancer gene 1 (BRCA1) acts to promote 5' end resection preferencing towards HR (Scully et al., 2019, Aleksandrov et al., 2020, Daley and Sung, 2014). Both BRCA1 and 53BP1 signal differing chromatin-modifiers to facilitate NHEJ or HR depending on the pathway of choice. ATM also has important roles in DDR to promote accessibility to DSB damage in heterochromatin via phosphorylation of KRAB-associated protein 1 (KAP-1), resulting in widespread chromatin relaxation and accessibility (Goodarzi et al., 2008). Chromatin modulation also acts to halt transcription at sites of DSBs. In HR repair, chromatin remodelling complex nucleosome remodelling and deacetylase (NuRD), which is recruited early in DSBs, acts to downregulate any active transcription around the DSB site (Chou et al., 2010). Additionally, the spatial distribution of chromatin in the nucleus indicates that actively transcribed genes are more central to the nucleus compared to non-coding sequences in heterochromatin which tends to be more peripheral (Qiu, 2015, Villeponteau, 1997). Heterochromatin has also been shown to abrogate the direct effects of IR via absorption of free radicals (Chiolo et al., 2011, Qiu, 2015). Additionally, nucleosomes have also been associated with a reduction in local DSBs following IR, suggesting a protective influence on local sites of DNA (Brambilla et al., 2020).

The chromatin conformation within GSCs has been studied showing that euchromatin was highly abundant in cells compared to relatively low levels of

heterochromatin (Zhao et al., 2008). Whilst GSCs have been demonstrated to have high levels of euchromatin, it is heterochromatin that appears to act as a protector against IR-induced DNA damage. Interestingly, GBM cells have also been found to utilise the histone deacetylases 4 and 6 (HDAC4, HDAC6) to promote stemness and radioresistance through promoting DSB repair and chromatin condensation (Marampon et al., 2017).

#### 1.4.4.1 Mapping chromatin profiles

Identifying the profiling of the chromatin landscape can be performed using several sequencing methods to give specific information on the chromatin context. One such commonly used method is assay for transposase-accessible chromatin coupled to sequencing (ATAC-seq) (Buenrostro et al., 2013). In brief, ATAC-seq makes use of Tn 5 transposases which are capable of cutting and tagging open regions of chromatin to allow these sites to be sequenced. ATAC-seq needs relatively few cells (~50,000 cells) to be performed and has a good signal to noise ratio compared to other methods but does have the risk of Tn5 insertion bias for mapping. Additionally, whilst ATAC-seq is effective in identifying areas of euchromatin enrichment, it does not specifically identify regions of heterochromatin. For mapping of heterochromatin regions, techniques such as DNase I Hypersensitive Sites Sequencing (DNase-seq) which primarily digests accessible chromatin locations and not highly nucleosome-bound sites, are a more effective means of looking at heterochromatin specifically. Chromatin immunoprecipitation sequencing (ChIP-seq) for heterochromatin marks can also be another means of identifying heterochromatin and euchromatin by mapping euchromatin or heterochromatin-associated histone markers such as H3K27ac; associated with euchromatin or H3K9me3; associated with heterochromatin (Kharchenko et al., 2011, Skene and Henikoff, 2017). Several other sequencing methods for chromatin profiling are also available, however these will not be discussed here. An excellent summary paper of addressing the mapping of chromatin has been written by Chawla et al. (Chawla et al., 2021).

## 1.5 Detecting DSBs

The detection of DSBs has been a key field of study for quantification of the DDR in healthy and cancer cells as well as an important aspect of defining cellular response to IR. Many methods exist to study DSBs in cells, which will be discussed here. Traditionally, DSBs can be visualised using immunofluorescence (IF) techniques via the identification of DSB DDR proteins that can act as a surrogate for the DSB. Total DSB load can also be measured using neutral comet assays. A number of methods for mapping DSB locations have been developed in recent years as a means of giving more precise locations of where DSBs are occurring, including breaks ligation in-situ sequencing (BLISS) and INDUCE-seq which will be discussed.

### 1.5.1 Immunofluorescence markers

IF markers have long been used as a means of quantifying the DDR to DSBs and as surrogates for DSB frequency. Two of the most commonly used markers for quantifying DSBs are γH2AX and 53BP1. These both are classified within the category of ionising radiation-induced foci (IRIF). The most common and earliest IRIF used is γH2AX, which produces identifiable foci within the nucleus following IR (Rogakou et al., 1998). Whilst γH2AX demonstrates a clear DDR following IR, these foci are known to span megabases due to the many histone modifications that occur in response to the local DSB (Rogakou et al., 1999, Friesner et al., 2005). The foci that are rapidly detectable within minutes following IR and γH2AX has been described as identifying DSBs in a 1:1 manner for DSB to foci (Kuo and Yang, 2008). Foci for γH2AX demonstrate a dose-dependent increase in accumulation following IR, making them an attractive target for monitoring DSBs as a whole (Friesner et al., 2005). The IRIF 53BP1 has also been used to measure DSB formation and has been found as another effective marker with good colocalisation with γH2AX (Rothkamm et al., 2015). As previously mentioned, 53BP1 acts as a recruitment protein at DNA DSBs, as well as promoting NHEJ at DSBs and also has actions in amplification of ATM activity (DiTullio et al., 2002). Regarding 53BP1, foci number per cell and foci resolution has been demonstrated to predict radiosensitivity, correlating with the surviving cell fraction following IR exposure (Markova et al., 2015). Like many techniques for detecting DSBs, measurement of IRIFs is limited by the necessity of fixing cells at

a single timepoint and identifying DSBs in a single snapshot. Some methods using 53BP1 have been used in an attempt to overcome this. Unlike γH2AX, 53BP1 is not a phosphorylated protein and so it is possible to use a truncated 53BP1 reporter via plasmid transfection to track foci development and resolution in live cells. Using a window chamber in a Ewing's sarcoma model, Yang et al (Yang et al., 2015) demonstrated the effective use of truncated and fluorescently labelled 53BP1-transfected cells in measuring DSBs following treatment with olaparib via live *in vivo* imaging.

Overall, IRIF can provide a helpful visual interpretation of the DSBs occuring within cells, however there are a number of limitations when considering the measurement of DSBs via IRIF. Whilst γH2AX and 53BP1 have demonstrated correlation of foci with DSBs, it must be remembered that these foci are still DDR proteins rather than direct DSBs. Therefore any process that might abrogate the response of either of these may result in an incomplete picture of the DSBs occurring within the cell. Additionally, both γH2AX and 53BP1 foci are representative of a very large region and so, whilst they allow for clear visual detection through microscopy, conclusions about DSB locations are very limited. Furthermore, 53BP1 has also been indicated to have a wider role in the cell which may not correlate only with DSBs. Whilst 53BP1 generally correlates well with γH2AX, other large 53BP1 foci known as 53BP1 nuclear bodies have been investigated and identified as larger regions representing regions of under-replicated DNA sequestered for repair after replication (Fernandez-Vidal et al., 2017). This does complicate the interpretation of these foci as to whether they remain a measure of DSBs or of the wider DDR.

### 1.5.2 Neutral comet assay and pulsed-field gel electrophoresis

Another means of determining DSB frequency is the use of DNA fractionation to detect broken DNA fragments by electrophoresis. This includes the neutral comet assay and pulsed-field gel electrophoresis (PFGE). The neutral comet assay detects DSBs by taking lysed cells, embedding these in agarose gel and using electrophoresis to cause broken DNA strands to migrate, creating a "comet tail" from the nucleus. The DNA has been fluorescently labelled to allow a visualisation of the intact nuclear DNA and the DSBs making up the tail (Roy et al., 2021, Olive and Banáth, 2006). The quantification of DSBs can then be

achieved by measuring the comet tails of individual nuclei. The tail moment is a measurement which quantifies the amount of DNA present within the tail and tail length (DNA percentage in tail x tail length). The percent of DNA within the tail can be quantified relative to the nucleus (Olive and Banáth, 2006). For PFGE, cells are first embedded in agarose then lysed whilst in agarose. The samples embedded already in agarose known as plugs are then run as a gel via electrophoresis. Here, the intact DNA remains at the top of the gel and DNA with DSBs migrates down the gel depending on DNA length and uses *S. cerevisiae* DNA as a control for fragment length (Pond and Ellis, 2019). Unlike the neutral comet assay, PGFE can be used for DNA transfer for immunoblot (Kawashima et al., 2017). PFGE can provide some means of extracting regions of DNA that have undergone DSBs. Additionally, steps can be taken as a means of quantifying DNA DSBs associated with replication forks (Kawashima et al., 2017).

Whilst both the neutral comet assay and PGFE can give a quantitative result of DSBs, they are limited in their ability to adequately address DSB location. Both are also under the limitation of DNA migration within the gels where migration can be affected by chromatin condensation which may confound results. The neutral comet assay does allow DSB quantification at a single cell level, but DSB sites are non-identifiable. The comet assay also requires large fractions of IR to detect DSBs (Olive and Banáth, 2006). As for PGFE, whilst different lengths of DSBs can be fractionated, DSB lengths may vary significantly and so identifying all relevant DSBs is not feasible, with many of the DSBs being at potentially untranscribed regions (Vítor et al., 2020, van Waarde et al., 1996). DNA fragments from PGFE can be very long from 200 kbp to 2000 kbp, limiting the accuracy of the specific DSB location and therefore location-specific conclusions (Pond and Ellis, 2019).

### 1.5.3 Indirect sequencing methods

Methods to detect DSBs at regions of key proteins or at induced locations have been considered as a means of determining DSB location, both *in vitro* and *in vivo*. Methods that have been considered have used means to localise to proteins and inducing DSBs within live cells; these will be discussed below.

#### 1.5.3.1 ChIP-seq of DSB proteins

ChIP-seq has been studied as a method aiming to detect DSB location using DSB surrogate markers. For this ChIP-seq identifies DNA bound to the protein of interest for isolation and subsequent sequencing. Using DNA bound to γH2AX, mapping of DSB location has successfully given broad views of key locations where DSBs are occurring, including identifying DSBs linked to TopIIß (Madabhushi et al., 2015). However, as already mentioned, phosphorylation of H2AX around DSBs spans across regions at a level of megabases and spreads across these regions (Iacovoni et al., 2010). Therefore, whilst general regions of DSBs can be determined, the specificity where these DSBs are occurring remains relatively low. Additionally, though γH2AX is generally seen as specific to DSBs, there is evidence that γH2AX can also localise to nucleotide excision repair sites at SSBs which again lowers the specificity to DSB locations (Marti et al., 2006, Iacovoni et al., 2010).

#### 1.5.3.2 GUIDE-seq and high-throughput genome-wide translocation mapping

In addition to the developments of DSB mapping *in vitro* which will be discussed below, there have also been methods developed in attempts to detect DSBs *in vivo*. These will not be discussed at length given that this thesis primarily concerns *in vitro* cell DSB mapping. Two such techniques; GUIDE-seq and high-throughput genome-wide translocation mapping (HTGTS) have been developed. GUIDE-seq uses CRISPR RNA-guided nucleases (RGNs) to induce DSBs in living cells which are then tagged by DSB DNA in an end-joining process similar to NHEJ. The DSB DNA tags are then amplified and sent for sequencing to give the locations (Tsai et al., 2015). For HTGTS, this mapping method, has undergone a number of additions to optimise sequencing (linear amplification mediated HTGTS: LAM-HTGTS (Hu et al., 2016) and improved HTGTS: iHTGTS (Yin et al., 2019)) but the basic protocol involves the use of “bait” DSBs from nuclease I-SceI generation. These bait DSBs act to capture “prey” endogenous baits, namely broken ends existing within the genome. These bait-prey junctions are ligated and tagged with an adapter with the type depending on the HTGTS iteration. Tagged bait-prey DSBs then undergo polymerase chain reaction (PCR) followed by sequencing (Chiarle et al., 2011, Hu et al., 2016). Both GUIDE-seq and HTGTS represent indirect mapping methods due to the introduction of guide RNA or DNA

sequences, though are considerably more direct than ChIP-seq. However, both GUIDE-seq and HTGTS rely on transfection or transduction as a step to integrate these sequences meaning that cells with poor transfection rates may not effectively have DSBs picked up. This method is also biased towards detecting DSBs repaired via NHEJ and therefore may not reliably represent DSBs that would trigger HR or other repair pathways (Hu et al., 2016).

## 1.5.4 Direct sequencing methods

In recent years, efforts have been made to develop effective methods to directly sequence DSBs *in vitro*. There are now a number of methods that have been developed to identify DSB location within the genome which will briefly be discussed.

### 1.5.4.1 BLESS, iBLESS and qDSBseq

One of the earliest methods used to directly identify DSB location was break labelling, enrichment on streptavidin, and sequencing (BLESS) which identified DSBs by use of in situ ligation with a biotin-bound hairpin linker. The biotin is then acquired by the streptavidin beads and then another hairpin linker distal to the DSB sequence is attached to the fragment. These marked DSB fragments are then PCR-amplified using primers that are particular to the attached linkers and DNA is then sequenced and purified (Crosetto et al., 2013). Two methods have aimed to improve upon the original protocol; namely immobilised BLESS (iBLESS) and quantitative DSB sequencing (qDSBseq). Regarding iBLESS, cells are initially immobilised in agarose beads as a means of protecting cells from the shearing forces that might be experienced in DNA preparation. The DSBs are then treated as per BLESS, using lysing, biotin-bound hairpin linkers and streptavidin capture (Biernacka et al., 2021). For qDSBseq, in an effort to quantify DSBs, spike-ins are used to quantify the total DSB yield by providing a normalisation value from site-specific endonucleases that are introduced (Zhu et al., 2019). Additionally, qDSBseq provides accompanying software for sequencing analysis to allow for quantification from DSB reads. Whilst the iBLESS and qDSBseq protocols improve upon the original BLESS method, these protocols do remain time and resource intense, requiring roughly 60 hours of laboratory work for processing and upwards of 10 million cells (Yan et al., 2017, Biernacka et al., 2021).

### 1.5.4.2 DSBcapture and END-seq

DSBcapture and END-seq are further DSB mapping techniques that have come following the development of BLESS. DSBcapture requires fixed cells and uses single-end sequencing of DSB fragments rather than the double ended technique used in BLESS. This is with the aim of providing a more diverse library of DSBs. DSBcapture uses A-tailing of blunted cells and incorporation of P5 Illumina adapters with no requirement for spike-in libraries as with qDSBseq (Lensing et al., 2016). Additionally, DSBcapture reports a higher yield of DSBs compared to BLESS and has demonstrated the capability of identifying DSBs at G4-rich regions in human epidermal keratinocyte cells. Similar to DSBcapture, END-seq also makes use of A-tailing and Illumina adapter ligation, however DSBs from live cells are labelled within agarose plugs and lysed *in situ* as a means of minimising artificially induced DSBs that have been thought to be caused by cell fixation (Canela et al., 2016, Biernacka et al., 2018). Like DSBcapture, END-seq reports an improvement in specificity and sensitivity to BLESS and an apparent improvement in DSB accuracy. Again, both DSBcapture and END-seq do require high cell yields of 1 million and 10 million cells respectively (Bouwman and Crosetto, 2018).

### 1.5.4.3 Breaks ligation in-situ sequencing

Breaks ligation in-situ sequencing (BLISS) represents another sequencing method for detecting DSBs in an aim to improve upon the original BLESS concept. BLISS uses T7-adapter sequences to label DSB ends along with a barcode to denote the sample and a unique molecular identifier (UMI) to allow for better quantification of DSBs. By using the T7-adapter sequences, tagged DSBs can be linearly amplified via *in vitro* transcription. The DSB-tagged DNA is then isolated in library preparation and sequencing (Yan et al., 2017). BLISS allows for DSBs to be identified *in situ* in either fixed cells or tissues, allowing for use in clinical specimens. BLISS also reports the use of cell numbers as low as $10^3$ compared to other techniques, though in the experience of this thesis, optimum results were obtained using 2 million cells per condition (Bouwman and Crosetto, 2018, Yan et al., 2017). In suspension BLISS (sBLISS) has also been developed as an aimed improvement on the original BLISS protocol which allowed for better cell harvesting at low cell requirements and does not require the addition of an

agarose plug, again reducing the initial cell content required (Bouwman et al., 2020). BLISS aims to address some of the challenges associated with DSB sequencing and the problems related to PCR via the use of UMIs. Though PCR is not fully excluded as a partial confounder using this method.

#### 1.5.4.4 INDUCE-seq

Finally, INDUCE-seq, published in 2022 (Dobbs et al., 2022), represents a further development in the mapping of DSBs. Many of the above direct sequencing methods require high cell numbers at input and are dependent to some degree on PCR amplification steps. A number of methods such as qDSBseq and BLISS have aimed to address this through spike-ins or UMIs, however this does not entirely remove the limitation of PCR in directly quantifying DSBs (Bouwman and Crosetto, 2018). The INDUCE-seq method reports a 1:1 read to DSB for detecting DSBs and is a PCR independent technique. For INDUCE-seq, DSBs are labelled *in situ* with a full length P5 Illumina adapter, after which the DNA is then fragmented. The fragmented DNA is then tailed with half-functional P7 adapters, which results in DSB-labelled fragments showing a P5 adapter at the DSB region and a half-functional P7 adapter at the fragmented end. Other non DSB-tagged fragments will only have P7 half-functional adapters attached. DNA fragments are then sequenced using an Illumina flow cell, where the P5 adapters are recognised by the flow cell and the half-functional P7 adapters are sequenced to give the DSB location (Dobbs et al., 2022). In addition to this more direct method of detecting DSBs, INDUCE-seq requires a relatively low cell number of 0.1 million cells. Cells do, however need fixation *in situ* and investigations into satisfactory plating mediums are still a matter of exploration.

## 1.6 Characterising DSBs in GSCs

Having discussed a wide array of methods in characterising DSBs and potential influencers on DSB location, this thesis will explore DSBs and their location specifically within GSCs. A number of other cancer cell lines have been investigated with regards to profiling DSBs including cancer cell lines in breast, leukaemia, mesothelioma and sarcoma (Ballinger et al., 2019, Brambilla et al., 2020, Yan et al., 2017). Additionally, normal cells have also been profiled to investigate DSBs including neural cells and NSCs which hold particular interest to

this work given the important links of NSCs with GSC populations (Ballarino et al., 2022). To our knowledge, this study is the first in primary patient-derived cell lines in GBM, with previous studies performed in commercial cell lines or in normal tissue.

As previously described, there is good evidence that GSCs have developed aberrant upregulation of DDR pathways which would appear to contribute to radioresistance (Ahmed et al., 2015, Mandal et al., 2011). Underlying this is the evidence that GSCs have increased levels of RS compared to differentiated progeny which appears to pre-empt DDR by priming key DSB repair proteins ATR as well as Chk1 (Carruthers et al., 2018). The underlying mechanisms for elevated RS remain ill-defined, though a number of potential causes of RS have been described above, many of which overlap with at-risk locations for DSB development such as non-canonical structures, replication-transcription conflicts and torsional stress.

Whilst DSBs are held up as highly deleterious, many cell processes do require physiological DSBs to maintain normal cell function such as chromatin modulation, transcription and processes requiring genetic recombination. Therefore, these processes must be highly monitored and preserved in order to promote continuing healthy cell function. In light of this, understanding to what degree GSCs maintain these functions is key in understanding how endogenous DSBs contribute to cell survival. DSBs are at the heart of the malignant process for GSCs and a main driver of tumourigenesis. The adaptability of GSCs to manage DSBs is key to their survival and to GBM recurrence.

The mapping of DSB location and DSB density in GSCs will better characterise the role of DSBs in treatment resistance and cell survival. Having identified that GSCs shared important characteristics with NSCs including self-renewal properties and elevated RS, comparing DSBs across GSC and neural cell populations is highly appealing. Additionally, understanding DSBs in GSCs with reference to other cancer cells will also frame the findings seen in GSCs. GSCs demonstrate markers of persistent DSBs, even at baseline (Carruthers et al., 2018). Therefore, there may be a subgroup of endogenous DSBs present that do not risk cell death but rather prime the DDR for management of lethal DSBs.

Finally, comparing DSBs in radioresistant GSCs with their radiosensitive differentiated progeny will also further characterise whether there are identifiable differences between radiosensitive and radioresistant populations in DSB locations and density contributing to survival. Another important consideration is whether there are particular DSB sites in GSCs following radiation that are preferentially repaired. This may be due to either pathway preference, given the well described inclination of GSCs towards HR, or related to targeting particular locations for repair as has previously been described in regions of high transcription (Lim et al., 2014, Chaurasia et al., 2012). Investigating DSBs in the context of IR has been notoriously challenging and, to date, the locations of IR-induced breaks remain uncharacterised. Whilst broad regions are detectable via ChIP-seq and PGFE, location-specifics remain elusive.

## 1.7 Research hypotheses and objectives

The precise relationship of DSB location and cell survival in the context of IR is an unexplored landscape. Only limited studies have investigated DSB location mapping in the context of cancer treatment resistance and IR. This thesis will characterise the “breakome”- referring to the mappable DSB landscape, in GSCs in three primary cell lines and their association with at-risk genomic locations and processes. The GSC breakome will be contextualised using matched GSC-derived differentiated progeny as well as commercial cell line data. The GSC breakome will be investigated in terms of the differential radiosensitivity seen between GSCs and differentiated progeny and the differences following IR will be quantified and evaluated for links to differential radiosensitivity.

Hypotheses:

1) Given that GSCs are capable of repairing theoretically lethal DSBs, the sites at which endogenous DSBs occur are relevant to GSC survival and radiosensitivity.

2) The location of these DSBs will be influenced by both physiological and pathological cell processes.

3) Exposure to IR will result in detectable changes in the GSC breakome that are relevant to cell survival.

Objectives:

- Map sites of DSBs in GSCs and differentiated cells using BLISS with reference to differentiated progeny populations and non-GBM cell lines

- Investigate previously described at-risk gene characteristics and genomic locations for DSBs in GSCs

- Compare the DSB landscape in the context of the transcriptome and in differential euchromatin enrichment

- Identify changes in DSB locations across radioresistant GSCs and radiosensitive differentiated cells

- Quantify differences between GSCs and differentiated cells before and after radiation

# Chapter 2 Materials and methods

## 2.1 Cell culture

Glioma cell lines were primary cell lines originating from resected tumours grown *in vitro*. All original cell lines had previously met with approval under the local regional ethics committee as agreed by compliance under the UK human tissue Act. Cell lines E2, G7 and R10 were originally gifted to the Chalmers lab by the Watts laboratory (Fael Al-Mayhani et al., 2009). These were grown in culture medium in T75 flasks. GSCs were cultured as monolayer on Matrigel™-coated flasks using advanced DMEM F12 (Gibco™) media with supplements: 1% L-glutamine (200mM Gibco™), 1% B27 supplement (50x Gibco™), 0.5% N2 supplement (100x Gibco™), 20ng/ml epidermal growth factor (Sigma) and 20ng/ml fibroblast growth factor 2 (Sigma). Differentiated cells were grown in MEM (Gibco™) with supplements: 10% foetal calf serum (Sigma) 1% L-glutamine (200mM Gibco™), and 1% sodium pyruvate (100mM Gibco™). Cell culture was performed in a category II tissue culture room using class II laminar flow hoods and sterile handling of specimens to avoid culture contamination.

Cells were kept in humidity at 37°C at 21% oxygen and 5% carbon dioxide. Passages after freezing were kept as low as possible with a passage of less than 8 for experiments used. For splitting and passaging cells, cells were washed in 5 ml of autoclaved phosphate buffered solution which was aspirated and then treated with 0.7 ml room temperature (RT) StemPro Accutase™ Cell Dissociation Reagent (Gibco™) and allowed to completely detach for up to 5 minutes at 37ºC. Cells were passaged when 70-80% confluency had been reached. Cells were counted using a haemocytometer by adding 10µl of dissociated cells for counting and then the mean of two chambers was counted, taking into account the 3 large squares, giving the total number of cells per ml as n $x10^4$. Alternatively, cells were counted using the CellDrop™ Automated Cell Counter with 10µl volumes using brightfield setting with settings for counting optimised by Dr Joanna Birch.

For cell freezing, cells were detached and brought into a single cell suspension from monolayer growth and centrifuged at 1800 rpm for 5 minutes. Cells were frozen in approximately $0.5x10^6$ cell batches in 1 ml cryovials. Cells were frozen

in 10% DMSO, 90% cell media solution. Cells were initially frozen in -80°C using either cotton wool wrapping or Mr. Frosty™ Freezing Container or a Corning® CoolCell™ 5 mL LX Cell Freezing Container. Cells were kept at -80°C for up to 4 months for use or transferred to liquid nitrogen after initial freezing had completed for long term storage. For thawing cells from liquid nitrogen, cryovials were thawed rapidly at 37°C and diluted in media up to 10 ml. Cells were spun as described and media was aspirated from the cell pellet. The pellet was resuspended in fresh media and cells were plated on T75 flasks. After cells had adhered, media was changed to fresh media.

## 2.2 Cell irradiation

Irradiation of cells was performed using the XStrahl RS225 irradiator with x-rays generated at 195 kV using a current of 15 mA. The dose rate was determined by the distance from the source and time with tables provided for determining the required time required dose delivery. For example, at a distance of 400 mm from the source, the dose rate delivery was 1.5 Gy per minute. Cells that were treated for 0Gy IR were taken out at the same time as the treated cells for the same length of time as the cells were irradiated.

## 2.3 Breaks ligation in situ sequencing

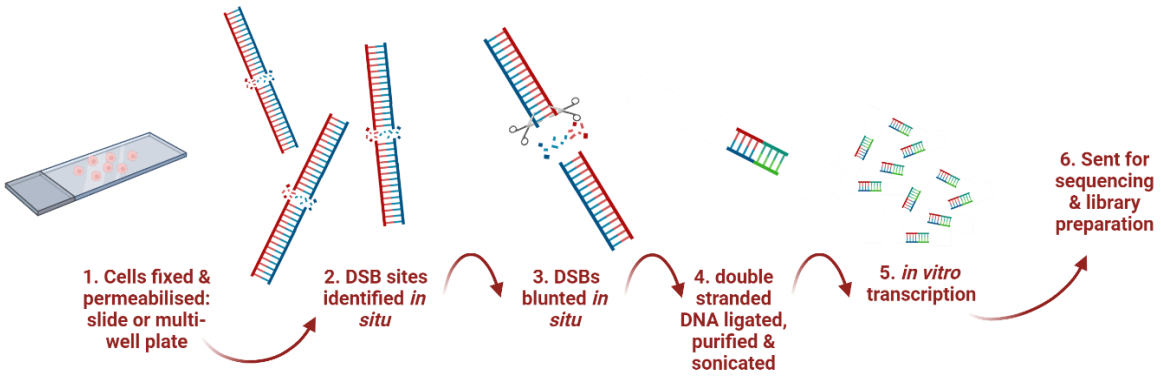
BLISS sample and library preparation was performed by Karen Strathdee for experiments E2, G7, R10 GSC verses differentiated and R10 24 hours irradiated cells (

**Figure 2.1 BLISS schematic**

(a) BLISS schematic outlining steps for DSB sequencing. 1 & 2: Cells are fixed and permeabilised for identification of in situ breaks. 3 & 4: DSB ends are blunted to allow for dsDNA adapters to ligate to DSBs. 5 & 6: The DNA is extracted using the adapters and sequenced immediately downstream of the tagged DSBs. It is then linearly amplified by T7-mediated *in vitro* transcription and then Illumina library prepped and sequenced. Image generated using Biorender.com. (b) DSBs are attached with the following DNA in order: barcode for identifying sample, UMI, RA5 adapter and T7 promoter for *in vitro* transcription. Image adapted from Yan et al (Yan et al., 2017). (c) BLISS DSB adapter visual aid for DSB blunting. DSBs are blunted in situ. DSB double stranded adapters are ligated onto DSBs. DSBs are attached with the following DNA in order: barcode for identifying sample, UMI, RA5 adapter and T7 promoter for *in vitro* transcription. Image adapted from Yan et al (Yan et al., 2017). (d) Alignment and processing of fastq files for downstream analysis. Fastq files generate, quality control of fastq files is performed then UMI and barcodes detected to extract reads, reads are trimmed and aligned to reference genome, reads are filtered for mapping quality , repetitive regions removed, UMIs appended to reads that have successfully mapped, UMIs are aggregated to account for PCR duplication, .BED files are generated for downstream analysis. Image adapted from Yan et al (Yan et al., 2017).

). Experiments were performed in triplicate, however E2 repeat 3 was deemed unreliable and so was excluded from analysis. The protocol used for BLISS was optimised in house by Karen Strathdee from the original paper by Yan et al (Yan et al., 2017). The irradiated E2S samples and library preparations were performed by Sarah Derby.

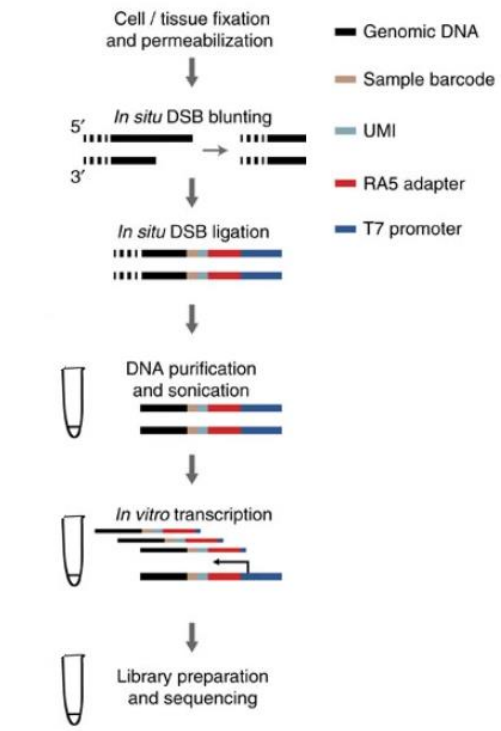
(a)


1. Cells fixed & permeabilised: slide or multi-well plate
2. DSB sites identified *in situ*
3. DSBs blunted *in situ*
4. double stranded DNA ligated, purified & sonicated
5. *in vitro* transcription
6. Sent for sequencing & library preparation

(b)


Ligation site
T7 RNA polymerase promoter
Illumina RA5 adapter
UMI
Sample barcode
5' – CGATTGAGGCCGGTAATACGACTCACTATAGGGGTTCAGAGTTCTACAGTCCGACGATC NNNNNNNN GTCGTCGC –
3' – AAGCTAACTCCGGCCATTATGCTGAGTGATATCCCCAAGTCTCAAGATGTCAGGCTGCTAG NNNNNNNN CAGCAGCG –

(c)


Cell / tissue fixation and permeabilization
*In situ* DSB blunting
5′
3′
*In situ* DSB ligation
DNA purification and sonication
*In vitro* transcription
Library preparation and sequencing
Genomic DNA
Sample barcode
UMI
RA5 adapter
T7 promoter

(d)

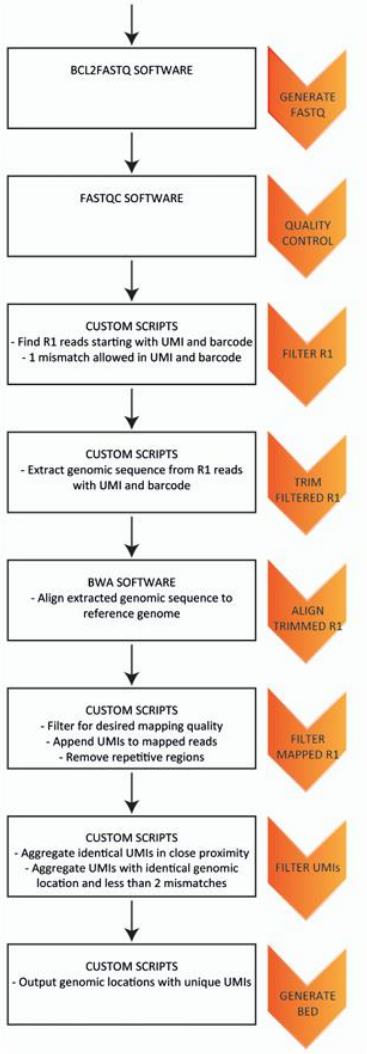

BCL2FASTQ SOFTWARE
GENERATE FASTQ
FASTQC SOFTWARE
QUALITY CONTROL
CUSTOM SCRIPTS
- Find R1 reads starting with UMI and barcode
- 1 mismatch allowed in UMI and barcode
FILTER R1
CUSTOM SCRIPTS
- Extract genomic sequence from R1 reads with UMI and barcode
TRIM FILTERED R1
BWA SOFTWARE
- Align extracted genomic sequence to reference genome
ALIGN TRIMMED R1
CUSTOM SCRIPTS
- Filter for desired mapping quality
- Append UMIs to mapped reads
- Remove repetitive regions
FILTER MAPPED R1
CUSTOM SCRIPTS
- Aggregate identical UMIs in close proximity
- Aggregate UMIs with identical genomic location and less than 2 mismatches
FILTER UMIs
CUSTOM SCRIPTS
- Output genomic locations with unique UMIs
GENERATE BED

**Figure 2.1 BLISS schematic**

(a) BLISS schematic outlining steps for DSB sequencing. 1 & 2: Cells are fixed and permeabilised for identification of in situ breaks. 3 & 4: DSB ends are blunted to allow for dsDNA adapters to ligate to DSBs. 5 & 6: The DNA is extracted using the adapters and sequenced immediately downstream of the tagged DSBs. It is then linearly amplified by T7-mediated *in vitro* transcription and then Illumina library prepped and sequenced. Image generated using Biorender.com. (b) DSBs are attached with the following DNA in order: barcode for identifying sample, UMI, RA5 adapter and T7 promoter for *in vitro* transcription. Image adapted from Yan et al (Yan et al., 2017). (c) BLISS DSB adapter visual aid for DSB blunting. DSBs are blunted in situ. DSB double stranded adapters are ligated onto DSBs. DSBs are attached with the following DNA in order: barcode for identifying sample, UMI, RA5 adapter and T7 promoter for *in vitro* transcription. Image adapted from Yan et al (Yan et al., 2017). (d) Alignment and processing of fastq files for downstream analysis. Fastq files generate, quality control of fastq files is performed then UMI and barcodes detected to extract reads, reads are trimmed and aligned to reference genome, reads are filtered for mapping quality , repetitive regions removed, UMIs appended to reads that have successfully mapped, UMIs are aggregated to account for PCR duplication, .BED files are generated for downstream analysis. Image adapted from Yan et al (Yan et al., 2017).

### 2.3.1 Cell harvesting and crosslinking

For each condition, two T75 flasks of $1x10^6$ were plated for a total of $2x10^6$.cells per condition. Cells were taken at baseline or following IR of 10Gy at the timepoint of interest (E2 GSC samples at 6 hours post-IR). Cells were trypsinised to a single cell suspension of 5 ml and counted as described. Cells were then spun down a 2000 rpm in 15 ml falcons at RT for 5 minutes. Cells were resuspended at a concentration of $1x10^6$ cells per ml in pre warmed media. Cells were fixed by adding 16% PFA (paraformaldehyde, methanol-free) to the suspension to give a final concentration of 4% PFA. The PFA-cell suspension was incubated for 10 minutes exactly at RT whilst gently rotating on a roller shaker. Glycine 2 M was added to suspension to give a concentration of 125 mM and incubated on a roller shaker for 5 minutes. The suspension was then incubated for a further 5 minutes on ice. The suspension was transferred to 2 ml eppendorfs and was centrifuged at 400 g for 10 minutes at 4˚C. The pellet was then resuspended at a final concentration of $1x10^6$ cells per ml in cold PBS. Cell suspension was then stored at 4˚C for up to 2 weeks prior to commencing template preparation.

## 2.3.2 Template preparation from cross-linked cell suspension

### 2.3.2.1 Cell lysis

A total of $2x10^6$ cells in 2 ml were taken and transferred to 2 ml Protein LoBind tubes (eppendorf). These were spun at 400 g at 4°C for 5 minutes, ensuring a pellet had formed at each spin step. The supernatant was then poured off and the pellet resuspended in 400 µl of 1st lysis buffer by gently flicking the side of the tube (Table 2.1). Cells were incubated on ice for 60 minutes after which the tubes were spun at 400 g RT for 5 minutes, the supernatant poured off and the pellets resuspended in 400 µl of 2nd lysis buffer (37ºC). Tubes were then incubated at 37°C for 60 minutes, gently shaking at 400 rpm. The tubes were then spun at 400 g RT for 5 minutes and washed with the following solution: CutSmart Buffer (NEB #B7204, 10x) diluted in solution to 1x with deionised distilled water (DDW) and 0.1% Triton X-100 (Sigma): 100 µl CutSmart Buffer, 900 µl DDW, 1 µl Triton X-100. Tubes were washed twice, spinning between washes. Washes were discarded after each spin.

### 2.3.2.2 DSB *in situ* blunting, labelling and linker ligation

For blunting, 200 µl of blunting mix was added per sample and incubated for 60 minutes at RT, shaking at 400 rpm. Samples were then spun for 5 minutes at 400 g at RT and then washed twice with 400 µl of CutSmart buffer as before and spun for a further 5 minutes at 400 g at RT. The freshly prepared BLISS linker ligation mix (Table 2.1) was added to the samples and pellets resuspended. Individual BLISS linkers were added to corresponding conditions (1A = GSCs, 2A = differentiated cells, 3A = GSCs irradiated, 4A = differentiated irradiated cells) The samples were then incubated overnight for 18-24 hours at 16°C.

The samples were then pelleted, spinning at 600 g for 10 minutes at RT. The pellets were washed twice as before with 400 µl of CutSmart buffer and spun between washes. The supernatant was poured off and the pellet resuspended in 200 µl of TAIL buffer (Table 2.1). To each sample 20 µl of proteinase K was added (800 U/ml, NEB #P8197S) and the samples were incubated overnight at 55°C, shaking at 800 rpm.

### 2.3.2.3 DNA purification, sonication and shearing

Providing the samples appeared clear, the proteinase K was heat-inactivated for 10 minutes at 95°C. The samples were then cooled to RT. Then, working in the fume hood, 220 µl of RT phenol/chloroform/isoamyl-alcohol (PCI) was added to each sample, the tube closed tightly and vigorously shaken for 30 seconds. The closed tubes were then spun for 15 minutes at 20,000 g at RT. The upper phase (approximately 220 µl) of each sample was then collected in the fume hood and transferred to DNA LoBind tubes, with the remaining PCI safely disposed of. A 220 µl volume of chloroform was added to the new tube, tightly closing the tube and vigorously shaking for a further 30 seconds. The samples were then spun again for 15 minutes at 20,000 g at RT. In the hood, the upper phase was again collected, and 3M sodium acetate was added at 1/10$^{th}$ of the collected volume. Glycogen (20mg/ml) was added to give a final concentration of 0.5 µg/ml (3.7 µl per 100 µl). Finally, 100% cold ethanol was added at a volume of 2.45x the sample volume and inverted 10 times. The samples were then placed at -80°C overnight.

The samples were taken out of freezing and then immediately spun for 90 minutes at 30,000 g at 4°C. The supernatant was then removed and pellets washed with 200 µl of 70% cold ethanol and spun again for 15 minutes at 30,000 g at 4°C, repeating this wash and then spun once. The pellets were then dried and dissolved in 100 µl Tris-EDTA (TE) and shaken at 1,100 rpm for 15 minutes at 50°C. The samples were then placed on ice.

The samples were sonicated using a BioRupter at 4°C at a cycle of 30 seconds on, 90 seconds off for 6 cycles with a pause at 3 cycles to mix the samples up and down.

The DNA was then purified using DNA SPRIselect beads (Beckman Coulter #B23317) by adding a 1:1 volume of RT beads to each sample and thoroughly mixed by pipetting up and down 10 times. The samples were incubated at RT for 15 minutes and then placed on a magnetic stand until the samples cleared. The supernatant was then manually aspirated and the beads washed twice with 200 µl of fresh 80% ethanol for 30 seconds. The beads were dried at RT for 5 minutes and the samples resuspended in 20 µl of TE buffer, removing from the magnetic

stand. The samples were incubated at RT for 15 minutes and then placed onto the magnet until the samples cleared. The 20 µl of TE-DNA supernatant was then transferred into at fresh 1.5 ml LoBind tube and 3 µl was sent for Qubit measurement of concentration and BioAnalyser run to assess fragment profiles.

## 2.3.3 Library preparation

### 2.3.3.1 In vitro transcription, DNA degradation and RNA clean up

After profiles were confirmed as adequate (profile peaks between 350-600 bp) the samples were taken for library preparation. Comparator samples were sequenced together for library preparation (i.e. GSC and differentiated replicate 1 were pooled together). For each, 200 ng of each condition was taken and made up to a volume of 7.5µl with nuclease free water. These matched DNA concentrations were then used to make up the in vitro transcription mix (Table 2.1) in 0.5 ml DNA LoBind tubes. The samples were then incubated in a thermocycler for 14-16 hours at 37°C with the lid set to 100°C.

After this, 1 µl of DNAse I (2 U/µl RNase-free #AM2222) was added to each 20 µl in vitro transcription (IVT) reaction and incubated for a further 15 minutes in the thermocycler at 37°C. The IVT reactions were then transferred to 1.5 ml DNA LoBind tubes and brought up to a volume of 50 µl with nuclease free water. Then, 50 µl of RT RNAClean XP beads (Beckman Coulter #A63987) were added to the sample and pipetted up and down 10 times and incubated at RT for 10 minutes. The samples were then placed on a magnetic stand until the solution cleared. The supernatant was discarded, and the samples were washed twice with 200 µl of fresh 80% ethanol for 30 seconds. The beads were then air-dried for 5-10 minutes, avoiding cracking of the pellet. The beads were removed from the magnet and resuspended in 7 µl of nuclease-free water, pipetting up and down 10 times. The samples were incubated at RT for 10-15 minutes and returned to the magnetic stand for 5 minutes until the solutions cleared. Then 5 µl of the sample could be transferred to a new 0.5 ml DNA LoBind tube on ice for adapter ligation.

#### 2.3.3.2 Adapter ligation

The 5 µl purified RNA was placed on ice and 1 µl of RA3 Illumina adapter was added to each sample by pipetting up and down. The samples were then placed in a preheated thermocycler, incubating for 2 minutes at 70°C with the lid at 70°C. The samples were then immediately placed on ice. Then, 4 µl of the adapter ligation master mix (Table 2.1) was added to the sample for a total of 10 µl and incubated on the thermocycler at 25°C for 2 hours with an unheated lid, after which the samples were transferred onto ice.

#### 2.3.3.3 Reverse transcription

To each sample, 2 µl of reverse transcription primer (RTP, 10 uM) was added and samples were incubated on a thermocycler for 2 minutes at 70°C. The samples were then immediately placed on ice for 1 minute and then 13 µl of the RTP master mix was added to each sample whilst on ice. The samples were then incubated in a thermocycler with the lid set at 50°C for the following cycle: 50°C for 50 minutes, 80°C for 10 minutes, 4°C hold.

#### 2.3.3.4 Library indexing and amplification

For the PCR amplification, samples were transferred into 200 µl PCR tubes and on ice, 5 µl of the required Illumina primer was added (i.e. for repeat 1 = RPI-1, repeat 2 = RPI-2, repeat 3 = RPI-3). The indexing master mix was then added to the sample, ensuring no bubbles. The samples were then placed in a thermocycler for the following PCR cycling: step 1- 98°C for 30 seconds, step 2- 98°C for 10 seconds, step 3- 60°C for 30 seconds, step 4- 65°C for 45 seconds, 8 cycles, step 5 - 65°C for 10 minutes, step 6 - 12°C hold.

After cycling, the samples were transferred to 1.5 ml DNA LoBind tubes for DNA SPRIselect bead purification. A volume of 0.8x DNA beads was added to the samples and pipetted up and down 10 times. This was incubated at RT for 15 minutes, resuspending halfway. The samples were then place on a magnetic stand until clear and the supernatant aspirated. The beads were washed twice with 200 µl of 80% ethanol for 30 seconds and air-dried at RT. The samples were then removed from the magnetic stand and resuspended in 22 µl nuclease free water and incubated for 15 minutes at RT. The samples were then placed back

on the magnet for 5 minutes until clear. The 20 µl of elute was then transferred into a PCR tube and the above PCR was repeated as described, adding the 2nd indexing mix (Table 2.1).

#### 2.3.3.5 Library purification

The samples were then transferred into 1.5 ml DNA LoBind tubes and diluted to a volume of 100 µl in preparation for the double sided clean up. A further 50 µl of RT DNA SPRIselect beads were added to samples and pipetted up and down 10 times and incubated for 15 minutes at RT. The samples were placed on a magnetic stand until clear. The 150 µl supernatant was then transferred to a new DNA LoBind tube, leaving the very large fragments behind. To this new tube a volume of 25 µl of beads was added and pipetted up and down 10 times and incubated for a further 15 minutes and then placed on a magnetic stand until the samples cleared. The supernatant was then manually aspirated and discarded, removing the smaller fragments, keeping the beads on the magnet. The beads were then washed twice with 300 µl of 80% ethanol for 30 seconds.

The beads were then allowed to air dry for 5 minutes at RT on the magnet. The samples were then removed from the magnet, resuspended in 15 µl of nuclease free water and incubated at RT for 15 minutes. The samples were then placed on a magnetic stand until the solution cleared and 14 µl of the supernatant was transferred to a new 1.5 ml DNA LoBind tube for each sample. The final 1 µl of dsDNA was sent for Qubit and BioAnalyser assessment to determine if library preparation had been successful.

**Table 2.1: BLISS mixes for methods**

| | |
|---|---|
| **1st lysis buffer** | |
| 10 mM Tris-HCl | 1 ml of 1 M Tris HCl pH 8 |
| 10 mM NaCl | 200 µl of 5 M NaCl |
| 1 mM Ethylenediaminetetraacetic acid (EDTA)<br>Sigma-100g #EDS | 200 µl of 0.5 M EDTA |
| 0.2 % Triton™ X100<br>Sigma | 200µl of 100% Triton™ X100 |
| pH 8, store +4˚C | make up to 100 ml DDW |
| **2nd lysis buffer** | |
| 10 mM Tris-HCl | 1 ml of 1 M Tris HCl pH 8 |
| 150 mM NaCl | 3 ml of 5 M NaCl |
| 1mM EDTA | 200 µl of 0.5 M EDTA |
| 0.3% sodium dodecyl sulphate (SDS) | 3 ml of 10% SDS |
| pH 8, store 25˚C | make up to 100 ml DDW |
| **Blunting mix 1** | |
| Nuclease-free water<br>Ambion #AM9932 | 75 µl |
| Blunting buffer 10x<br>Quick Blunting Kit Buffer NEB #B1201S | 10 µl |
| BSA 10 mg/ml diluted from 50 mg/ml<br>BSA 50 mg/ml Thermo #AM2616 | 1 µl |
| dNTPs 1 mM<br>dNTP mic NEB #N1201AA | 10 µl |
| Quick Blunting enzyme mix<br>Quick Blunt Enzyme mix NEB #E1201 | 4 µl |
| **BLISS linker ligation** | |
| Nuclease-free water<br>Ambion #AM9932 | 66 µl |
| T4 ligase buffer 10x<br>Thermo #EL0011 | 10 µl |
| ATP 10 mM<br>Thermo #R0441* | 12 µl |
| BSA 50 mg/ml<br>Thermo #AM2616 | 3 µl |
| BLISS linker at 10 uM unique per condition (1A, 2A, 3A, 4A)<br>Obtained from Crosetto lab, prepared by Karen Strathdee | 4 µl |
| T4 ligase highly conc. 5 U/µl<br>Thermo #EL0011 | 5 µl |
| **TAIL buffer** | |
| 10 mM Tris | 100 µl of 1 M Tris HCl pH 7.5 |
| 100 mM NaCl | 200 µl of 5 M NaCl |
| 10 mM EDTA | 200 µl of 0.5 M EDTA |
| 0.5% SDS | 0.5 ml of 10% SDS |
| pH7.5 | make up to 10 ml DDW |
| **In vitro transcription mix** | |
| Sonicated DNA mix | 7.5 µl |
| rNTP master mix (2 µl of each)<br>MEGAscript™ T7 high yield transcription kit Invitrogen #AM1334 | 8 µl |
| 10x MEGAscript Reaction buffer<br>MEGAscript™ T7 high yield transcription kit Invitrogen #AM1334 | 2 µl |
| T7 Enzyme Mix MEGAscript<br>MEGAscript™ T7 high yield transcription kit Invitrogen #AM1334 | 2 µl |
| RiboSafe RNAse Inhibitor 40 U/µl<br>Bioline #BIO-65027 | 0.5 µl |

| Adapter ligation master mix | |
|---|---|
| 10x RNA Ligase buffer<br>NEB #B0216L | 1 µl |
| RNaseOUT (40 U/µl)<br>Invitrogen #100000840 | 1 µl |
| T4 RNA Ligase truncated<br>NEB #M0242L | 1 µl |
| DDW | 1 µl |
| **RTP master mix** | |
| 5 x SSIV buffer<br>Invitrogen #LT-02241 | 5 µl |
| 25 mM dNTPs (fresh dilution)<br>Invitrogen dNTP mix 25 mM each #R1121 | 1 µl |
| 100 mM DTT (Invitrogen) | 2 µl |
| Nuclease-free Water<br>Ambion #AM9932 | 1 µl |
| RNaseOUT (40 U/µl)<br>Invitrogen #100000840 | 2 µl |
| SuperScript IV 200 U/µl<br>Invitrogen #LT-02241 | 2 µl |
| **Indexing master mix** | |
| Nuclease-free water<br>Ambion #AM9932 | 15 µl |
| NEBNext Ultra II Q5 mix (NEB, Q5) | 50 µl |
| RP1 primer (common primer for all libraries RP-one) | 5 µl |
| **2nd indexing mix** | |
| PCR product from round 1 | 20 µl |
| NEBNext® Ultra II Q5® mix<br>NEB Q5 #M0544S | 25 µl |
| RP1 primer (common primer for all libraries)<br>Illumina #15013198 | 2.5 µl |
| RPI-1/ RPI-2/ RPI-3 (the same as in PCR 1)<br>Illumina #15013198 | 2.5 µl |

A table of mixes used in BLISS assay. Components and volumes stated above. Reference numbers and companies displayed on right under component name.

## 2.4 BLISS sequencing and analysis

The BLISS sequencing was performed by Glasgow University Polyomics department to a depth of 60 million reads per sample. Fastq quality was assessed by polyomics and fastQC files were also received along with fastq files. Samples E2, G7, R10 and R10 IR 10 Gy at 24 hours were single-end sequenced. Sample E2S IR 10 Gy at 6 hours was paired-end sequenced. Fastq formatting, alignment and preprocessing for samples E2, G7 and R10 GSC vs differentiated cells was performed by Tracy Ballinger to give bam, bed, bedgraph and bigWig files for analysis. Samples were initially aligned to the Genome Reference Consortium Human Build 19 (hg19: GRCh37.p13 February 2009 release) but later realigned to Genome Reference Consortium Human Build 38 (hg38: GRCh38.p14

Ensembl release 110) using “LiftOver” tool in RStudio (Maintainer, 2023). Irradiated experiments in E2 and R10 were aligned by Sarah Derby using Tracy Ballinger’s pipeline (see supplementary files) in the Edinburgh high performance computing cluster “Eddie”. Steps of the BLISS alignment pipeline are detailed below.

### 2.4.1 BLISS fastq processing

For pipeline alignment two files were used for defining the necessary files and outputs. The .sh file “myglobals.sh” defined key variables: export directory, reference genome, fastq access directory, temporary directory, genomic annotation files and location of “scan_for_matches” software. The .sh file “run_bliss_pipeline.sh” defined names of the input and output files: FQNAME= name of the input fastq file, LABEL= defined label for cell line and repeat number, DIRNAME= defined label for the directory name for cell line, PATTERNS = the name of the file with the patterns required for matching to BLISS linkers (see previous section 1.2). Further variables designated in “run_bliss_pipeline.sh” were: sdir = source directory, outdir = output directory and tmpdir= temporary directory.

The “run_bliss_pipeline.sh” runfile was used to call up the required BLISS steps for formatting and alignment. This pipeline followed the following steps:

1. Find data files: find corresponding fastq datafiles and determine if the files are paired, printing the files found to the terminal.

2. “filter_fastq.sh”: Reformatting fastq files and filtering fastqs to leave only reads with the appropriate BLISS linker attached.

3. “align_reads.sh”: alignment of the reads to the reference file specified in “myglobals.sh”.

4. “filter_duplicates.sh”: filter out duplicates and get unique molecular identifiers (UMIs).

5. “get_bliss_summary_stats.sh”: get summary for data after filtering including the number of unique reads.

6. “filter_blacklist_cap.sh”: filter out blacklisted regions.

7. “bedgraphs_to_bigwigs.sh”: make bigwigs from bedgraphs for visualisation.

The files for all of the “.sh” files described above are available in supplementary files. These files were coded by Tracy Ballinger and adapted for the subsequent alignment runs by Sarah Derby.

### 2.4.2 BLISS downstream analysis

Downstream analysis for BLISS was performed primarily in RStudio 3.9-4.3. RStudio scripts for each chapter are also available as supplementary files. For each chapter there is a methods section outlining the analysis performed as per RStudio. The RStudio packages, functions and customised theme used are also available in the appendix. BLISS bedgraph files were used to plot DSB pattern at a whole cell, chromosomal and gene level using packages “circlize” (Gu et al., 2014), “karyoploteR” (Gel and Serra, 2017) and “annotatR” (Cavalcante and Sartor, 2017). To look at DSBs within genic regions, genome data was downloaded from Ensembl including chromosome, gene start and end sites, ensemble gene names, Ensembl annotation, gene description and GC content. These gene locations were intersected with DSB locations and the number of DSBs per gene was calculated using a loop. Given that gene length could be an important factor in the number of breaks harboured per gene, DSB frequency was normalised to DSBs per 1000 base pairs per gene (DSBs/kbp):

*DSBs/kbp = (DSBs in region of interest, i.e gene/length of region of interest, i.e. gene)/1000*

This was based on the paper “BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks” by Yan et al (Yan et al., 2017) which previously described gene break frequency in terms of break frequency per kbp. In order to minimise over-representation of short genes and regions, genes and regions less than 200 bp were not included in analysis. To investigate other regions of interest and non-genic regions, “binned” regions of the whole genome were employed. These were taken at sizes 50 kbp and 100

kbp. Here the whole genome was divided into 50 kbp or 100 kbp regions and the DSB frequency per region was calculated using bedtools intersect in commandline with BLISS bedfiles.

GSC lines had blacklisted regions excluded as part of preprocessing, as did the public datasets (Eichler et al., 2004, Amemiya et al., 2019). Centromeres are identifiable on karyoplots by red bands. The dip or absence of DSBs at centromeres reflected exclusion of centromeres from mapping as part of blacklists. Similarly, many of the short arms of acrocentric chromosomes demonstrate little to no mapping due to the high frequency of blacklisted regions in these areas (Eichler et al., 2004). Whilst centromeres and highly repetitive sites could be of great interest for identification of DSB frequency given their roles in cell division and chromatin, their exclusion was necessary due to the highly repetitive nature of these sites. Therefore, the absence of DSBs at these regions represents regional exclusion rather than an absence of DSBs due to limitations with our DSB mapping assay.

#### 2.4.2.1 Differential analysis

The differential analysis of DSB counts was performed using "DESeq2". DESeq2 used count normalisation via a median of ratios method. Scaling of normalised counts was performed using RStudio scale function to produce a z-score for analysis:

*Scaled z-score = (actual expression -mean expression)/standard deviation*

Gene DSBs and DSBs per 100 kbp regions were analysed as part of the differential analysis. Genes and 100 kbp sites with less than 5 DSBs in less than 3 samples were excluded from the differential analysis. Significance for differential analysis was defined as a log2 fold change of greater than 1 or less than -1 and an adjusted p-value of less than 0.05 following Benjamini-Hochberg procedure (BHP) correction for false discovery rate (FDR).

#### 2.4.2.2 Downstream statistical analysis

Downstream statistical analysis of BLISS data is specified in individual chapter methods. For correlation analysis of non-parametric data, Spearman rank

correlation was performed to generate rho values and p-values for non-parametric data. To allow comparison between repeats and identify differences between differentiated cells and GSCs and 0 Gy and 10 Gy samples, samples were also expressed as a fold change of the control (i.e. differentiated cells or 0 Gy samples).

*Fold change = sample count/control sample count*

For fold change differences, the mean of medians was used for statistical analysis. In analysis across two groups an unpaired t-test was performed with BHP correction. In analysis across groups of three or more one way analysis of variance (ANOVA) was performed. Post-hoc analysis was performed using a Tukey test with BHP correction. For both, a p-value of $<0.05$ was considered significant.

#### 2.4.2.3 Correlation with other datasets

Neural cell BLISS .bed sequencing files were freely available and accessed from https://doi.org/10.6084/m9.figshare.18530531.v2 from the published paper by Ballarino et al: "An atlas of endogenous DNA double-strand breaks arising during human neural cell fate determination" (Ballarino et al., 2022). Cancer cell line BLISS bedfiles MCF7 and K562 were obtained from Tracy Ballinger (Ballinger et al., 2019). Both neural cells and commercial cancer cell lines were originally aligned in hg19 and so bedfiles were realigned to hg38 using the LiftOver tool on RStudio as described above for BLISS datasets.

#### 2.4.2.4 RNA-seq

RNA-seq data was generated for E2, G7 and R10 GSCs and differentiated cells by Dr Emily Clough. Read alignment and processing was performed using Galaxy and details of alignment are available in Dr Emily Clough's thesis "Investigating mechanisms and indicators of sensitivity to replication stress-targeting therapies in glioblastoma". RNA-seq data was originally aligned to hg19 and therefore RNA-seq gene names only were used for RNA-seq-BLISS DSB correlation. Transcripts per million (TPM) was calculated from read counts using RStudio and the following formula:

*tpm3 <- function(counts,len) { x <- counts/len return(t(t(x)*1e6/colSums(x)))}*

## 2.5 Assay for Transposase Accessible Chromatin

ATAC-seq data libraries were generated by Karen Strathdee. Sequencing was performed by Glasgow University Polyomics department to a depth of 60 million reads per sample. Fastq quality was assessed by polyomics and fastQC files were also received along with fastq files. All samples were single-end sequenced. Libraries were generated for E2, G7 and R10 with the following conditions: GSC, differentiated, GSC-irradiated 10 Gy at 24 hours, differentiated irradiated 10 Gy at 24 hours. These were performed in triplicate.

### 2.5.1 ATAC-seq summary

A schematic of the ATAC-seq method is available below in Figure 2.2. In brief, ATAC-seq uses Tn5 transposase to identify accessible regions of open euchromatin. Tn5 transposase probes and cleaves the accessible regions between nucleosomes and tags cleaved DNA with dsDNA adapters. DNA marked with dsDNA adapters have barcodes attached to the adapter sites. This, now barcode-marked DNA, is now PCR-amplified. The PCR-amplified DNA is then sent for next generation sequencing (Buenrostro et al., 2013).

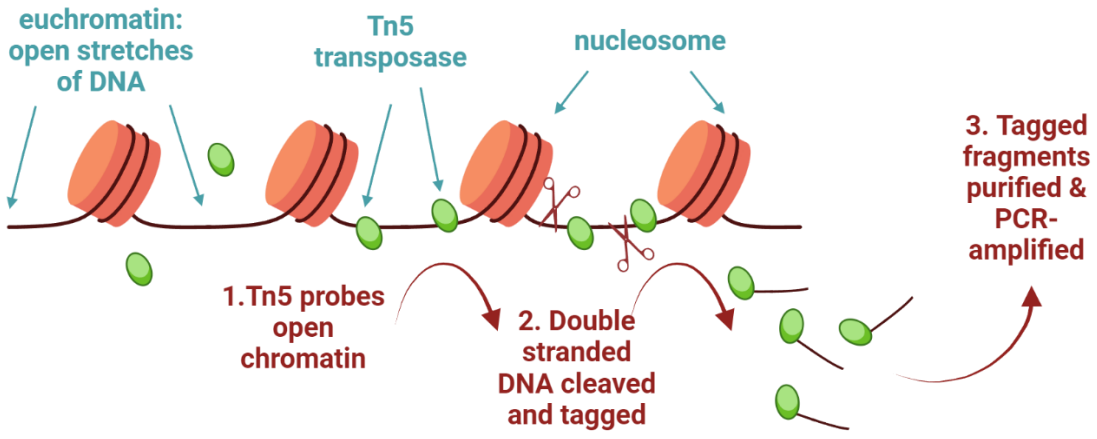



**Figure 2.2. Schematic of ATAC-seq**

Summary of ATAC-seq profiling steps. Euchromatin DNA is cleaved and tagged via Tn5 transposase with dsDNA adapters. Barcodes are attached to the adapter sequences. Adapter-attached DNA fragments are PCR-amplified. PCR-amplified DNA is sent for sequencing. Image generated using BioRender.com

### 2.5.2 ATAC-seq processing and analysis

ATAC-seq fastq processing was performed using the nf-core Nextflow ATAC-seq pipeline (Patel et al., 2023, Ewels et al., 2020). This was run on Windows Subsystem for Linux (WSL) using Visual Studio Code and was run using the .sh file "run_ATAC-seq___.sh" and a file sample sheet. The .sh summary files are available in the supplementary appendix. ATAC-seq pipeline nf-core outputs included fastQC data, MACS2 quality control assessment, .bam, .broadpeak files and consensus broadpeak files. Samples were aligned to hg38. The following steps were part of the ATAC-seq nf-core pipeline:

1. Quality control (QC) of raw fastq reads using FastQC v0.12.0 (Ewels et al., 2016).

2. Adapter trimming by Trim Galore! v0.6.5.

3. Alignment using Burrow-Wheeler Aligner (BWA- bwa-0.7.17.tar.bz2) (Li and Durbin, 2009).

4. Duplicates marked using picard v3.0.0.

5. Filtering of reads to remove blacklisted regions, marked duplicates, mismatched reads with 4 or more mismatches, unmapped reads using SAMtools v1.18, BAMtools v2.5.2 and BEDtools v2.31.0 (Quinlan and Hall, 2010, Barnett et al., 2011, Li et al., 2009).

6. QC of aligned reads using picard 3.0.0.

7. Create normalised bigWig files scaled to 1 million reads using BEDtools v2.31.0 and bedGraphtoBigWig v398.

8. Enrichment of peaks using deepTools v3.5.3 (Ramírez et al., 2016).

9. Peak calling of broadPeaks using MACS2 v2.2.9.1 (Zhang et al., 2008).

10. Annotation of peaks regarding features of interest using HOMER v4.11 (Heinz et al., 2010).

11. Create consensus peak calls across samples using BEDtools v2.31.0.

12. Merge filtered alignments across replicates using picard v3.0.0.

## 2.6 Whole genome sequencing

Whole genome sequencing (WGS) DNA preparation was performed by Emily Clough and sequencing of DNA was performed by Novogene™ utilizing paired end sequencing, generating libraries to a depth of 60 million reads. Details of WGS sequencing steps are available on the Novogene™ website.

### 2.6.1 WGS analysis

WGS quality control and initial analysis was performed by Novogene™ with alignment to reference genome hg38. Files containing single nucleotide polymorphisms (SNPs), structural variants (SVs) and CNVs were provided by Novogene™. CNV files were used to correlate DSB frequency with copy number high or low genes.

## 2.7 INDUCE-seq

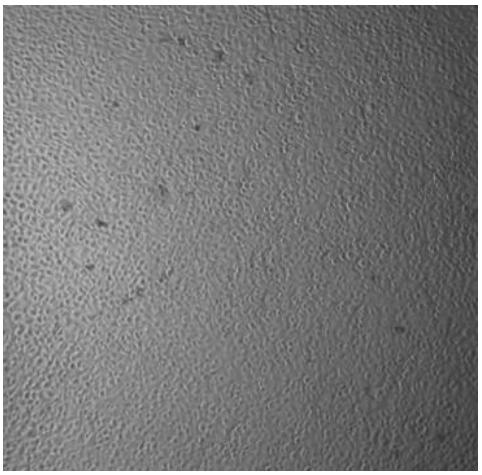
The INDUCE-seq DSB mapping technique was performed in collaboration with the Simon Reed lab in Cardiff. Cells were cultured and fixed in Glasgow and sent on ice via courier to Cardiff where they were prepared, sequenced and processed in Cardiff by Kierney O'Dare and Dr Patrick Van Eijk.

### 2.7.1 Sample preparation

For fixing and transport of cells, 96 well Greiner plates (cell culture microplate, item number: 655090) were coated first with poly-D-lysine (PDL, 0.1 mg/ml, Invitrogen #A3890401) that was diluted to 50 ug/ml with sterile Dulbecco's PBS. Wells were incubated in a laminar flow culture hood at RT for 1 hour with 50 µl of PDL working solution, after which the PDL was removed. The wells were rinsed 3 times with 100 µl of distilled water and then allowed to dry in a hood, uncovered for 2 hours. The dry plates were then wrapped in Parafilm™ and stored at 4°C until the day of use. Plates were made up no more than 2 days prior to cell fixing.

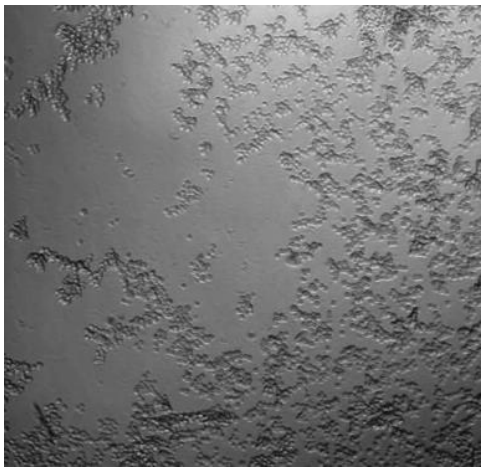
The two initial runs for INDUCE-seq were unsuccessful due to poor plate adherence of GSCs to PDL plating despite the addition of a spin step of 96-well plates for 5 minutes at 300 g to encourage cell fixing (Figure 2.3). Run 1 was done with the G7 cell line on PDL-coated plates only. Cells were grown and treated in T75 flasks, then counted and replated onto the 96 well format. G7 GSCs displayed significant cell loss following washing steps and fixation. For run 2, cells were plated and grown directly onto the 96 well plate with the aim of minimising trypsinisation steps which might have been resulting in potential cell loss. For treatment of cells using this method, a lead shield had been planned to keep controls and IR-treated on the same plate. Two cell lines were tested using this method: E2 and R10, as a means of determining the best option moving forward. Prior to IR treatment it was apparent that the GSC lines remained unable to properly adhere, displaying spherical growth in both groups; E2 GSCs more so than R10 GSCs (Figure 2.3). These cells were not taken forward for further treatment and processing as monolayer growth was required for processing.

(a)



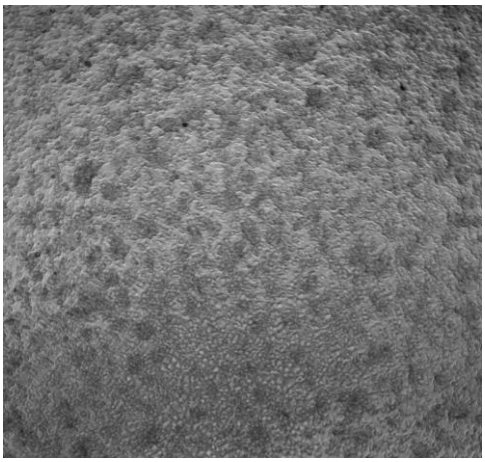
G7 differentiated PDL coated plates

(b)



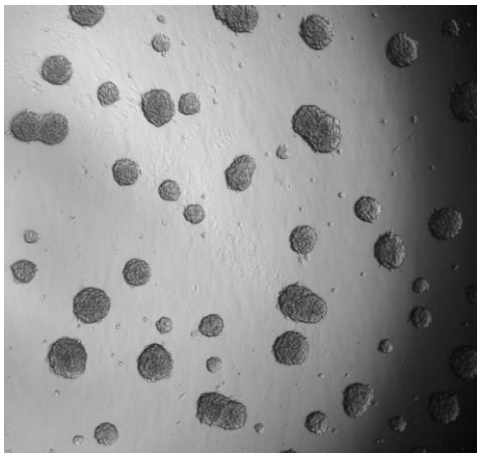
G7 GSC PDL coated plates

(c)



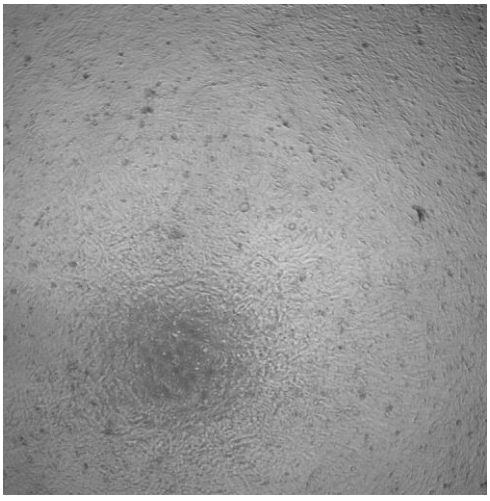
E2 differentiated PDL grown direct on plate

(d)



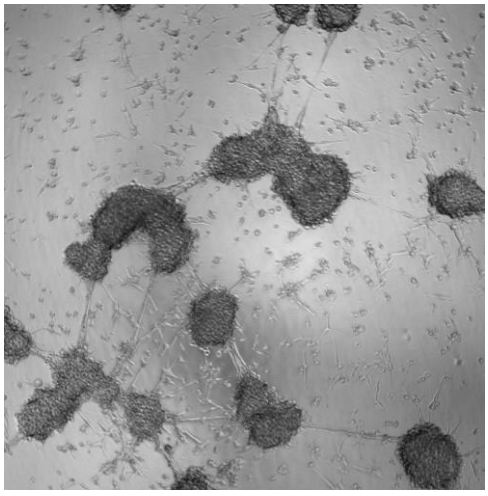
E2 GSC PDL grown direct on plate

(e)



R10 differentiated PDL grown direct on plate

(f)



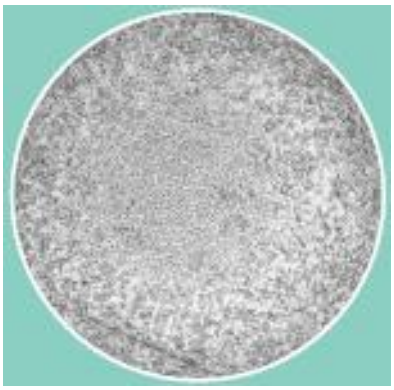
R10 GSC PDL grown direct on plate

**Figure 2.3 Representative images of trial 1 and 2 of INDUCE-seq cell plating**

Representative images of plating adherence of cells on PDL with 5 minute 300 g plate spin and fixed in 4% PFA. Images taken at 5X magnification. (a-b) Trial 1 of PDL plating. (a) G7 differentiated cells plated post IR. (b) G7 GSCs plated post IR. (c-f) Trial 2 of plating and treating directly onto 96-well plates. If successful, plan was for lead plate shield during treatment. (c) E2 differentiated cells. (d) E2 GSCs. (e) R10 differentiated cells. (f) R10 GSCs.

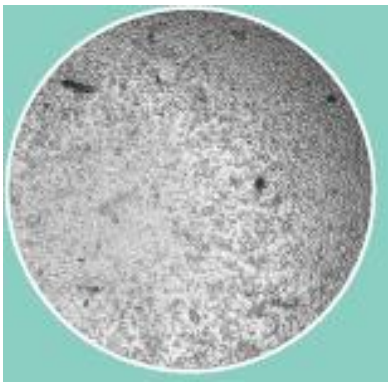
Therefore, for the third run to increase cell adherence, laminin was also used to coat plates following PDL coating after discussion with the Reed lab. For this, laminin (1 mg/ml) was prepared with ice cold PBS to a concentration of 10 µg/ml. The wells were then coated 50 µl and incubated for 2 hours are 37°C. The plates were then immediately used for plating. In addition, a separate plate was also prepared by adding 10 µl/well of FCS and incubating for 30 minutes prior to fixing to encourage cell settling of GSCs. This also succeeded in promoting cell settling, however, given that FCS is also used to differentiate GSCs, it was the laminin only plates that were taken forward for further processing (Figure 2.4). R10 cells were taken forward as a means of comparing both differentiated and GSC lines with BLISS results as well. Additionally, on review of the earlier images (Figure 2.3) R10 GSCs appeared to have the best chance of giving the even monolayer that was required for further processing.

(a)



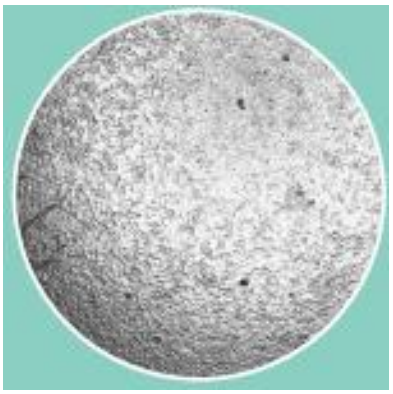
R10 differentiated control laminin plates

(b)



R10 GSCs control laminin plates

(c)



R10 differentiated irradiated laminin plates

(d)

R10 GSCs irradiated laminin plates

**Figure 2.4 R10 laminin-plated cells sent for INDUCE-seq- run 3**

Images taken by Reed lab prior to processing. Plates coated with PDL followed by laminin with 5 minute 300g plate spin to maximise plate settling. (a) R10 differentiated control cells. (b) R10 GSCs. (c) R10 differentiated IR 10Gy 24 hour cells. (d) R10 GSC IR 10Gy 24 hour cells.

For cell culture R10 GSCs and differentiated cells were cultured as previously described. For INDUCE-seq sample processing, cells had to be submitted in multiples of 8, therefore 2 repeats per condition were prepared with the following conditions being: differentiated 0 Gy IR, differentiated IR 10 Gy 24 hours, GSC 0 Gy IR, GSC irradiated 10Gy 24 hours. Cells were grown and irradiated in T75 flasks, plated to a concentration of $1x10^6$ cells. Cells were taken 30 minutes prior to the desired timepoint and trypsinised and counted as described. Cells were plated on 96 well plates at a concentration of 200,000 cells per well, pertaining to one condition and one replicate. Cells were plated in a columnar fashion (Figure 2.5).

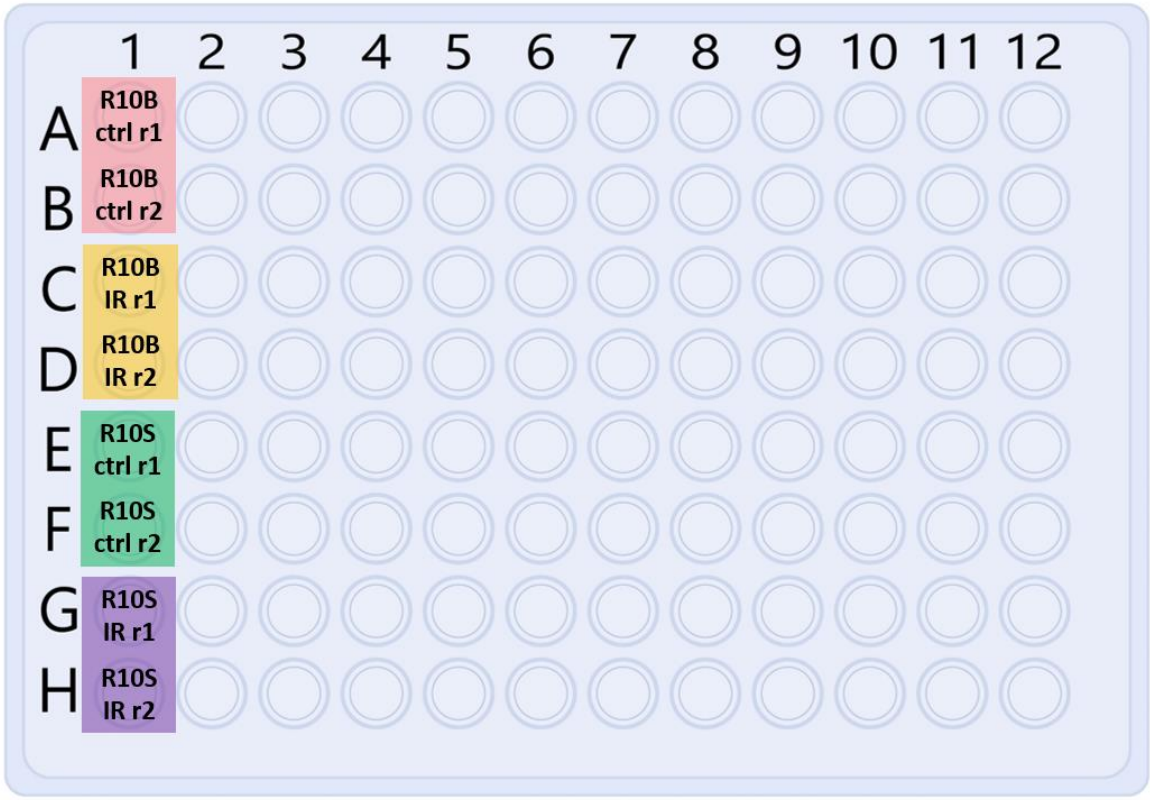



**Figure 2.5: 96 well plate layout for INDUCE-seq processing**

Cells were plated at 200,000 cells/well in 96 well Griener plates coated with PDL and laminin. Experiments were performed in n=2 repeats per condition. Cells were plated in columnar fashion as described. R10B ctrl: red 0 Gy differentiated IR cells at 24 hours. R10B IR: orange 10 Gy differentiated IR cells at 24h. R10S ctrl: 0 Gy GSCs IR cells at 24 hours. R10S IR: purple GSCs 10Gy IR cells at 24 hours.

From this stage, plating and washing was done with manual pipetting to avoid major disturbance of the cells with automatic pipetting. Cells were allowed to settle for 30 minutes in an incubator at 37°C after which they were then spun at 300 g for 5 minutes on the plates. Cells were then fixed in the hood on ice by gently aspirating media and washing once with 100 µl PBS which was then aspirated. A further 100 µl of PBS was then added to cells and 100 µl of 8% PFA, making a total concentration of 4% PFA. Cells were incubated in PFA for 10 minutes at RT. The PFA was then aspirated and disposed of as per building policy and cells were manually washed with PBS twice. Cell adhesion was checked under a microscope between wash steps. Samples were stored in 200 µl of PBS at 4°C until transport and tightly sealed with Parafilm™. Samples were transported via courier on ice.

### 2.7.2 INDUCE-seq analysis

Sample preparation following fixation and sequenced sample processing was performed at the Reed lab in Cardiff by Kierney O'Dare and Dr Patrick Van Eijk. A detailed protocol of the INDUCE-seq method is available from the original published manuscript "Precision digital mapping of endogenous and induced genomic DNA breaks by INDUCE-seq" (Dobbs et al., 2022).

In brief, samples were permeabilised, blunted and labelled in situ using full-length P5 adapters. DNA was then extracted, purified and fragmented, after which library preparation was performed with half-functional P7 adapters. This would then leave relevant DNA fragments with one P5 full length adapter and one half functional P7 adapter. These fragments would be sequenced using an Illumina flow cell, giving a "1 break for 1 read" result. For data pre-processing, the steps performed by the Reed lab were as follows: fastq trimming of adapter sequences, quality control analysis of fastq files, sequence alignment and read filtering to hg38 in SAM/BAM format, conversion to .bed and bedgraph files for break calling. Files provided by the Reed lab were .bam, .bed, .breakends (precise genomic location of individual breaks determined from sequenced reads). .breakcounts (frequency of breaks at each genomic location) and .bedgraph and .bigwig files of breaks per genomic bin of 100 kbp.

## 2.8 Immunofluorescence staining of DSB markers

### 2.8.1 Confocal cell preparation

R10 GBM GSC and differentiated cells were seeded at $8x10^4$ cells per well on 6 well plates with 22 mm glass coverslips and mounted on Matrigel™ for GSC cultures. At 24 hours cells were treated with 0 Gy IR or 10 Gy IR. For fixation, cells were treated with 1 ml/well extraction buffer (0.5 M Hepes = 2.5 ml of 100 ml 1 M Sigma #H0887, 5 M NaCl = 0.5 ml of Sigma-250 g #S7653, 0.5 M Sucrose = 30 ml of Sigma-250 g #S7903,100X triton = 100 µl of Sigma-250 ml #T8787, 0.5 M EDTA = 250 µl of Sigma-100 g #EDS, DDW = 16.5 ml, MgCl2 = 150 µl of Sigma-100 G # M8266) for 5 minutes. Cells were then washed twice with 2 ml of PBS and fixed in a laminar flow culture hood with 1 ml/well 4% PFA for 15 minutes at 24 hours post IR. The PFA was aspirated and discarded as per building policy and

cells were washed and covered with PBS. Plates were kept in PBS at 4°C for up to two weeks prior to staining.

For primary staining of cells, coverslips were blocked with a blocking solution for 30 minutes (Block: 0.1% triton PBS, 5% FBS and 0.5% bovine serum albumin: BSA Sigma) with 1ml/well. The required primary antibody was made up in DAKO (at a dilution of 1:100 to 1:200 diluted in DAKO Real Antibody Dilutent Agilent Technologies #S2022 Table 2.2) at a volume of 100 µl per coverslips. Primary staining was performed using light-tight trays lined with damp paper towel where annotated parafilm was used for mounting coverslips. 100 µl of primary antibody per coverslip was applied to the parafilm and coverslips were placed cell-side down onto the antibody. Coverslips were incubated in the dark at 4°C for 24 hours.

For secondary staining of cells, coverslips were removed from the parafilm and placed cell side up on 6 well plates for washing twice with 0.05% triton. Secondary antibodies were also made up in DAKO at 100 µl per coverslip (1:500 dilution in DAKO Table 2.2). Parafilm was used for mounting with 100 µl of antibody being applied and the coverslip placed cell side down onto the antibody. Cells were incubated in the dark for 1 hour at 37°C.

Coverslips were then washed in 6 well plates cell side up three times with 0.05% triton and subsequently mounted using Vectashield with DAPI (Vector #H-1200) and fixed with Biotium Covergrip Coverslip Sealant (Cambridge Biosciences #BT23005) onto white tipped slides. Slides were stored at -20°C and imaged within 1 week of staining.

### 2.8.2 Confocal image analysis

Immunofluorescence images were obtained by using a confocal microscope at 40x oil-mounted magnification (Zeiss 780) and z-stacks of 3 slices were taken to optimise foci capture. Images were saved as .czi and exported as maximum intensity .tiff images.

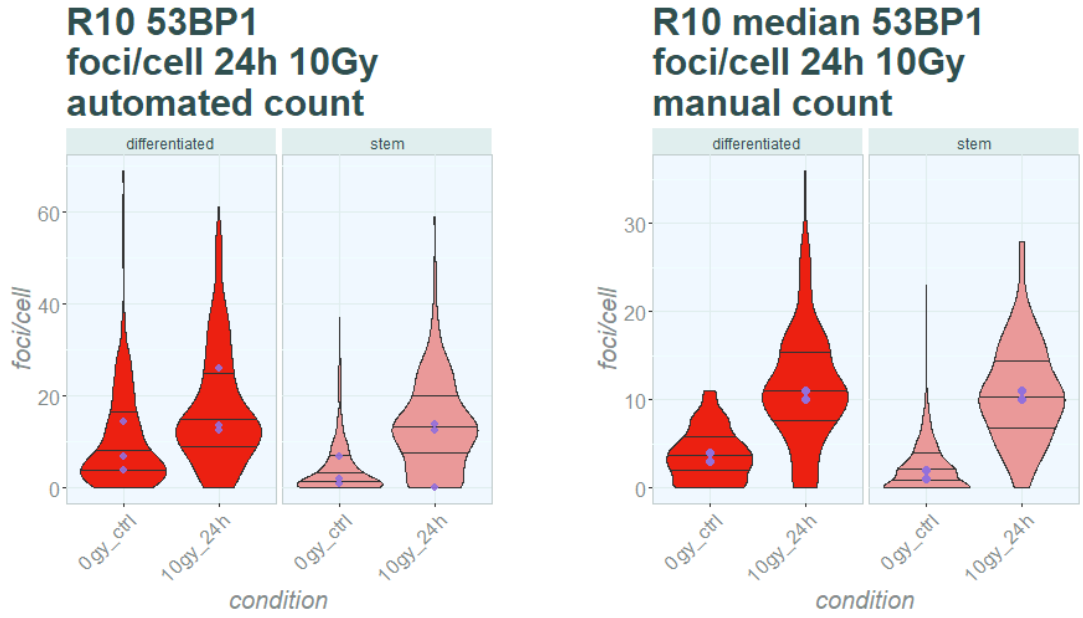
An ImageJ macro was developed in .txt format to semi-automate foci counting and foci overlap (Schindelin et al., 2012, Schneider et al., 2012). The ImageJ

macro was developed in collaboration with Dr Mark Jackson. Foci size and foci intensity were set manually by the user to be varied depending on user preference and staining. For each image cell nuclei were defined by DAPI staining and foci within the prespecified size and threshold intensity were counted and locations logged in the ImageJ log for both green and red foci. Maximum intensity images were used. Integrated density for cells was also measured using the same macro for blue, red and green channels (see supplementary appendix for macro script).

*Integrated Density = sum of all pixel values in nucleus * area of one pixel*

For replicate 3, it was noted that there was a median of 0 foci in the 10 Gy 24 hours IR group for GSCs. Therefore, manual counting of 53BP1 foci was also used to assess reliability of foci measurements. This demonstrated a disparity between 53BP1 automated count foci and manually counted foci. On inspection of the images, it was evident that some cells displayed broad IF across the nucleus for 53BP1 antibody staining (Figure 2.6). Whilst foci were identifiable with appropriate thresholding, this was not accounted for by the ImageJ macro in use. Therefore, for 53BP1 foci captured via confocal microscopy these were counted manually.

(a) (b)




(c) (d) (e) (f)

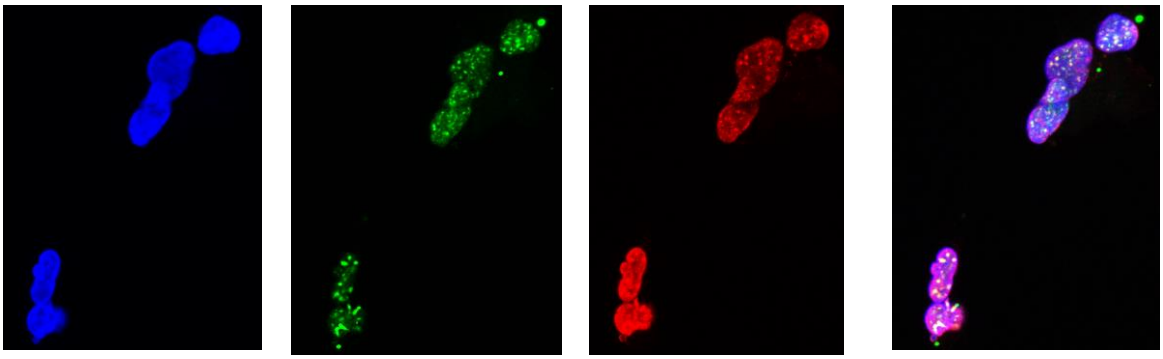


**Figure 2.6 53BP1 foci counting using automated ImageJ foci counter macro and manually counted foci**

Total counts displayed as violin plot across three repeats with repeat medians displayed as purple dots. Interquartile ranges displayed on violin plots. Dark red (left) for differentiated cells. Pink (right) for GSCs. (a) Automated counts calculated on ImageJ. (b) Manually counted results. (c-f) Example of 53BP1 staining that precluded foci counting with broad 53BP1 staining across multiple cells despite the presence of foci in repeat 3. (c) DAPI staining of nucleus in blue. (d) γH2AX staining in green. (e) 53BP1 staining in red. (f) merged image.

## 2.8.3 Opera cell preparation

For using the Opera Phenix™ high content screening system, E2 GSCs were plated on 96 well plates (Greiner), coated with 50 µl/well of Matrigel™ for 30 minutes for GSC adherence. Cells were plated at a concentration of $7.5x10^3$ cells/well for a minimum of 24 hours until adequate adherence and spread was reached.

No cells were plated on the outermost wells but empty wells were filled with 200 µl of PBS to maintain plate humidity. Cells were then irradiated at 10 Gy or 0 Gy irradiated and fixed at 6 hours. The permeabilization and fixation process was performed using the same process above: permeabilising and extraction, PBS washes, 4% PFA fixation and storage in PBS as described in the previous section. For primary antibody staining, cells were blocked for 30 minutes with the previously described blocking buffer. Primary antibody concentrations were diluted in DAKO at a concentration of 1:200 (Table 2.1) and wells were covered with primary antibody at a volume of 50 µl/well. Plates were light sealed with tin foil and incubated at 4°C for 24 hours. The wells were then washed manually with 100 µl/well of 0.05% triton and secondary antibody applied. Secondary antibody was made up at a concentration of 1:500 in DAKO and additionally DAPI (4', 6-Diamidino-2-Phenylindole, Dihydrochloride, 10 mg, Thermo Fisher #D1306) was added to the secondary block at a concentration of 1:300. Wells were incubated with 50 µl/well for 1 hour at 37°C and then washed twice with 100 µl of 0.05% triton and light sealed until imaging. Imaging was performed within 3 days of cell staining.

**Table 2.2 Immunofluorescence antibodies, stains and dilutions**

| Target | Antibody | Dilution |
|---|---|---|
| **γH2AX** | Anti-phospho-Histone H2A.X (Ser 139) clone JBW301 (05-636) Millipore (sigma) mouse | 1:100 or 1:200 |
| **53BP1** | Anti-53BP1 BP1 (05-726) Millipore (sigma) mouse | 1:100 |
| **53BP1** | Anti-53BP1 antibody (ab21083) abcam rabbit | 1:100 |
| **53BP1** | Anti-53BP1 antibody #4937 cell signalling rabbit | 1:200 |
| **goat anti-mouse 647** | Alexa fluor 647 (A21236) invitrogen | 1:500 |
| **goat anti-rabbit 633** | Alexa fluor 633 (A21071) invitrogen | 1:500 |
| **goat anti-mouse 488** | Alexa fluor 594 (A11017) invitrogen | 1:500 |
| **DAPI** | 4', 6-Diamidino-2-Phenylindole, Dihydrochloride, 10 mg, Thermo Fisher #D1306 | 1:300 |

Antibody stains for immunofluorescence experiments as displayed. Antibody dilution displayed on right most column. Company and code number displayed.

## 2.8.4 Opera image analysis

Cells were imaged on the Opera Phenix™ high content screening system which was performed under the supervision of Lynn McGarry. Appropriate regions for

well analysis were identified and 22-30 images per well were obtained across 5 wells per condition repeat at a magnification of 20x using a water-based objective. Set up for image analysis for foci counting was designated under the advice of Lynn McGarry in order to capture cell nuclei, identify relevant foci and foci overlap and report mean fluorescence intensity/cell results.

*Mean Fluorescence Intensity = Sum of the values of all nuclear pixels / Number of pixels in nucleus*

### 2.8.5 Statistical analysis

Statistical analysis for both confocal image results and Opera image results was performed in RStudio. Graphs were generated using “ggplot2”(Wickham, 2016) and “tidyverse”(Wickham et al., 2019) was used for data parsing. For statistical testing, the mean of medians across three repeats was used. ANOVA and post-hoc Tukey testing with BHP correction was employed to assess for significance, with a p-value of $<0.05$ being deemed a significant statistical result.

## 2.9 Generation of graphs, figures and statistical analysis

Graph and figure generation was performed in RStudio as was statistical analysis. For each results chapter RStudio scripts were generated for reproducibility. These are available in the supplementary files and listed in the appendix. A list of the packages, functions and theme used is also available in the supplementary file appendix.

# Chapter 3 Characterising DSBs in GBM

## 3.1 Introduction

To characterise the role of DSBs and their location in GSC treatment resistance, the baseline pattern of endogenous DSBs was established. This will be outlined in detail in this chapter.

### 3.1.1 Characterising DSB patterns in GSCs

As previously introduced, GSCs are highly resistant to radiation and other therapies (Lathia et al., 2015, Carruthers et al., 2018). GSCs have been postulated as the tumour cells responsible for inevitable GBM recurrence despite maximal therapy with surgery and chemo-radiotherapy (Emlet et al., 2014). A key feature of GSCs is an elevated level of RS, which has been associated with aberrant DDR and subsequent radioresistance (Carruthers et al., 2018, Bartkova et al., 2010, Bao et al., 2006). Interestingly, elevated RS levels have also been identified in neural progenitor cells which may confer some important shared mechanisms (Ballarino et al., 2022, Wei et al., 2016).

The elevated level of RS in GSCs has been proposed as a means of priming the aberrant DDR to allow rapid repair from radiotherapy-induced DSBs (Carruthers et al., 2018). Elevated RS is known to be associated with DSBs and therefore RS may not only drive radiation resistance but also enhance mutagenic potential (Tsegay et al., 2019, Irony-Tur Sinai and Kerem, 2018). DSBs are known to be highly deleterious with even a single persistent DSB being capable of triggering cell death in normal cells (Featherstone and Jackson, 1999).

Interestingly, GSCs also demonstrate markers of persistent DSBs, even at baseline (Carruthers et al., 2018). Therefore, there may be a subgroup of endogenous DSBs present that do not result in cell death but rather prime the DDR for enhanced repair of lethal DSBs. Mechanisms of DSB formation play an important role as these can be both physiological and pathological. For example, RS from resultant replication-transcription conflicts is associated with GSCs as well as DSBs. Additionally, non-canonical DNA structures such as R-loops and G4 have been associated with cancer and DSB formation (Kumari et al., 2019,

Hamperl et al., 2017). Furthermore, topological stressors on the DNA require active response from topoisomerases to induce DSBs as a means of alleviating torsional stress from transcription elongation, replication and DNA reorganisation (Bunch et al., 2015). All of these may play a role in shaping the GSC baseline breakome.

Finally, an important consideration is whether there are particular DSB sites following radiation that are preferentially repaired such as highly transcribed regions, areas in accessible euchromatin or regions responsible for replication origin firing. GSCs can survive high doses of radiation. The DDR in GSCs has been established as aberrantly upregulated but it is not established whether the DDR response is globally upregulated or whether the DDR is particularly directed to repair regions promoting cell survival. The DSBs caused by IR are largely assumed to be stochastic and, whilst these persistent DSBs are relatively rare, the global change in the breakome following IR may help in determining how the DDR response in GSCs changes from baseline. To quantify the differences between baseline DSBs and following IR, a baseline description of DSBs and location must first be established.

Therefore, in this chapter the baseline pattern of DSB locations in GSCs will be described to act as comparator to other cell groups including GBM differentiated progeny cell lines, non-GBM lines and DSBs persisting following IR.

### 3.1.2 DSBs in previously published data

Whilst the origin of GSCs have remained somewhat controversial, GSCs have been postulated to arise from NSCs at the subventricular zone, giving them their ability to repopulate tumour burden and self-renew (Lee et al., 2018). This makes neural cells a highly relevant cell population to investigate alongside our GSC lines. Neural cells have already been investigated using BLISS by Ballarino et al. (Ballarino et al., 2022), with the data now publicly available. Other commercial cancer cell lines have been investigated using BLISS including cancer cell lines MCF7, a breast cancer cell line, and K562, an erythroleukaemia-derived cell line (Ballinger et al., 2019). The breakomes of these neural cells and non-GBM cancer cells, were compared across our three GSC breakomes. The patterns

and locations of DSBs in GSC and non-GSC lines was characterised to give an appreciation of unique features of the GSC breakome in a wider context.

### 3.1.3 Chromatin profiling

The chromatin context; encompassing the euchromatin-heterochromatin state and chromatin modulation, in cancer is understood to have major driving influences on mutation frequency and gene expression (Polak et al., 2015, Beck et al., 2012). Chromatin remodelling in GBM specifically has been cited as a driver of treatment resistance (Liau et al., 2017, Chen et al., 2022). Regarding DSB induction and chromatin distribution, proximity to nucleosomes has been identified as having an apparent protective relationship against IR, specific to DSBs (Brambilla et al., 2020). This also raises the question as to whether chromatin remodelling acts to facilitate early DSB repair or may protect against the induction of damage. Interestingly, chromatin remodelling has also been associated with RS resolution by promoting replication fork stability and replication error processing (Fournier et al., 2018). Characterising DSBs in the context of chromatin architecture may assist in better defining the pattern of GSC DSBs. ATAC-seq is a method of defining the location of accessible areas of euchromatin within the genome. In brief, ATAC-seq uses Tn5 transposase which acts to induce nicks in the DNA and allow integration of adapters into the DNA which allows areas with adapters to be isolated for sequencing. These isolated locations are described as “peaks” of euchromatin (Grandi et al., 2022).

### 3.1.4 Highly broken DSB regions

Sites of high DSB density were isolated as a means of determining any region-specific features that might correlate to the increase in DSB density. The sites or “hotspots” with the highest DSB density across GSC lines were identified and compared between the 3 GSC lines and their differentiated progeny. These differentiated lines, derived from the original corresponding primary cell lines, allowed for a comparative analysis between radioresistant GSC populations and radiosensitive differentiated cell lines. Differences between GSCs and differentiated cell lines are explored throughout this thesis in addition to comparison of GSCs with neural cell line and commercial cancer cell line data.

### 3.1.5 Aims

This chapter gives an overview of the DSB landscape in three GBM GSC lines (GSCs: E2, G7 and R10). Broad patterns of BLISS-detected DSBs across the genome were observed across these three GSC lines and in addition, publicly available BLISS breakome data from neural cells and two commercial cancer cell lines were also described. An overview of euchromatin enrichment across GSC genomes was outlined alongside GSC DSB patterns. Finally, DSB hotspots were isolated across GSCs and also across corresponding differentiated cell populations to identify overlapping locations between cell lines and populations.

- Give an overview of genome-wide BLISS-detected DSBs across E2, G7 and R10 GSC cell lines

- Compare DSB patterns at a whole genome level for publicly available BLISS data in neural cell lines: NES, NPC and NEU and commercial cancer cell lines: MCF7 and K562.

- Outline the genome-wide euchromatin enrichment in GSCs alongside DSB patterns.

- Identify if gross characteristics are shared between regions of high DSB density between GSC cell lines.

## 3.2 Methods

### 3.2.1 Breaks ligation in-situ sequencing DSB mapping

To profile DSB location, BLISS was performed in the three cell lines E2, G7 and R10 GSCs and differentiated cells as per the protocol described by Yan et al. (Yan et al., 2017). This method is described in detail in chapter 2.

## 3.2.2 Publicly available datasets

### 3.2.2.1 Neural cell data

The publicly available neural cell data by Ballarino et al. (Ballarino et al., 2022) describes DSBs in three neural cell lines at varying levels of differentiation. These three lines from least to most differentiated were: NES, NPC and finally NEU lines. These neural cells were derived from the NES human cell line where cells were cultured in medium to differentiate cells into NPC and NEU subgroups. To investigate DSB distribution in these lines, downloaded bed files were transferred to hg38 from hg19 using LiftOver tools via RStudio.

### 3.2.2.2 Commercial cancer cell data

As previously mentioned, profiling of DSBs using BLISS has been performed in other cancer cell lines such as K562; a line derived from erythroleukaemia and the breast cancer line MCF7. Bed files from the data provided by Ballinger et al. (Ballinger et al., 2019) were obtained and the LiftOver tool was used to transfer files from hg19 to hg38 alignment via RStudio.

## 3.2.3 ATAC-seq datasets

ATAC-seq libraries were prepared by Karen Strathdee in matched GSC and differentiated cell lines in E2, G7 and R10. Details of the alignment and pre-processing are available in chapter 2. In brief, ATAC-seq used Tn5 transposase to probe open regions of euchromatin and tag these with dsDNA adapters which were barcoded and PCR-amplified. These were sent for next generation sequencing and subsequent processing was performed using Nextflow nf-core ATAC-seq pipeline analysis. In brief, the nf-core ATAC-seq pipeline included adapter trimming, alignment to hg38, filtering of duplicates, blacklists and unmapped reads followed by generation of the following files: .bigWig, .broadPeak, .bam, and consensus peaks from MACS2.

## 3.2.4 Genome-wide analysis

For the above datasets, sequenced DSB data was investigated and processed using RStudio (versions 3.9-4.3). A mapping overview across chromosomes 1 to 22 was performed using “karyploteR” packages (Gu et al., 2014, Gel and Serra,

2017) to identify large genomic differences and regions of interest. Sex, alternate assembly chromosomes and mitochondrial chromosomes were excluded on the basis that lines were derived from both male and female subjects and that alternate assembly chromosomes were not shared across cell types. Cell lines were plotted against the whole genome to describe break pattern at a whole genomic level. Plots were generated by using density plots from bedgraph DSB files to give an overview of DSBs across lines. Plots were also represented across chromosomes by relative chromosome length using bigWig files as density plots. Blacklisted regions were excluded from analysis which included highly repetitive regions such as centromeres, telomeres and many parts of the short arms on acrocentric chromosomes, hence the apparent DSB deserts at these sites (Amemiya et al., 2019). Plots of DSB frequency were also generated as genomic line plots using "circlize" RStudio package (Gu et al., 2014). Plots of DSB frequency were generated from 50 kbp region bins as a genomic line plot. For this, the hg38 genome was divided into 50 kbp regions from the chromosome start site to the end of each chromosome. Total DSBs per 50 kbp region were calculated and plotted along the length of the circos karyoplots.

For DSBs across chromosomes, DSBs/kbp per chromosome was calculated for all cell lines and normalised for DSBs per individual repeat. To represent DSB density across chromosomes, chromosomal DSB density was expressed as a fold change of the overall DSB density per individual repeat.

DSB density per chromosome:

*DSBs/kbp = (DSBs in chromosome/length of chromosome)/1000*

DSB density per cell line repeat:

*DSBs/kbp = (DSBs in cell line repeat/length of genome)/1000*

Fold change was calculated per chromosome as chromosomal DSB density/cell line DSB density for cell lines E2 GSCs, G7 GSCs, R10 GSCs, mcf7, k562, NES, NPC and NEU neural cell lines. Results were tabulated to demonstrate range and also displayed as a boxplot and jitter to show individual results per chromosome.

### 3.2.5 Highly broken DSB regions

The whole genome was divided into 50 kbp bin regions as per previous BLISS studies (Ballinger et al., 2019). For each bin, the total number of DSBs was calculated. Genome binning was used to investigate the most frequently broken regions. For analysis of the locations with the greatest number of DSBs per cell line 50 kbp bin regions were used. Mapping of DSBs across highly broken regions was performed using "karyploteR" and "RIdeogram" packages(Gu et al., 2014, Gel and Serra, 2017, Hao et al., 2020) for shared regions and identifying related structures. For the two 50 kbp regions with the highest DSB density, density was also plotted using Integrative Genome Viewer (IGV). Binned regions of 50 kbp were used as a means of giving a balance between a broad overview of DSBs within these regions without losing potentially important surrounding features by using a smaller bin size and in line with previously published BLISS data (Ballinger et al., 2019). These were sorted into order of descending frequency of DSBs and maps were generated. Gene sites were also annotated on karyoplots to assist in identification of landmarks relative to break frequency. For 50 kbp analysis, GSC and differentiated progeny cells were used as a means of establishing shared or contrasting spread of highly broken regions. Established differentiated GBM lines were used for comparison to GSC lines.

## 3.3 Results

### 3.3.1 Endogenous breaks overview

#### 3.3.1.1 GSCs show shared break sites across lines

All GSC lines were plotted against the whole genome to describe the breakome at a whole genomic level. Figure 3.1 represents DSB patterns by cell line and chromosomes across repeat 1 for the three GSC lines using DSB density across chromosomes. Repeats were treated individually in this case as the overall DSB pattern was similar within repeats and this allowed a comparison between GSC lines. As previously noted, blacklist regions such as centromeres and short-arms of acrocentric chromosomes (chr13, chr14, chr15, chr21 and chr22) do not demonstrate DSBs due to low mappability (Amemiya et al., 2019). At a broad

genomic overview level, the patterns of DSBs across cell lines appeared similar. At chromosome 11q arm, there was one particularly notable shared peak across E2, G7 and R10 GSCs. On chromosome arm 13q, E2 and R10 GSC lines demonstrated a peak that was not shared across G7.


E2S_1
G7S_1
R10S_1
chr1
chr2
chr3
chr4
chr5
chr6
chr7
chr8
chr9
chr10
chr11
chr12
chr13
chr14
chr15
chr16
chr17
chr18
chr19
chr20
chr21
chr22
E2S_1
G7S_1
R10S_1
chr11

**Figure 3.1. Repeat 1 for GSC lines E2, G7 and R10 DSB density by individual chromosomes**

Density plot of DSBs across repeat 1 for 3 GSC cell lines E2, G7 and R10. DSB density represented across chromosomes by relative chromosome length using bigWig files for DSB density. Lines from top to bottom: E2 GSC repeat 1 (red), G7 GSC repeat 1 (green), R10 GSC repeat 1 (purple). Karyoplots are displayed below DSB density. Red markers on karyoplots represent centromeres. Chromosomes 1 to 22 displayed. Peak at chromosome 11q highlighted in red box and zoom.

From Figure 3.2; a condensed DSB density plot, there were further regions demonstrating shared peaks across the different GSC cell lines. As with Figure 3.1, q11 continued to demonstrate a peak shared across the three cell lines (Figure 3.2). The peak between E2 and R10 at 13q became less apparent, however, a peak in chromosome 6 was more evident using this overview. All three cell lines demonstrated some variation of DSB density across chromosomes. Overall, R10 and G7 both demonstrated a drop in DSBs across chromosome 10 which was not seen in E2 (Figure 3.2). Conversely, chromosome 7 demonstrated a higher relative DSB density in R10. All three cell lines showed a relative drop in DSB density from baseline in chromosome 18. Cell line G7 appeared to show less DSBs overall in chromosomes 13 and 22. Overall, these findings demonstrate some shared patterns in DSB density between GSC lines.

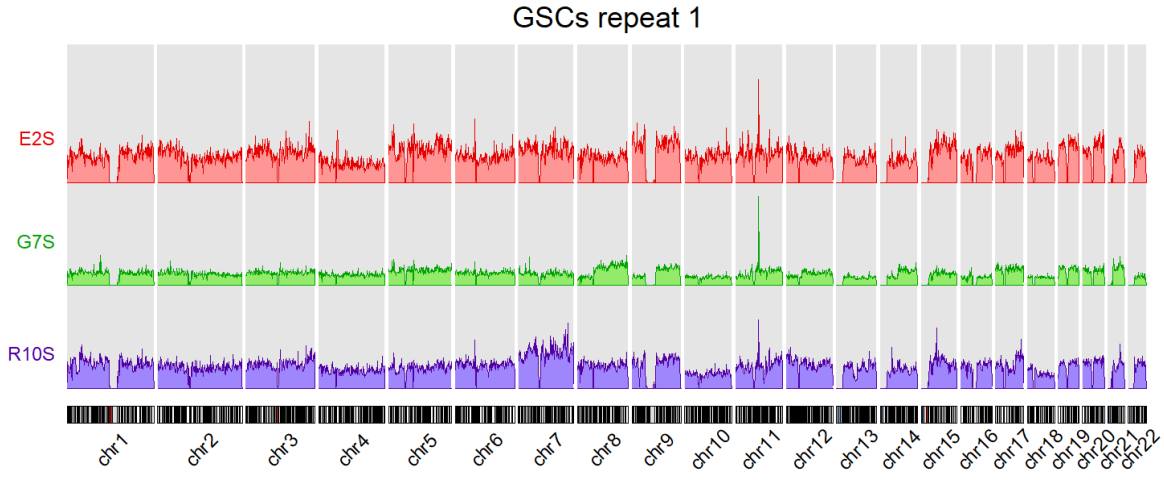



**Figure 3.2. Repeat 1 for GSC lines E2, G7 and R10 DSB density across chromosomes 1-22**

Density plot of DSBs across chromosomes 1-22 in numerical order. Chromosomes from repeat 1 in order of left to right. From top to bottom: E2 GSC repeat 1 (red), G7 GSC repeat 1 (green), R10 GSC repeat 1 (purple). Karyoplot of chromosomes 1-22 displayed below.

## 3.3.2 DSBs in other cell lines

A number of BLISS DSB datasets are publicly available and were used to investigate broad DSB patterns across other cells as a comparison to the three GSC lines. Publicly available BLISS datasets including data from neural cell lines NES, NPC and NEU (Ballarino et al., 2022) and two cancer cell lines K562 and MCF7 were used (Ballinger et al., 2019). The three neural cell lines represented three levels of neural cell differentiation, starting with NES as the least differentiated subtype. These were obtained from the self-renewing cell lines AF22, which displayed IF staining for Nestin and SOX2 (Falk et al., 2012). To obtain differentiated cells, NES cells were cultured in differentiating media for 6 days for NPCs and the NEU cells were cultured in differentiating media for a total of 35 days which expressed GFAP as an indication of differentiation.

### 3.3.2.1 Neural cell DSBs demonstrate greater uniformity across chromosomes than GSCs

Neural cell lines are displayed in Figure 3.3 and displayed a broadly similar density of DSBs across all lines with some peaks displayed across each line such as NEU and NPC demonstrating a peak at chromosome 10 and chromosome 21. Otherwise, there were few clear discernible peaks above DSB baseline in the neural cell lines. The NES cell line did not appear to demonstrate any clear peaks that were visible at this overview level and had no obviously apparent shared regions with NPC and NEU lines. Overall, DSB density across chromosomes appeared relatively similar to baseline for NPC and NEU, however NES appeared to show a small dip in DSB density at chromosome 19.

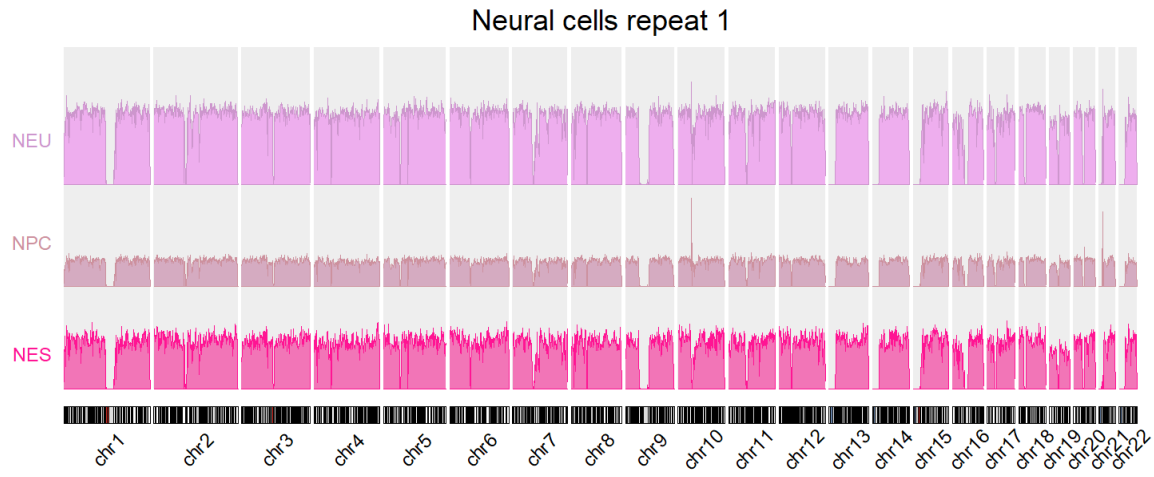



**Figure 3.3. Repeat 1 for neural cell lines NES, NPC and NEU DSB density across chromosomes 1-22**

DSB density plot of chromosomes 1-22 from neural cells left to right. DSB density across chromosomes from top to bottom: NEU (lilac), NPC (dark pink), NES (hot pink). Karyoplot of chromosomes 1-22 displayed below.

Neural cells were also plotted in Figure 3.4 as a DSB frequency plot given that few discrete DSB peaks were apparent on the displayed density plot. For the peak seen in NPC and NEU lines at chromosome 10, this remained clearly visible, though the peak at 21 was less apparent on the DSB frequency plot. Conversely, chromosome 7 appeared to show a peak shared across NPC and NEU lines, with a small peak in one of the NES repeats also. Overall, DSB peaks across NES lines remained distinct to NPC and NEU lines. At chromosome 20, NES lines showed a peak across both repeats. Across the two NES repeats there were peaks visible in both, though fewer were shared between the repeats than in NPC and NEU lines where most peaks that were visible at an overview level were visible in both repeats. These results demonstrated a greater uniformity of DSB density than that of the GSC lines across chromosomes.

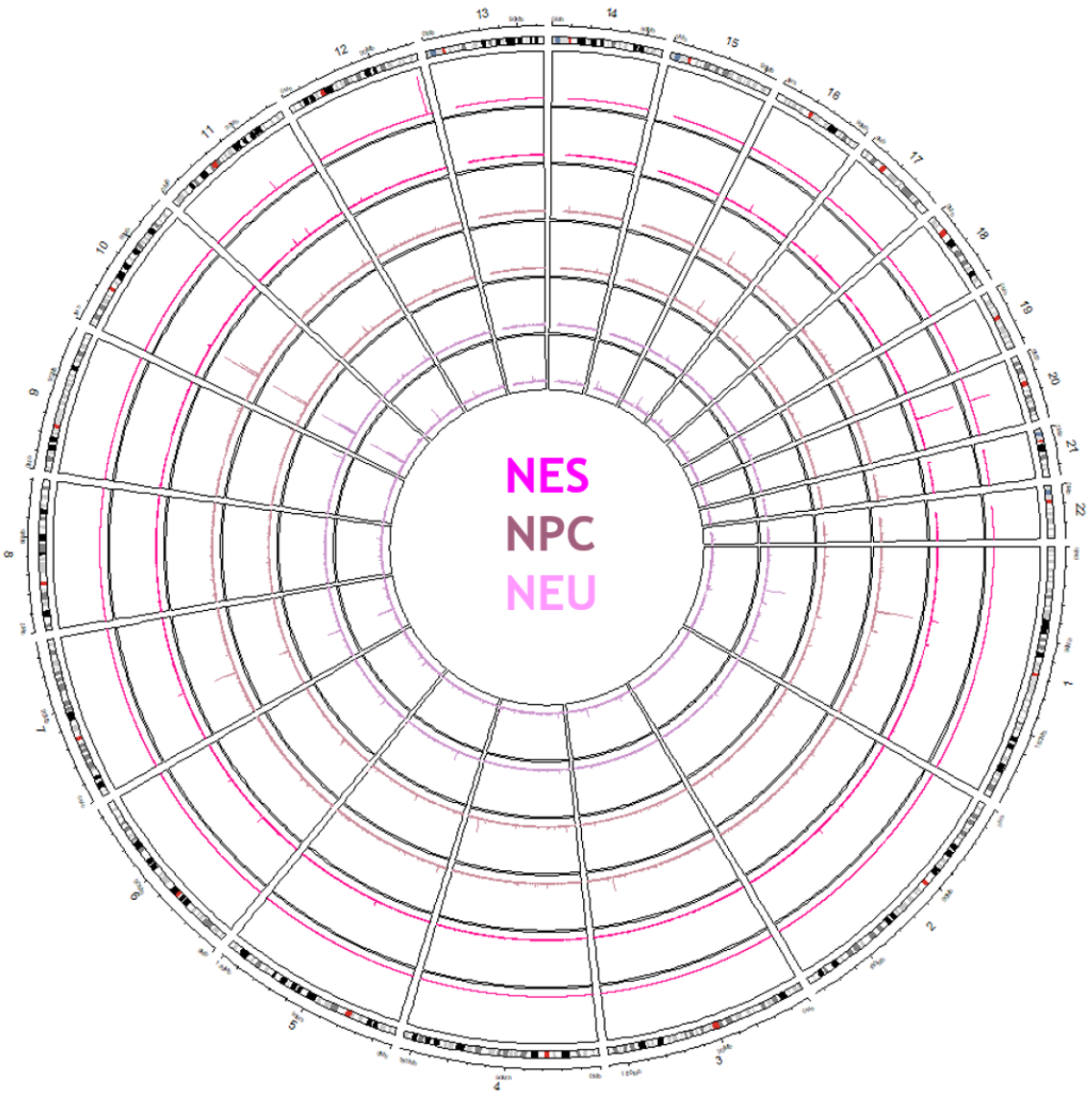



**Figure 3.4. Neural cell lines NES, NPC and NEU circos DSB frequency across chromosomes 1-22**

Frequency plot of DSBs across chromosomes per cell line repeat. Free scaling used to allow visualisation of DSB frequency occurring significantly above baseline. Neural cell lines displayed from inside to outside: NEU repeat 1 and 2 (pale pink), NPC (dark pink) repeat 1 and 2, NES repeat 1 and 2 (hot pink). Y-axis displays DSB frequency across chromosomes. Karyoplot of chromosomes represented as most outer ring. Centromeres visible as red band on karyoplot.

#### 3.3.2.2 DSB distribution in commercial cancer cell lines demonstrates variability in DSB distribution across chromosomes

As previously stated, profiling of DSBs using BLISS has been performed in other cancer cell lines including K562; a line derived from erythroleukaemia and the breast cancer line MCF7. Using the data provided by Ballinger et al. (Ballinger et al., 2019), an overview of DSB frequency is represented graphically in Figure 3.5 across K562 and MCF7 cell lines. Both MCF7 and K562 display distinct DSB peaks across multiple chromosomes. This is most notable in the MCF7 line at

chromosome 1, 3, 17 and 20. The cell line K562 also demonstrated peak regions above baseline at chromosomes 9, 13 and 22. Overall, baseline variability of DSBs appeared greater in both MCF7 and K562 lines than in neural cell lines and was similar to the variable pattern of DSB density seen in GSC lines. Figure 3.5 also displays a circos plot of the two commercial cancer cell lines with GSC cell lines for total DSB frequency across chromosomes. MCF7 appears to maintain the DSB peaks across chromosome 1,3, 17 and 20. Peaks in K562 at 9, 13 and 22 also remain visible. Overall, MCF7 and K562 cell lines demonstrate distinct DSB distribution from each other and also from GSCs. However, the shared peak in chromosome 11q seen in GSCs was also visible in both commercial cancer cell lines.

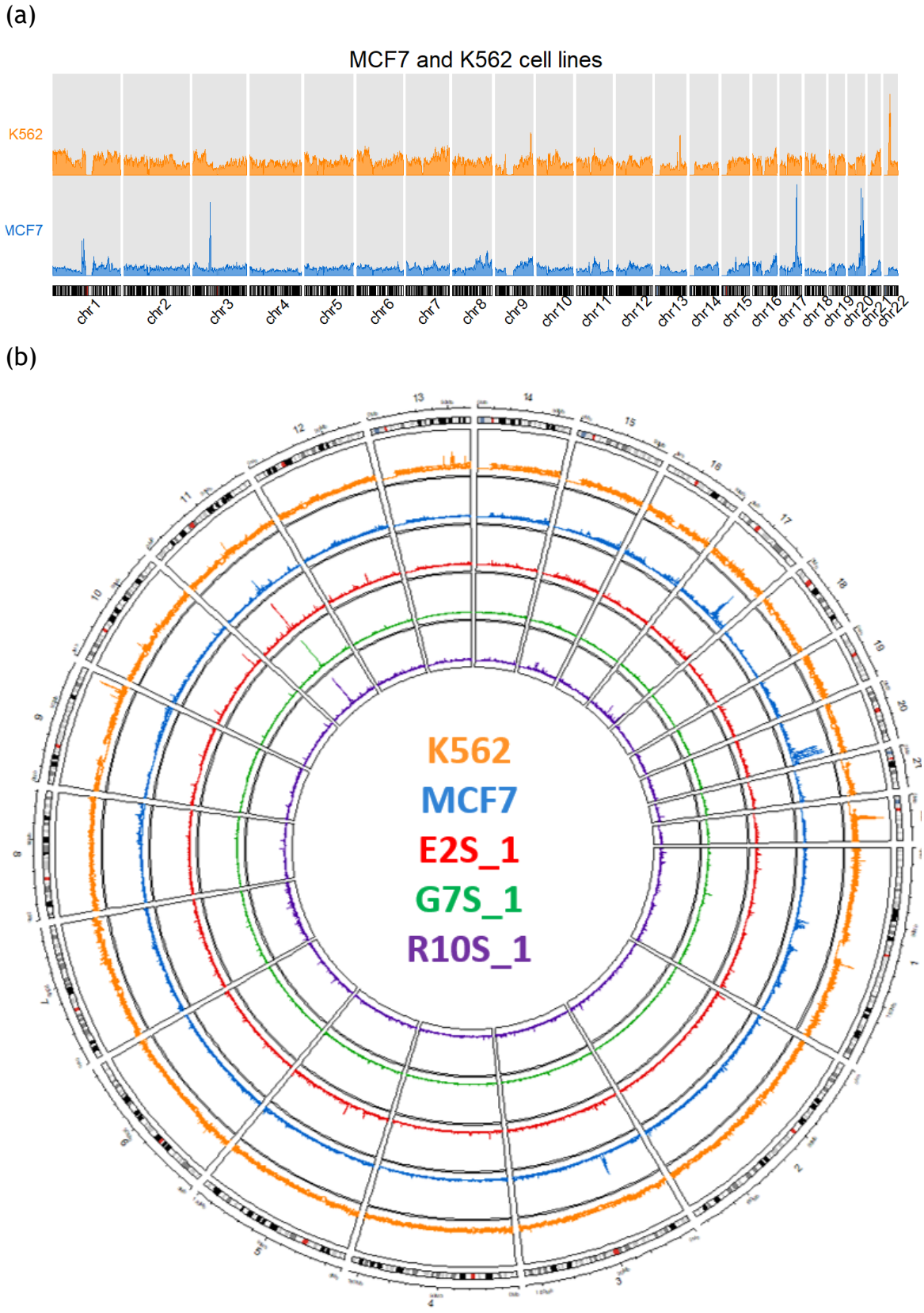



**Figure 3.5 DSB patterns across commercial cancer cell lines**

(a) Density plot of DSBs across commercial cell lines K562 (orange) and MCF7 (blue), y-axis displays DSB density across chromosomes 1-22. (b) Circos plot of DSB frequency across cancer cell lines. From inside to outside: R10 GSC repeat 1 (purple), G7 GSC repeat 1 (green), E2 GSC repeat 1 (red), MCF7 (blue), K562 (orange). Y-axis displays DSB frequency across chromosomes 1-22. Karyoplot displayed as outermost ring. Centromeres visible as red band across karyoplot.

In Table 3.1, the DSB density across chromosomes relative to global density per genome is displayed for all cell types including GSCs, commercial cancer cells and neural cells. **Error! Reference source not found.** also displays relative DSB density across chromosomes as a representation of overall spread for each cell line as well as DSB/kbp per chromosome. The range of relative DSB density across chromosomes in GSCs demonstrated the smallest range in E2S repeat 2 (0.659 to 1.256; chromosomes 22 and 5 respectively) and the largest range in R10 repeat 2 (0.642 to 1.730; chromosome 18 and 7 respectively). All three R10 GSC lines had relative density change of at least 1.6 or above in chromosome 7 and this was consistently the largest relative increase across the R10 GSC repeats. Chromosomes 18 and 10 had a relative decrease of 0.67 or lower across the three repeats in R10 GSCs. E2 GSCs demonstrated chromosome 22 having the greatest relative decrease in DSB density. G7 GSCs also demonstrated consistency across repeats where the greatest decrease in relative DSB density was seen in chromosome 22 across all repeats.

Of all the cell lines MCF7 had the largest range across chromosomes with a range of 0.562 (chromosome 18) up to 2.296 (chromosome 20). Cancer cell line K562 had a more modest variation of relative DSB density with a range from 0.589 (chromosome 14) to 1.253 (chromosome 7).

Neural cells demonstrated a relatively lower range variation compared to GSC and commercial cancer cells with the largest range in relative DSB density being 0.640 to 1.103 in NEU repeat 2 (NEU_2 chromosomes 22 and 6 respectively) though a generally greater overall DSB burden compared to GSCs and K562. Neural cells had a closely similar interquartile range of between 0.867 to 1.066. Chromosome 22 had the greatest relative decrease across NES, NPC and NEU cells. Chromosome 6 had the greatest relative increase in DSB density in all neural cell repeats apart from NPC repeat 2 where chromosome 10 had a relatively higher DSB density.

Across all cell lines, there was a significant difference in variance between lines of relative DSB density (Levene's test for homogeneity of variance $p = <0.0001$) but no overall difference between medians across lines in post-hoc analysis (Kruskal-Wallis $p=0.587$, effect size small = -0.0019).

**Table 3.1 Relative DSB density across chromosomes**

| | Median fold change across chromosomes | Interquartile range | | Range | |
|---|---|---|---|---|---|
| **E2 GSC rep 1** | 0.976 | 0.860 | 1.143 | 0.683 | 1.328 |
| **E2 GSC rep 2** | 0.978 | 0.859 | 1.148 | 0.659 | 1.256 |
| **G7 GSC rep 1** | 0.987 | 0.835 | 1.164 | 0.502 | 1.397 |
| **G7 GSC rep 2** | 0.993 | 0.832 | 1.214 | 0.516 | 1.432 |
| **G7 GSC rep 3** | 0.987 | 0.839 | 1.198 | 0.492 | 1.380 |
| **R10 GSC rep 1** | 0.985 | 0.816 | 1.094 | 0.655 | 1.661 |
| **R10 GSC rep 2** | 0.974 | 0.810 | 1.098 | 0.642 | 1.730 |
| **R10 GSC rep 3** | 0.989 | 0.823 | 1.063 | 0.666 | 1.633 |
| **K562** | 1.004 | 0.912 | 1.044 | 0.589 | 1.253 |
| **MCF7** | 0.978 | 0.861 | 1.108 | 0.562 | 2.296 |
| **NES rep 1** | 1.004 | 0.880 | 1.057 | 0.711 | 1.110 |
| **NES rep 2** | 1.015 | 0.887 | 1.058 | 0.717 | 1.102 |
| **NPC rep 1** | 1.007 | 0.892 | 1.063 | 0.664 | 1.090 |
| **NPC rep 2** | 1.011 | 0.883 | 1.068 | 0.642 | 1.091 |
| **NEU rep 1** | 1.002 | 0.887 | 1.060 | 0.657 | 1.097 |
| **NEU rep 2** | 0.996 | 0.867 | 1.066 | 0.640 | 1.103 |

DSB density across GSC, commercial cancer cell lines and neural cells lines calculated across chromosomes with normalisation for chromosome length (DSB/kbp) across chromosomes 1 to 22. DSB fold change calculated across overall DSB density per repeat (DSB/kbp). Median fold change, range and standard deviation across chromosomes 1 to 22 per cell line displayed. A supplementary table of DSB density per repeat has also been provided in supplementary material.

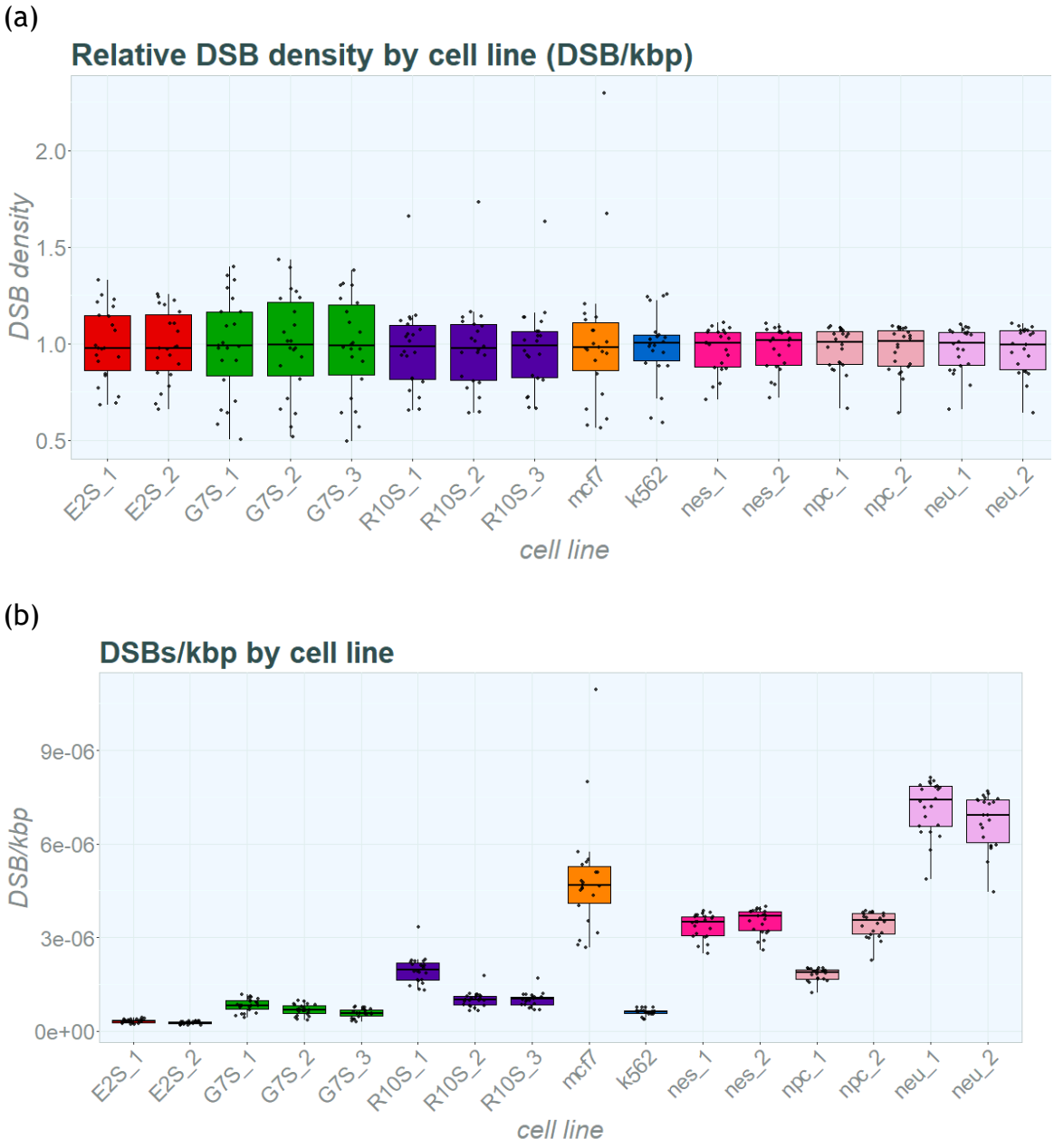



**Figure 3.6 DSB density fold change across chromosomes by cell line**

(a) Relative DSB density across chromosomes 1-22 per cell lines. Range, interquartile range and median represented on boxplots. Individual relative DSB density per chromosome overlaid as black dots. E2 GSCs represented in red, G7 GSCs represented in green, R10 GSCs represented in purple, MCF7 line represented in orange, K562 line represented in blue, NES neural lines represented in hot pink, NPC neural lines represented in pale pink, NEU lines represented in dark pink. (b) DSB/kbp per chromosomes 1-22 per cell line repeats overlaid as black dots on boxplots. GSCs have also been plotted separately in supplemental figures at DSBs/kbp level.

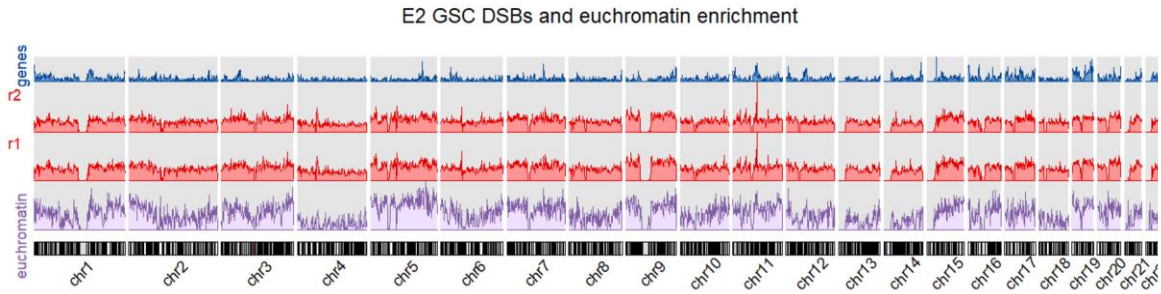
## 3.3.3 Chromatin mapping in GSCs

As previously discussed, chromatin state has key roles in cell expression, function and survival (Martí et al., 2021, Aleksandrov et al., 2020, Mack et al., 2019). Interestingly, chromatin state may also exhibit protective elements against induced DNA damage from IR (Brambilla et al., 2020), making this an appealing avenue to investigate when considering DSB location mapping in GSCs.

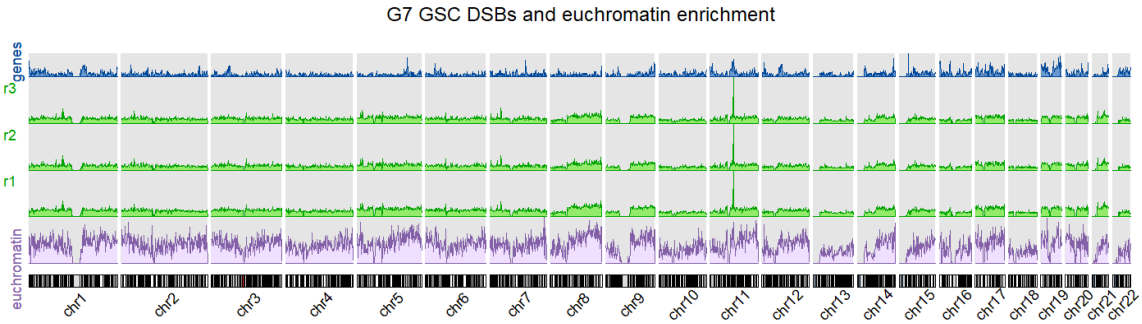
### 3.3.3.1 Euchromatin profiles and DSB density broadly vary with gene density

To visualise euchromatin enrichment profiles and DSB density, BLISS data and ATAC-seq datasets for euchromatin profiling were plotted together. Figure 3.7 demonstrates the distribution of DSBs measured by BLISS across GSCs alongside their respective euchromatin profiles. For each of the GSC lines, the repeats were presented together. DSB and euchromatin patterns partially follow gene site frequency, however this was not wholly the case. For example, chromosome 12 had a densely genic region which was not obviously reflected in DSB or euchromatin densities in any of the cell lines. Similarly, DSB densities did not appear to directly correspond to euchromatin enrichment. Whilst euchromatin should broadly be enriched at gene sites, there will be locations of genes that are in regions of heterochromatin due to methylation and heterochromatin marks. This may give the resultant pattern where highly methylated areas may be highly genic but not easily accessible due to epigenetic silencing. Regarding DSBs and euchromatin, there were some broadly shared patterns in DSB density and euchromatin enrichment. For example, E2 DSB density broadly dropped at chromosome 4 and 18, as did relative euchromatin across these chromosomes. In G7 there was a drop of euchromatin in 13 and 18 which was also reflected by DSB density at these chromosomes. Finally, chromosomes 10 and 18 had lower DSB densities in R10 and there was again a drop in euchromatin from baseline seen. As previously mentioned, R10 showed a higher DSB density in chromosome 7 than baseline. However, R10 euchromatin at chromosome 7 did not appear as highly distinct above baseline at an overview level. Across repeats, DSB densities remained highly consistent in E2, G7 and R10 lines.

(a)




(b)




(c)

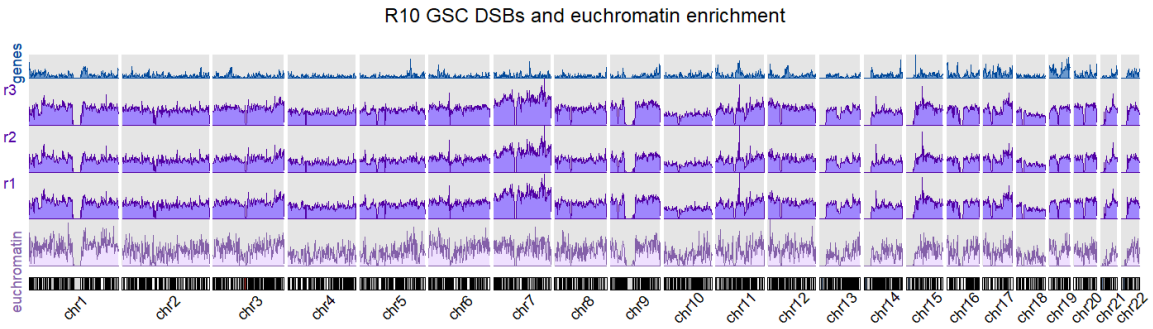



**Figure 3.7. DSB densities euchromatin enrichment across chromosomes 1-22 in GSCs E2, G7 and R10**

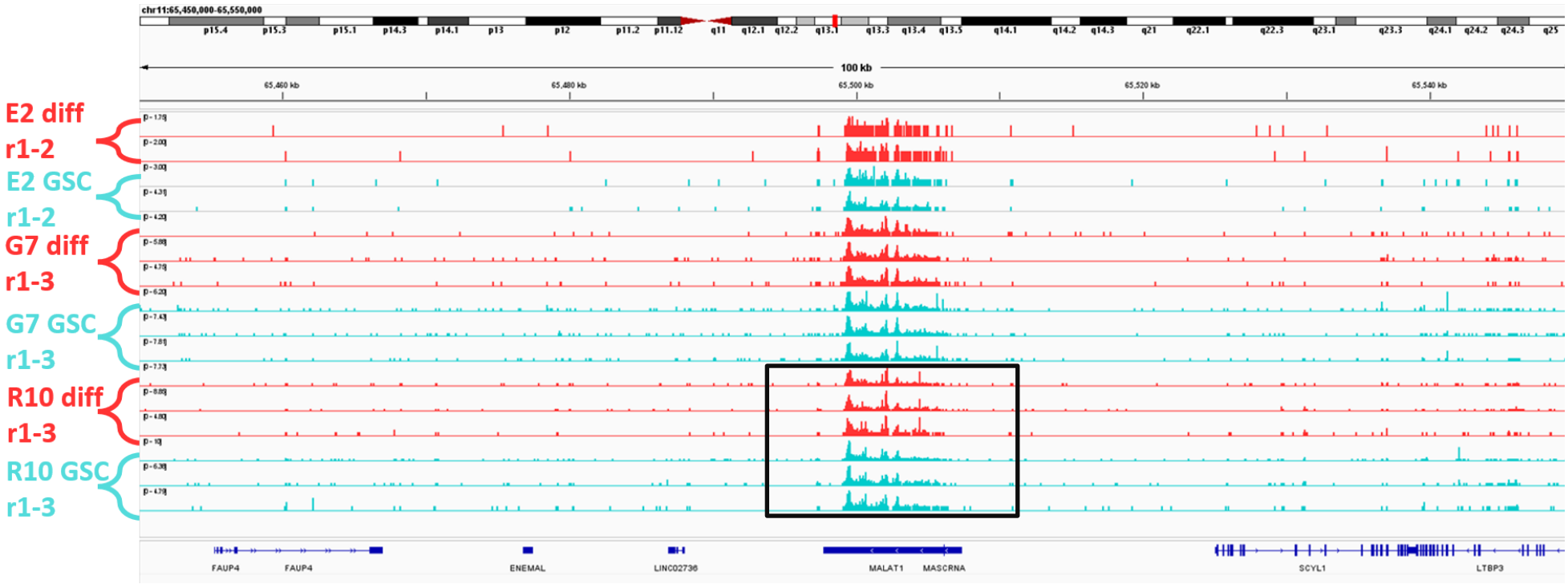
DSB density plots of GSCs in triplicate and euchromatin enrichment across chromosomes 1-22. Chromosomes displayed below as karyoplots from chr1-chr22. Euchromatin profiles displayed immediately above karyoplots in lilac. Individual GSC repeats of DSB density displayed above euchromatin enrichment profiles. Gene density profiles displayed on top line in blue for reference. (a) E2 GSCs repeats 1-2 (red). (b) G7 GSCs repeats 1-3 (green). (c) R10 GSCs repeats 1-3 (purple).

### 3.3.4 Frequently broken genomic regions are shared across GSCs and differentiated GBM populations

Regions with high DSB density were investigated to identify whether there were any shared features associated with high DSB density at these locations. To identify highly broken regions within cell lines, genomic bins were again used. In addition to mapping frequently broken GSC regions, DSB frequency was also

mapped in differentiated progeny cells. This was to identify whether there was variability in the regions with the highest DSB density in GSCs compared to differentiated cells that could account for changes in the DDR between them. 50 kbp regions were ranked from highest to lowest DSB density. Across all lines, the two regions with the highest DSB density were in 11q: bin regions chr11:65,500,000-65,550,000 and chr11:65,450,000-65,500,000 respectively. This was true across both GSCs and differentiated cells. This region demonstrated a consistent and conserved pattern of DSB density across GSC and differentiated lines (Figure 3.8). The DSB frequency was consistently concentrated on the long non-coding RNA (lncRNA) metastasis associated lung adenocarcinoma transcript 1 (*MALAT1/TALAM1*) gene region and showed similar distribution of DSB frequency across repeats and cell lines.

(a)


chr11:65,450,000-65,550,000
100 kb
65,460 kb
65,480 kb
65,500 kb
65,520 kb
65,540 kb
E2 diff
r1-2
E2 GSC
r1-2
G7 diff
r1-3
G7 GSC
r1-3
R10 diff
r1-3
R10 GSC
r1-3
FAUP4
FAUP4
ENEMAL
LINC02736
MALAT1
MASCRNA
SCYL1
LTBP3

(b)

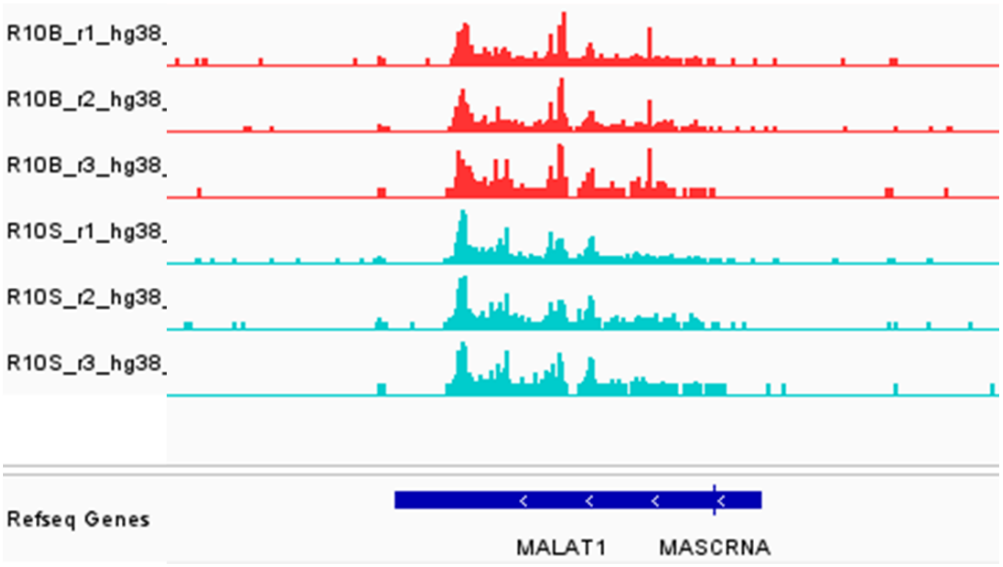



**Figure 3.8. Visualising DSB pattern across highest ranked bins for DSB frequency at chromosome 11q**

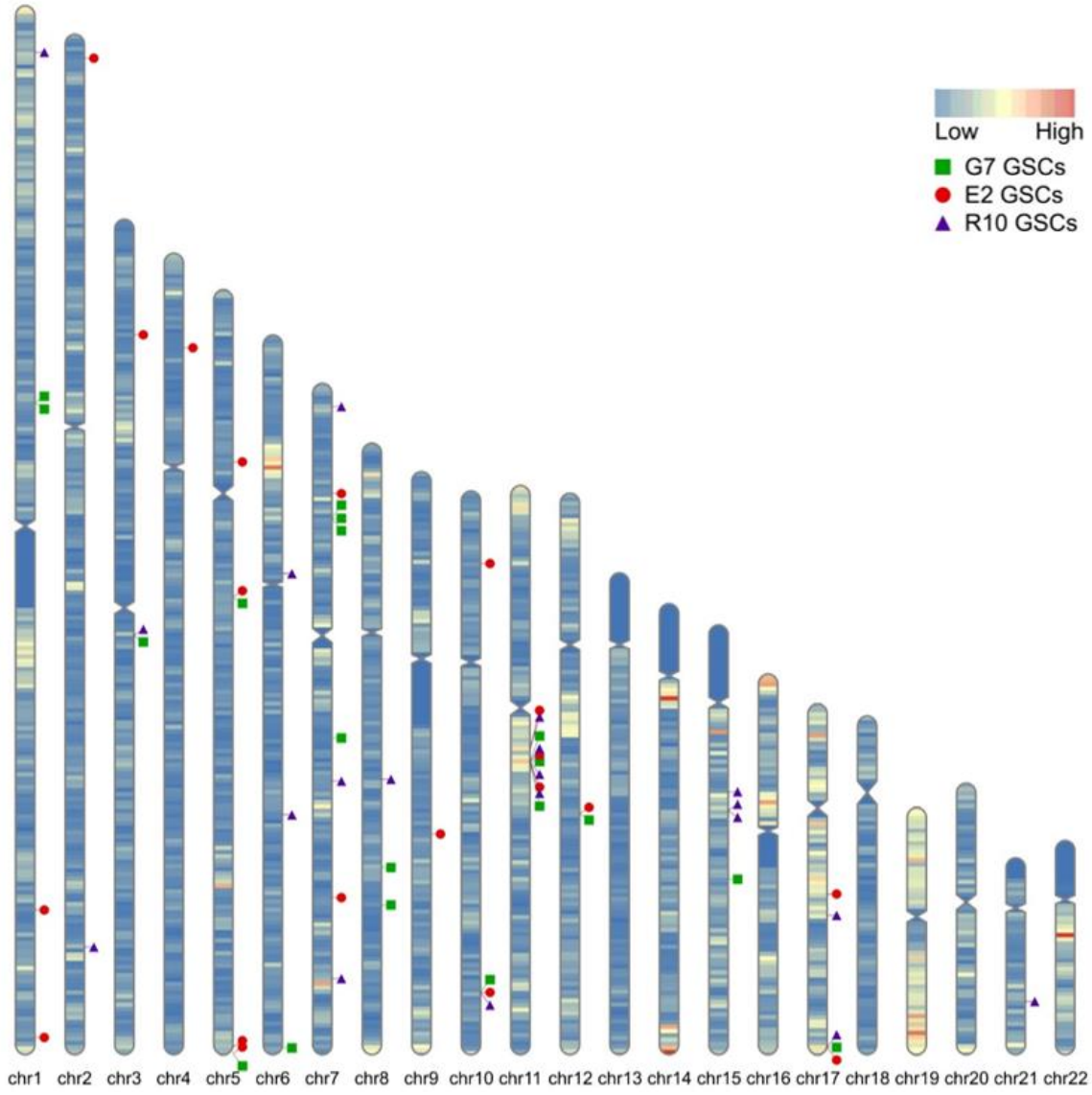
(a) DSB pattern plotted using integrated genome viewer (IGV) at chr11: 65,450,000-65,550,000 co-ordinates. DSB frequency represented across all GSC cell lines and differentiated cell lines and repeats. GSCs represented in turquoise, differentiated cells represented in red. Reference genes annotated in blue. From top to bottom: E2 differentiated cells repeats 1-2, E2 GSCs repeats 1-2, G7 differentiated cells repeats 1-3, G7 GSCs repeats 1-3, R10 differentiated cells repeats 1-3, R10 GSCs repeats 1-3. (b) Close up of R10 GBM DSB pattern across *MALAT1* gene.

The 20 regions with the highest DSB density were plotted across the genome for GSC lines E2, G7 and R10 (Figure 3.9 a). As previously noted, the region chr11: 65,450,000-65,550,000 was shared as a high DSB density site across E2, G7 and R10. Other sites that were shared across the three GSC lines were chr10: 119,000,000-119,050,000 and chr17: 81,500,000-81,550,000, both of which were in the top 20 regions with the highest DSB density.

The top 100 regions with the highest DSB density were also identified and the number of shared regions across GSC cell lines were plotted as a Venn diagram (Figure 3.9). Across the three cell lines there were 26 shared 50 kbp sites within the 100 regions with the highest DSB density. E2 and G7 GSCs shared over two thirds of the same highest DSB density regions. R10 shared 64 top 50 kbp sites with E2 and 62 top 50 kbp sites with G7. For GBM differentiated progeny, the 100 sites with the highest DSB density were identified. In both GSCs and differentiated progeny there was a majority overlap in the highest DSB density locations across the E2, G7 and R10 lines (Figure 3.9). E2 GSCs and

differentiated cells had the least number of shared regions with 66 50 kbp sites, whereas G7 and R10 had 80 and 83 shared sites respectively across GSC and differentiated cells. These findings demonstrated that there were several shared sites across cell lines and an association of DSBs occurring within the same genic regions across these.

(a)



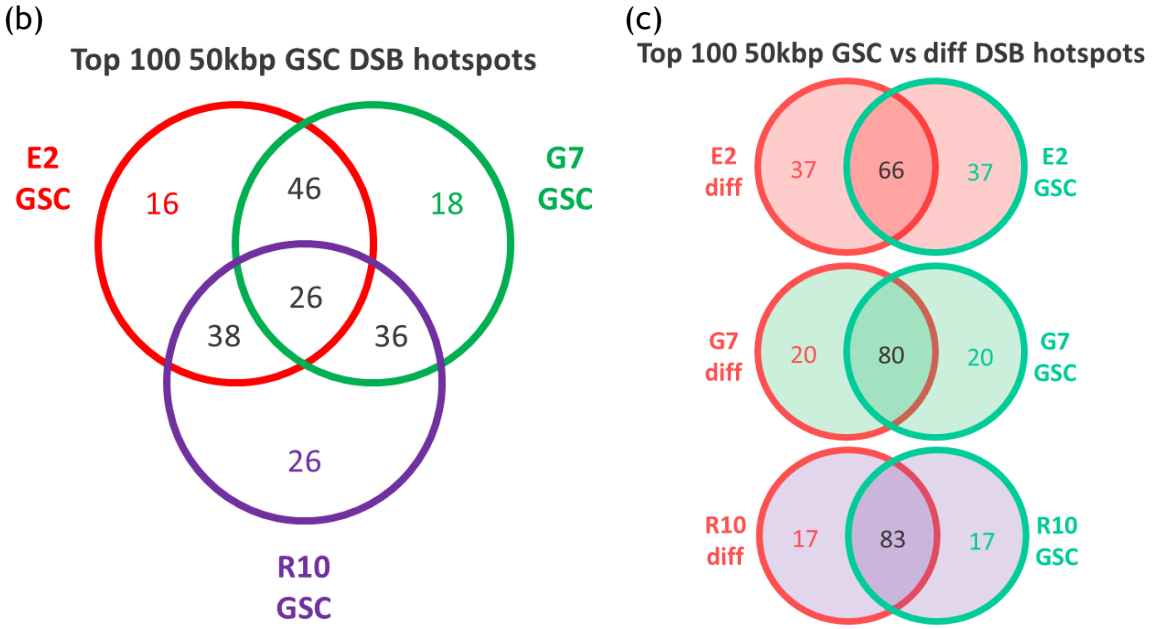



**Figure 3.9. Highest DSB density 50 kbp regions in E2, G7 and R10 cell lines**

For cell lines E2, G7 and R10, DSB density was calculated per 50 kbp region from highest to lowest. (a) The top 20 most frequently broken bins were extracted and locations plotted on an ideogram across E2, G7 and R10 GSCs to identify the location of the sites with the highest DSB density. E2 GSC top 20 regions annotated as red circles, G7 GSC top 20 regions annotated as green squares, R10 GSC top 20 regions annotated as purple triangles. (b.) Venn diagram of the top 100 regions of 50 kbp across E2, G7 and R10 GSC and their respective overlaps. (c.) Venn diagrams per cell line of top 100 regions and respective overlap sites between differentiated and GSC lines for E2, G7 and R10 respectively.

## 3.4 Discussion and conclusions

This chapter has given a broad overview of the endogenous DSB landscape in three primary GBM GSC populations. Patterns of DSBs at a whole genome level have also been observed in neural cells and commercial cancer cell lines MCF7 and K562. An overview of euchromatin enrichment for the corresponding GSCs has also been established and observed alongside DSBs. High DSB density sites have been identified and explored across GSCs and differentiated cells.

### 3.4.1 Patterns of DSBs in glioblastoma

Overall, DSB density across repeats demonstrated a consistency of distribution supporting the concept that endogenous DSBs do not occur randomly but rather are at identifiable regions across the genome. Most strikingly, GSC lines had a far greater variance in DSB density across chromosomes than their neural cell counterparts which showed very little in the way of differences across chromosomes. Additionally, DSB density across chromosomes varied, such as in

R10 GSCs where there was a broad increase in DSBs in chromosome 7 with a broad decrease in chromosome 10. This may be reflective of CNV regions where GBM tumours have been associated with copy number gain in chromosome 7 with concurrent copy number loss in chromosome 10 (Cohen et al., 2015). This chromosomal variation in DSB density was also notable in commercial cancer cell lines which also demonstrated a broader variation in relative DSB density across their respective breakomes compared to neural cells. This shared difference between GSCs and commercial cell lines may reflect the underlying nature of cancer genomes and their tendency towards genomic instability as one of the hallmarks of cancer (Hanahan, 2022, Negrini et al., 2010).

Across the three cell lines there were some identifiable similarities with consistently shared DSB regions such as the DSB clusters seen at chromosome 11q. Here there was a clear demonstrable peak of DSBs which was both maintained throughout repeats and across the three cell lines coinciding with the shared highest density DSB region.

## 3.4.2 DSBs in other cell types

### 3.4.2.1 Neural cells

The lineage of GSCs remains somewhat controversial, however they have been postulated to arise from NSCs (Ivanov and Hei, 2014). Whether GSCs originate from NSCs or represent other de-differentiated neural cell types, there are clearly important links to be explored. Identifying differences and shared features in DSB locations between our GSC populations with neural cells may help in identifying how mutated regions are affected by DDR mechanisms. Contrary to the GSC data, neural cells demonstrated a broadly consistent DSB frequency across chromosomes. There remained some sites of clear DSB peaks, however these were less frequent than in GSCs. NPC and NEU cells shared a DSB peak at chromosome 10 and 21, whilst NES cells appeared to generally show less in common with the other lines. This broadly similar DSB density across the neural cell breakome compared to GSCs may be indicative of the difference in GSCs and neural cells regarding genomic stability. Whilst neural cells are known to be at risk of RS, these cells remain otherwise genomically stable and functional. On the other hand, GSCs are well known to exhibit significant

genomic instability which may be reflected in the comparative variability of DSB density. Given that chromothripsis or genomic shattering occurs in an estimated 50% of GSCs, this may result in a far greater variability in DSB density across the GSC breakome (Cortés-Ciriano et al., 2020).

#### 3.4.2.2 Commercial cancer cell lines

In addition to neural cells, the two other cancer cell types, MCF7 and K562 were also profiled. Both demonstrated a more irregular baseline of DSBs compared to neural cells. Given that this was a shared attribute across both these and the GSC lines, there may be a greater variation in DSB frequency across cancer cell lines compared to non-cancer cell lines, although further analysis of both non cancer and cancer data would be required to determine this. Certainly, DSBs are widely described as deleterious, with the potential to result in genomic instability; a hallmark of cancer (Negrini et al., 2010). Curiously, both MCF7 and K562 lines also shared a peak of DSBs with GSCs at chromosome 11q, suggesting an area of interest for further investigation. Both cancer cell lines displayed large DSB peaks, unique to their cell lines. These are not within the scope of this thesis to investigate and further analysis for these cell lines is available in the paper by Ballinger et al (Ballinger et al., 2019). Like GSCs, DSB density varied considerably across chromosomes. Given these clear differences between the GSC and cancer cell lines in comparison to the neural cell lines, this was in keeping with the broad instability of the cancer genome.

### 3.4.3 Chromatin distribution and DSBs

There are relatively limited conclusions that can be taken from the mapping of euchromatin enrichment here. Euchromatin enrichment levels varied across chromosomes and cell types, likely reflecting the differences in transcriptional expression and the dynamic activity of chromatin activity within a cell. Broadly speaking, euchromatin regions were generally within genic regions which in general also appeared to have higher DSB density. However, there were also clear exceptions to this, indicating that euchromatin was not solely linked to gene location and vice versa. Euchromatin enrichment in the context of DSBs will be discussed further in later chapters where euchromatin sites were examined at a higher resolution.

### 3.4.4 Highly broken DSB regions

Having investigated the breakome at a broad genomic level, regions of high DSB frequency were also considered in more detail. As described, the most frequently broken region across all three cell lines was chr11: 65,450,000-65,550,000 and the region across the lncRNA gene *MALAT1/TALAM1* (Figure 3.8). This region accounted for the shared break peak in chromosome 11 and demonstrated a consistent DSB pattern which was present across cell lines and across GSCs and differentiated cells. The lncRNA *MALAT1* has been implicated widely in several cancers (Wang et al., 2018, Cervena et al., 2022). Certainly, *MALAT1* expression in GBM has been associated with poorer outcomes and survival (Chen et al., 2017). Though the role of *MALAT1* remains unclear, it is understood that *MALAT1* interacts with several small RNAs (snRNA: small nuclear RNAs, miRNA: microRNAs). The RNA MALAT1 can act as a competitive RNA and sequester smaller miRNAs, preventing them reaching the desired targets (Cai et al., 2018, Xu et al., 2021b). Additionally, the structure of the MALAT1 RNA transcript is known to be highly complex with a number of helices, pseudoknots, multiway junctions and a 3' triple helix, preventing rapid degradation (McCown et al., 2019). Furthermore, the RNA MALAT1 has been reported to contain G4 motifs which are also known to be associated with DSBs (Linke et al., 2021, Mou et al., 2022). Given the highly complex nature of the lncRNA secondary structure, it could be hypothesised that the DNA structure may also be prone to non-canonical secondary structures which may have some bearing on DSB frequency or distribution. Similarly, lncRNA are well known to interact with DNA and causing seconding structures such as R-Loops. This propensity for DSBs at this site may reflect R-Loop accumulation at these regions. Overall, the 50 kbp regions with the highest frequency of DSBs consistently showed DSBs concentrating within genic regions (Figure 3.8). This will be discussed in following chapters and highlights genic regions as important areas of investigation for DSBs. Finally, many of the highest DSB density 50 kbp regions were shared across E2, G7 and R10 GSC lines with 26 shared sites across the top 100 regions with the highest DSB density. Equally, there were several sites that were shared within two of the three cell lines indicating that some regions had a higher consistent DSB frequency in these three cell lines. Looking at the GSC and differentiated cells, there were several shared top DSB regions, with R10 GSCs and differentiated cells sharing 83 of the 100 sites, G7 sharing 80 and E2 sharing

66 of the top 100 highest DSB density sites. This may indicate that differences across GSC and differentiated DSB distribution may be relatively subtle and requires further analysis to determine whether imCharacterising DSBs in GBMportant distinguishing features exist within their respective breakomes.

### 3.4.5 Conclusions summary

- The breakome of GSCs is quantitatively and qualitatively different compared to neural cells, with greater variation in DSB density across chromosomes compared to neural cell lines.

- GSC DSB density indicates there are shared regions of DSBs within GSC lines, most notably at chromosome 11. DSB patterns within repeats of the same cell line are broadly very similar, indicating consistent regions of increased DSB frequency. DSB frequency across chromosomes between GSC cell lines demonstrate both conserved peaks that are shared as well as differential peaks between cell lines. GSC euchromatin enrichment varies across GSC lines and chromosomes indicating differences in the chromatin landscape between GSC lines.

- Like GSCs, cancer cell lines MCF7 and K562 share a peak of DSBs in chromosome 11 with GSC lines. Cancer cell lines also demonstrate a greater variation in DSB density across chromosomes compared to neural cell lines, potentially reflecting the genomic instability seen in cancer cells.

- The regions with the highest DSB density in GSC and differentiated cells indicate that DSBs are mainly concentrated at genic locations implying a potential role of gene-associated DSBs. Of the top 100 regions with the highest DSB density, 25% were shared across all three GSC lines. The highest DSB density regions in all three GSC lines were in chr11:65,500,000-65,550,000 and chr11:65,450,000-65,500,000

# Chapter 4 Exploring DSBs in genes and gene length

## 4.1 Introduction

The previous chapter discussed DSBs across the whole genome in addition to identifying regions of high DSB density in GBM cell lines. In the 50 kbp regions with the highest DSB density, DSBs were predominantly clustered around genic areas, making DSBs within genes an important area for further investigation.

### 4.1.1 Genes with high density DSBs

As the previous chapter indicated that genic regions may have an important role in endogenous DSB location, genic sites were taken forward for further study. Previous BLISS studies have shown that genes are important sites of interest for DSB locations in other cell lines (Wei et al., 2016, Yan et al., 2017). As previously discussed, *MALAT1* was highlighted by the 50 kbp regions as a site with a high DSB density. This also appeared to be the cause of DSB peaks at chromosome 11q. Chapter 3 also demonstrated a similar peak in chromosome 11q in cancer cell lines. The following chapter investigated DSB density in genes in GSCs and their differentiated progeny as well as in neural cell lines.

### 4.1.2 Gene length

Gene length has been postulated as important in neural cells with regards to DSB frequency (Wei et al., 2016). Wei et al (Wei et al., 2016) investigated gene length in neural cells as a contributor to DSB frequency, particularly in long neural genes. Long neural genes have been proposed to be at risk of DSBs due to the potential for RS due to replication-transcription conflicts. These long neural genes are known to take more than one cell cycle to transcribe, making transcription-replication conflicts inevitable (Helmrich et al., 2011). Long neural genes have also been proposed as a mechanism for the aberrantly upregulated DDR in GSCs (Carruthers et al., 2018). Given this hypothesis and the common shared features with neural cells it was important to determine whether gene length might also influence DSB frequency in our cells (Gimple et al., 2019).

Whilst gene length alone is clearly an important factor to consider, the potential increase in stochastic DSB occurrence purely related to a greater length of DNA should also be accounted for. Therefore, both absolute DSB frequency and DSB frequency adjusted for total gene length were used to investigate gene length. This chapter described the correlation of DSB frequency in relation to gene length in GSCs, with special interest in long neural genes.

### 4.1.3 Aims

This chapter identified the genes with the highest densities of DSBs in GSC lines and their differentiated progeny as well as DSB density in publicly available BLISS data for neural cell lines. The association of gene length with DSB frequency and density in GSCs and differentiated cells was explored.

- Identify the genes with the highest DSB density within GSCs and differentiated progeny cells and contrast with those genes with the highest DSB density in neural cell lines.

- Investigate the association of DSB frequency and density with gene length and in long neural genes in GSCs and differentiated cells.

## 4.2 Materials and Methods

### 4.2.1 Genes

To investigate DSB density in genes it was necessary to take length of gene into consideration given that DSB density could be influenced by absolute gene length where longer genes might have a greater random chance of harbouring a DSB compared to short genes. This was used as a means of comparing DSB frequencies across genes of different lengths. The DSB density was calculated per gene using the hg38 Ensembl database and list of genes (Martin et al., 2023). Intersecting DSBs with genes were calculated using RStudio to intersect genes with DSBs. In order to account for gene length, the DSB frequency per kbp of gene was calculated (DSBs/kbp):

*DSBs/kbp = (DSBs in gene/length of gene)/1000*

For further details see chapter 2 and RStudio chapter 4 appendix code. The mean density of DSBs per gene was calculated across repeats and genes were ranked from highest to lowest DSBs/kbp. The adjustment of DSBs per gene using gene length was at risk of overrepresenting DSBs in very small genes, therefore genes that were 200 bp or less were excluded from this analysis.

### 4.2.2 Gene length

For gene length, both absolute DSBs per gene and DSBs/kbp were used. Genes with no DSBs were excluded from analysis to remove genes that may have been unable to map by BLISS and short read sequencing. The absolute and adjusted DSB frequency per gene was mapped as a correlation plot and then as DSBs by gene length quartiles and finally as long neural genes and non-long neural genes. The long neural gene list was obtained from the genes identified by Wei et al (Wei et al., 2016) in their paper exploring long neural genes as sites of DSB clusters. The p-values displayed for Spearman rank correlation tests were highly significant, given that the number of datapoints provided was the total number of genes containing DSBs. Therefore p-values are provided as a means of comparing between gene length-adjusted and absolute values for correlation plots and should not necessarily confer biological significance. For gene length quartiles and long neural genes, DSB density had a non-Gaussian distribution and therefore medians per category were reported for statistical analysis. Medians per repeat were calculated. The mean of medians was used to calculate ANOVA results with a p-value of $<0.05$ deemed significant.

## 4.3 Results

### 4.3.1 Genes with high density DSBs

As described, genes in GSCs seemed to demonstrate a higher density of DSBs in GSCs than intergenic regions from 50 kbp bin results. Therefore, DSB density was investigated per gene in our GBM cell lines and also in neural cells and commercial cancer lines for a comparative overview.

#### 4.3.1.1 Genes with the highest DSB density are broadly shared across GSCs and differentiated progeny

The DSB density per gene was investigated in E2, G7 and R10 lines. The 10 genes with the highest DSB density in E2, G7 and R10 GSCs and differentiated progeny cells are displayed and discussed below (E2: **Error! Reference source not found.**, G7: **Error! Reference source not found.**, R10: **Error! Reference source not found.**).

The most frequently broken genes across all GBM cell lines were *MALAT1*/*TALAM1*, (**Error! Reference source not found.**, **Error! Reference source not found.**, **Error! Reference source not found.**) corresponding to the previous data demonstrating this region as a highly broken region. This was true in both GSC and differentiated lines. As previously discussed, *MALAT1* is a lncRNA which has been associated with cancer progression. This lncRNA gene locus is also spanned by the antisense *TALAM1* gene. Both *MALAT1* and *TALAM1* have been implicated in tumourigenesis, with *TALAM1* having synergistic and regulatory functions on *MALAT1* activity (Gomes et al., 2019). In addition to *MALAT1* and *TALAM1*, Poly(RC) Binding Protein 1 (*PCBP1*), was also represented in the 10 most frequently broken genes across all lines. *RMRP*, *ENSG00000270640* and *ENSG0000023496* (no Ensembl gene names available) were present in the 10 most frequently broken genes in both E2 GSCs and G7 GSC cell lines (**Error! Reference source not found.**, **Error! Reference source not found.**). For differentiated cells, E2, G7 and R10 all had *TMSB4X* and *NMNAT1P3* in the top 10 genes of highest DSB density.

Regarding GSCs and differentiated cells, many of the top 10 genes with the highest DSB density were shared across cell lines. For E2 GSC and differentiated cells, shared top genes were *TALAM1*, *MALAT1*, *NMNAT1P3*, *ENSG00000234961*, *PCBP1*, *ZNF619P1* and *RMRP* (RNA Component of Mitochondrial RNA Processing Endoribonuclease). Genes *ENSG00000270640* and *ENSG00000256804* were present within the top 10 genes with the highest DSB density in E2 GSCs. However, neither of these were present in the top 10 highest DSB density genes in E2 differentiated cells. Though not present within the top 10 highest DSB density genes, they were within the top 100 genes with the highest DSB density for differentiated cells. Only *SMC3P1* was present as a high DSB density gene in E2

GSCs and not also within the top 100 highest DSB density genes in differentiated cells. Regarding E2 differentiated cells, *ENSG00000276690, TMSB4X* and *H2AX* were all present within the top 10 highest DSB density genes. These genes were also present within the top 100 E2 GSC highest DSB density genes.

For G7, in addition to *MALAT1*/*TALAM1*, the lncRNA *NEAT1*, was also in the top 10 genes with the highest DSB density for GSCs and differentiated cells. The genes *RMRP*, *FOS* and *RPS12P16* were represented in the top 10 genes with the highest DSB density in GSCs. Though not within the top 10 genes with the highest DSB density for G7 differentiated cells, *RMRP*, *FOS* and *RPS12P16* were present within the top 100 genes in G7 differentiated cells with the highest DSB density. In differentiated G7 cells, *NMNAT1P3*, *TMSB4X* and *ID1* genes were in the top 10 highest DSB density genes. These three genes were also present in the top 100 highest DSB density genes for G7 GSC lines.

In R10, *RPL6P27*, *RCN1P2*, *PCBP2P2* and *TMSB4X* were all shared within the top 10 genes of highest DSB density in GSCs and differentiated cells. The R10 GSCs also had *SERBP1P5*, *TCEA1P2* and *NPM1P27* within the top 10 highest DSB density genes all of which were also in the top 100 highest DSB density genes in R10 differentiated cells. Similarly, *NMNAT1P3*, *ANXA2P2* and *YBX1P10* which were within the top 10 highest DSB density genes for R10 differentiated cells were also present within the 100 highest DSB density for R10 GSCs.

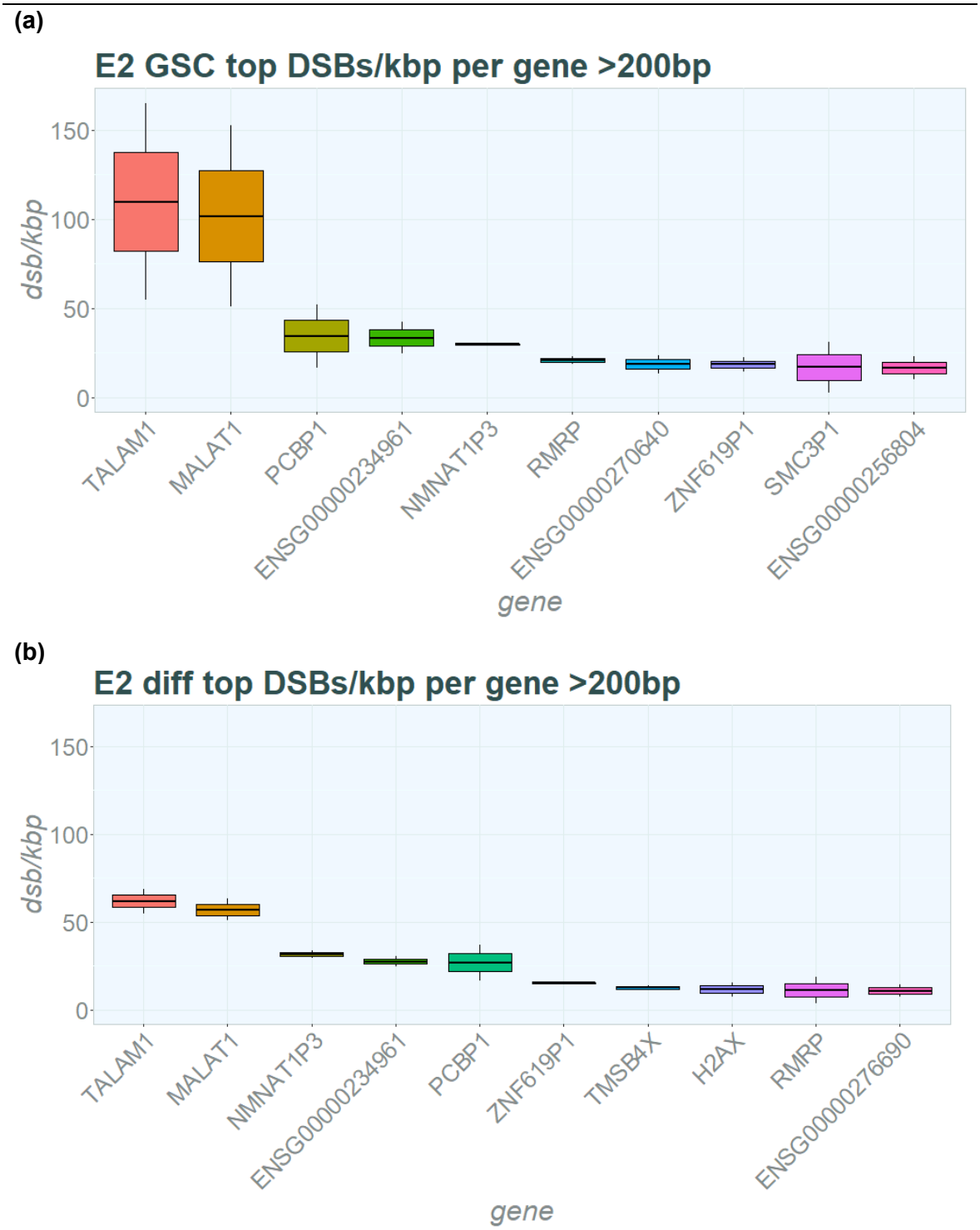
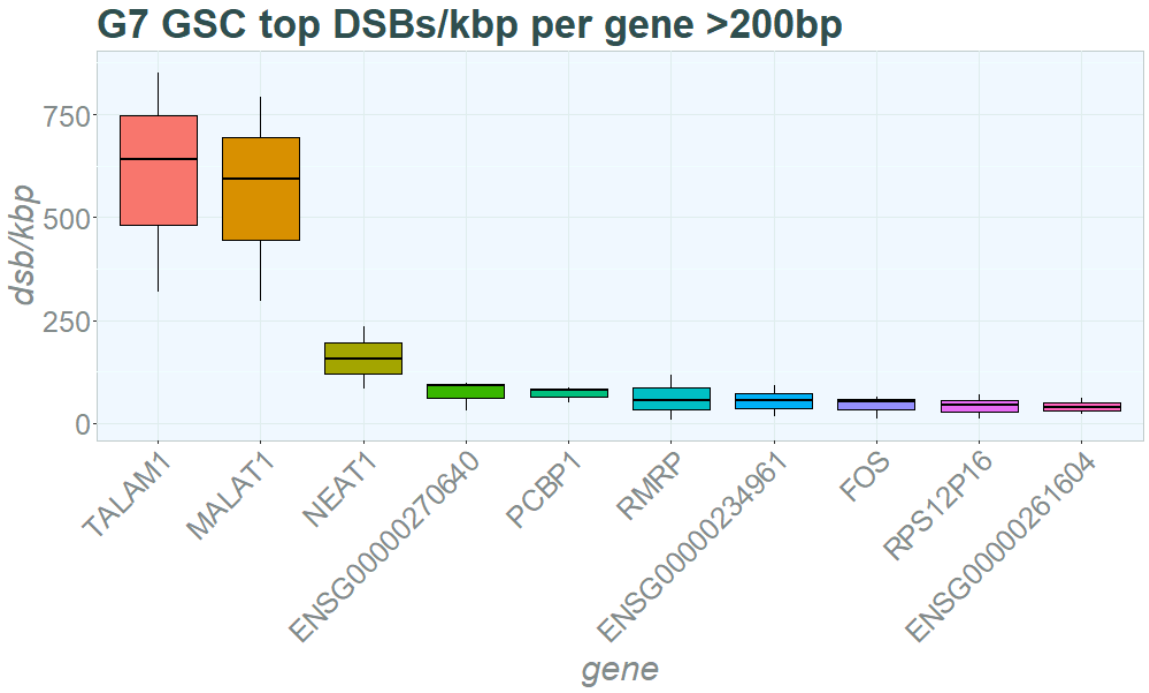



**Figure 4.1. E2 genes with the highest DSB density: top 10**

Top 10 genes with the highest DSB density in E2 GSCs and differentiated cells. Order of DSB frequency determined by ordered mean DSB frequency per gene after adjusting for gene length. Gene names presented as symbols unless no Ensembl gene name was available. Genes with no Ensembl gene name were called by Ensembl code instead. (a) E2 GSC top broken genes adjusted for gene length. (b) E2 differentiated cells top broken genes adjusted for gene length.

(a)




(b)

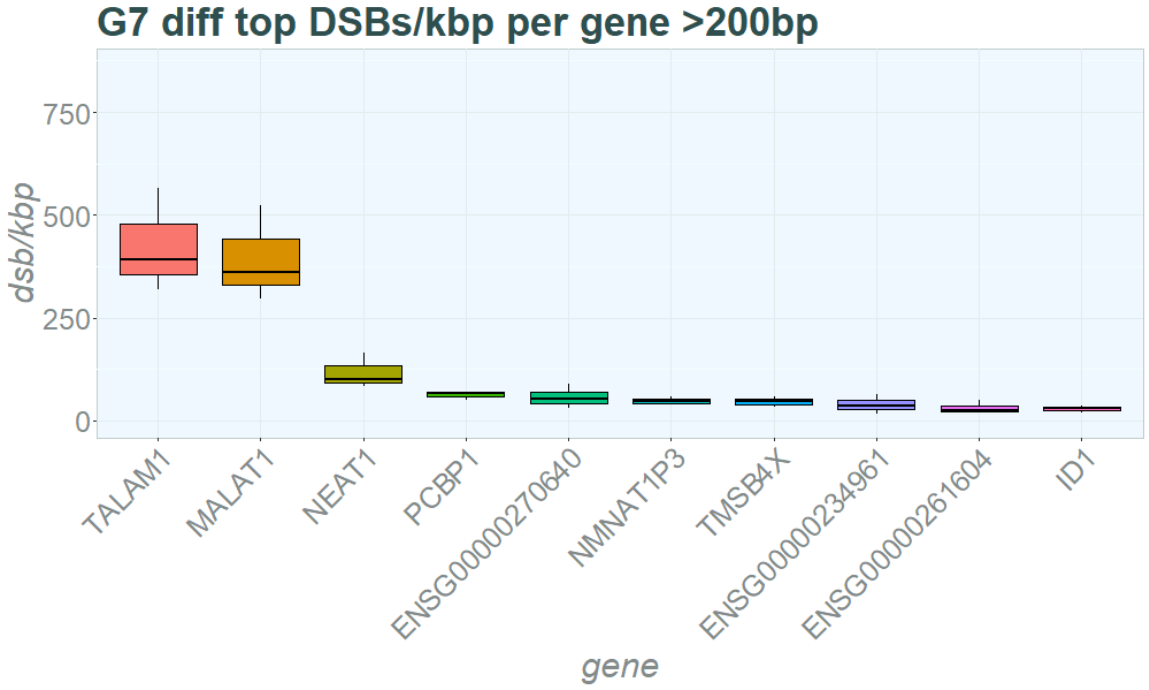
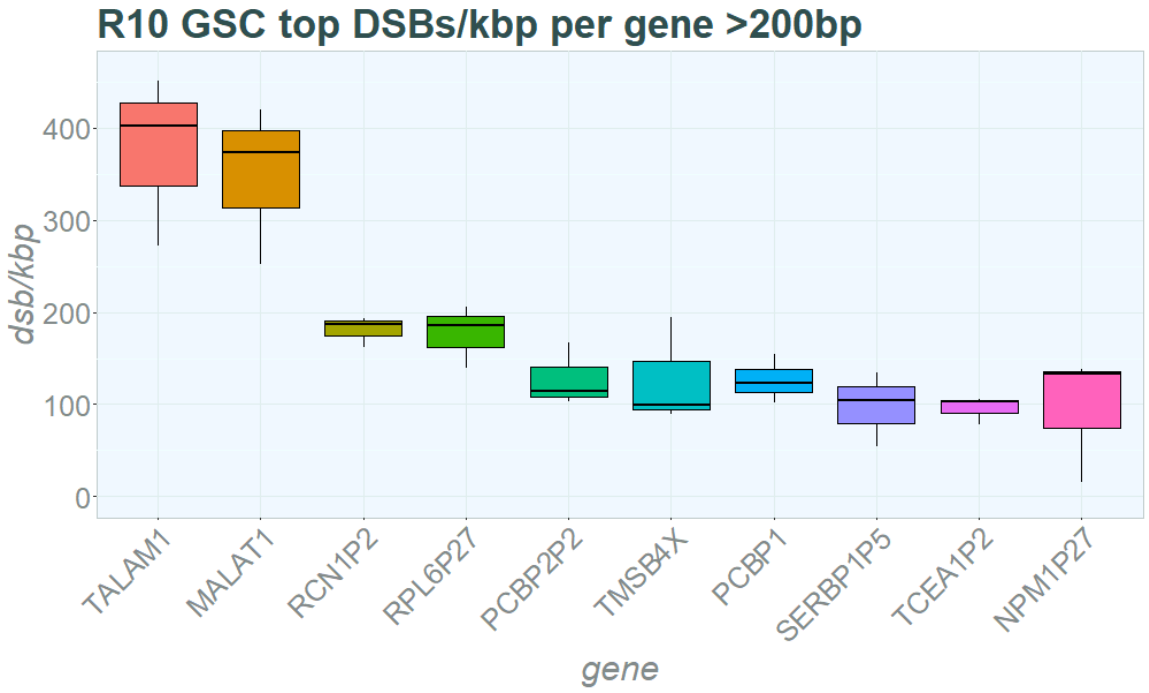



**Figure 4.2. G7 genes with the highest DSB density: top 10**

Top 10 genes with the highest DSB density in G7 GSCs and differentiated cells. Order of DSB density determined by ordered mean DSB density per gene after adjusting for gene length. Gene names presented as symbols unless no Ensembl gene name was available. Genes with no Ensembl gene name were called by Ensembl code instead. (a) G7 GSC top broken genes adjusted for gene length. (b) G7 differentiated cells top broken genes adjusted for gene length.

(a)




(b)

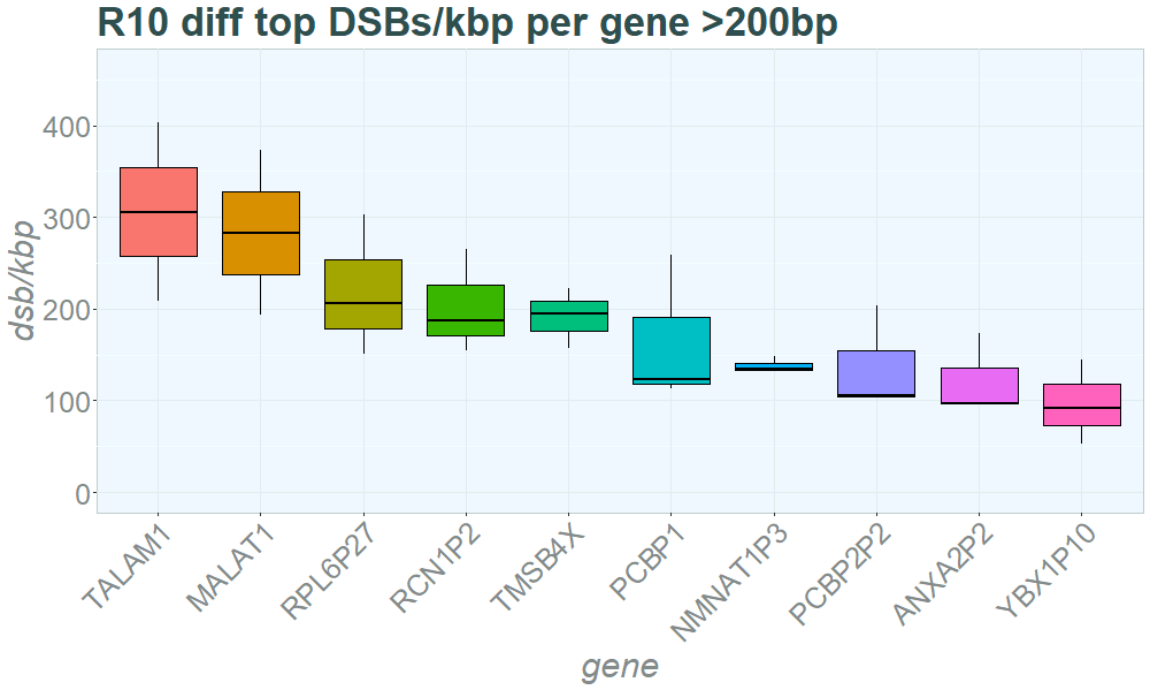



**Figure 4.3. R10 genes with the highest DSB density: top 10**

Top 10 genes with the highest DSB density in R10 GSCs and differentiated cells. Order of DSB density determined by ordered mean DSB density per gene after adjusting for gene length. Gene names presented as symbols. (a) R10 GSC top broken genes adjusted for gene length. (b) R10 differentiated cells top broken genes adjusted for gene length.

#### 4.3.1.2 Genes with high DSB density in neural cells share *MALAT1* as a high DSB density gene

In NPC and NEU lines, *MALAT1/TALAM1* sites were also represented in the top 10 genes with the highest DSB density but not in NES. Whilst *MALAT1* was not within the top 10 genes for highest DSB density in NES lines, it was within the top 100 genes. The neural cells also all had *RMRP* in the top 10 genes with highest density DSBs, as did E2 GSCs and differentiated cells and G7 GSCs. The gene *RMRP* is also a lncRNA and has been associated with cell-cycle progression (Vakkilainen et al., 2019).

Four small nucleolar RNA (snoRNA) were present within the genes with the highest DSB density in the neural cell lines, three of which were small Cajal body associated RNAs (scaRNAs: *SCARNA2*, *SCARNA13*, *SCARNA7*). These genes transcribe to non-coding RNAs and act in a number of roles including alternative splicing. Neural lines also shared snoRNA *SNORD17* as one of the top 10 genes with highest density DSBs across NES, NPC and NEU lines.

Several mitochondrial-related pseudogenes were also included across the top 10 genes with the highest density DSBs in NES and NPC lines. These included pseudogenes *MTND2P28, MTND1P23, MTATP6P1, MTCO2P12* and *MTCO3P12* in NES and *MTCO2P1* in NPC. Additionally, NPC also had the pseudogenes *HMGB3P32*, *ZNF619P1* and *RPS24P12* within the top ten genes represented.

There were relatively few overlapping genes within the top 100 genes with the highest density DSBs across the neural cell lines with only 9 shared genes across NES, NEU and NPC lines including *MALAT1/TALAM1*, *RMRP* and *SNORD17* as previously mentioned. Other shared genes were *SCARNA13*, *CRIM1-DT*, *RPPH1*, *Metazoa_SRP* and *BNIP3P41*. This was in contrast to GSC GBM lines which shared 62 genes with each other all within the top 100 genes with the highest density DSBs, though many of these only held Ensembl IDs and not symbol names. Only *MALAT1/TALAM1* was shared across all neural cells and GSCs within the top 100 genes with the highest DSB density.

Overall, these results indicated that a number of non-protein coding RNA genes were represented in the genes with the highest DSB density in neural cells, particularly in the least differentiated cell line NES. However, few genes were

shared between neural cell lines which was in contrast to GSCs which had many shared high density DSB genes across GSC and differentiated progeny as well as between GBM lines.

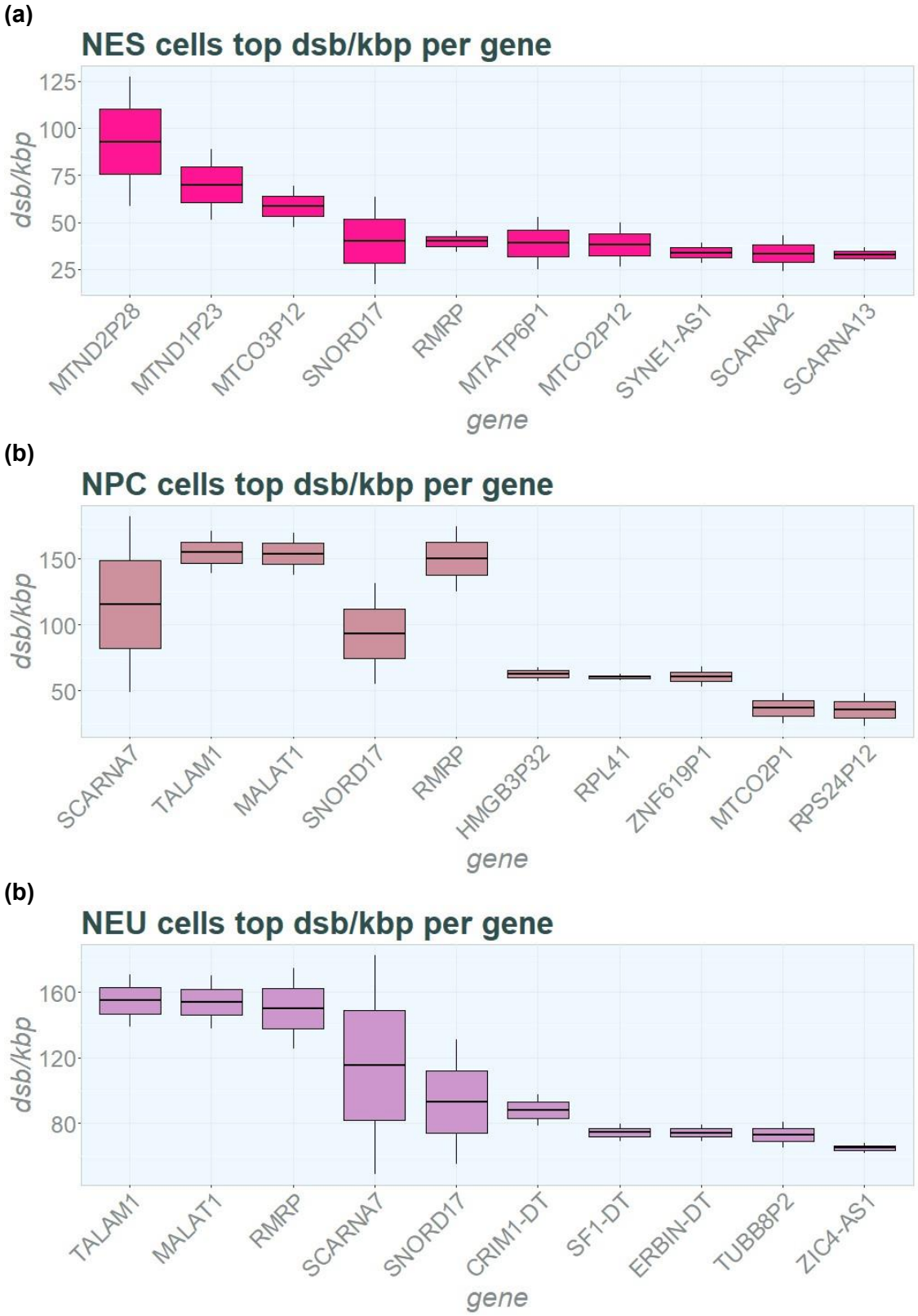



**Figure 4.4 Neural cell line genes with the highest DSB density: top 10**

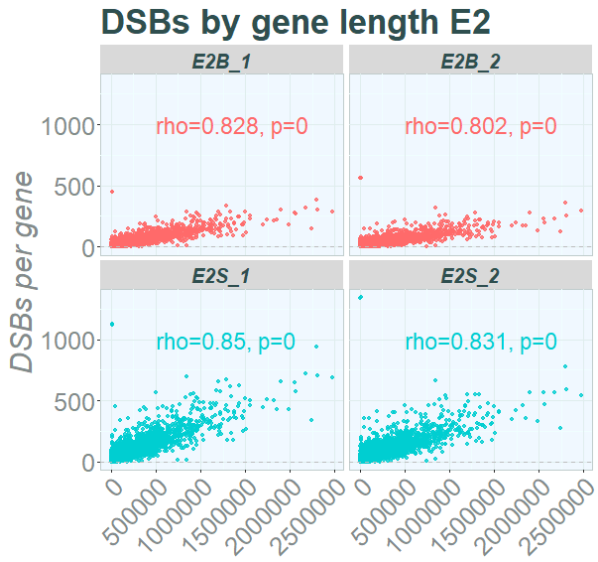
Top 10 genes with the highest DSB density in neural cells; NES: neuroepithelial stem cells, NPC: neural progenitor cells, NEU: post-mitotic neural cells. Order of DSB density determined by ordered mean DSB density per gene after adjusting for gene length. Gene names presented as symbols. (a) NES top broken genes adjusted for gene length. (b) NPC top broken genes adjusted for gene length. (c) NEU top broken genes adjusted for gene length.

### 4.3.2 DSB frequency increases with increasing gene length but DSB density does not in GBM lines

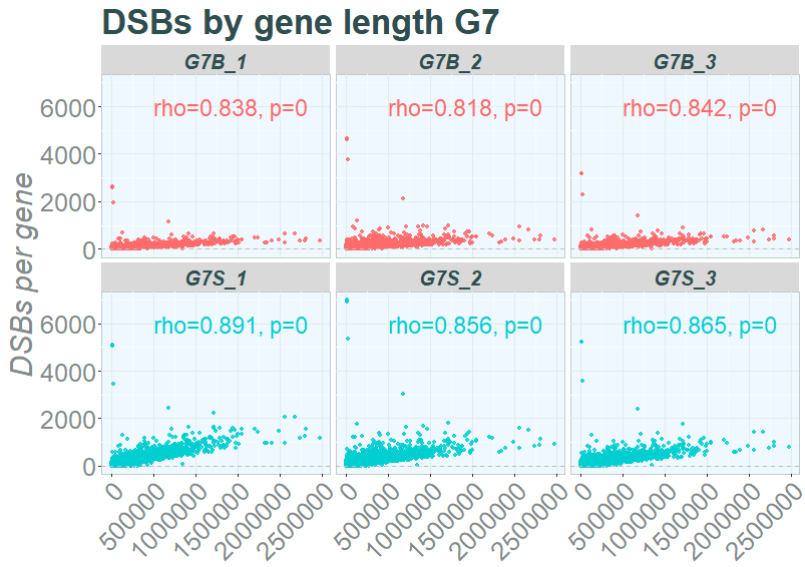
Long genes have been highlighted as sites prone to DNA damage and DSBs (Wei et al., 2016). Long neural genes are included in this group and have been identified as having an increased risk of breakage. These are often highly transcribed in neural cells and have been linked to a variety of disease processes. Both absolute DSB frequency and DSB density adjusted to total gene length was used to investigate gene length.

When absolute DSB frequency per gene was plotted against gene length, there was a consistent positive correlation (Figure 4.5). This was consistent across both GSCs and differentiated lines with repeats demonstrating positive rho values of greater than 0.8 (range: >0.81 - <0.92). GSCs demonstrated a trend towards moderately higher positive correlation compared to differentiated progeny though t-test of GSC vs differentiated rho values was non-significant (G7 lines: p=0.085).

**(a)**


DSBs by gene length E2
E2B_1
E2B_2
E2S_1
E2S_2
rho=0.828, p=0
rho=0.802, p=0
rho=0.85, p=0
rho=0.831, p=0
DSBs per gene
0
500
1000
0
500000
1000000
1500000
2000000
2500000

**(b)**


DSBs by gene length G7
G7B_1
G7B_2
G7B_3
G7S_1
G7S_2
G7S_3
rho=0.838, p=0
rho=0.818, p=0
rho=0.842, p=0
rho=0.891, p=0
rho=0.856, p=0
rho=0.865, p=0
DSBs per gene
0
2000
4000
6000
0
500000
1000000
1500000
2000000
2500000

**(c)**

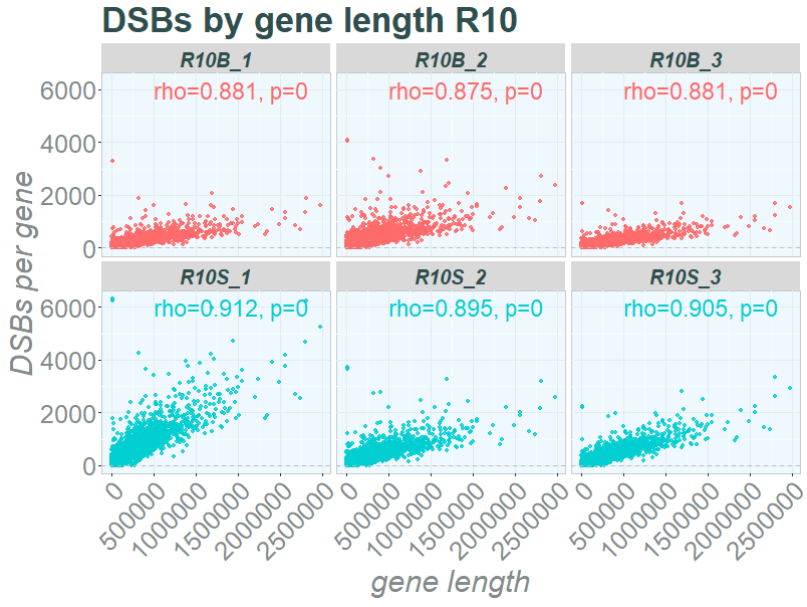
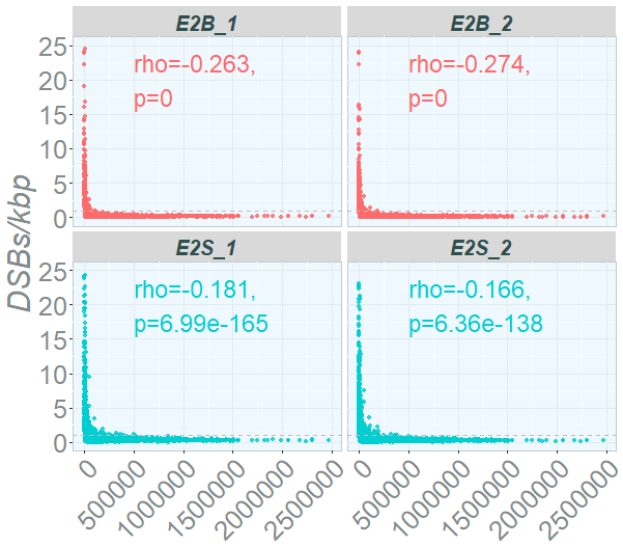

DSBs by gene length R10
R10B_1
R10B_2
R10B_3
R10S_1
R10S_2
R10S_3
rho=0.881, p=0
rho=0.875, p=0
rho=0.881, p=0
rho=0.912, p=0
rho=0.895, p=0
rho=0.905, p=0
DSBs per gene
gene length
0
2000
4000
6000
0
500000
1000000
1500000
2000000
2500000

**Figure 4.5. Total DSBs per gene by gene length GBM cells**

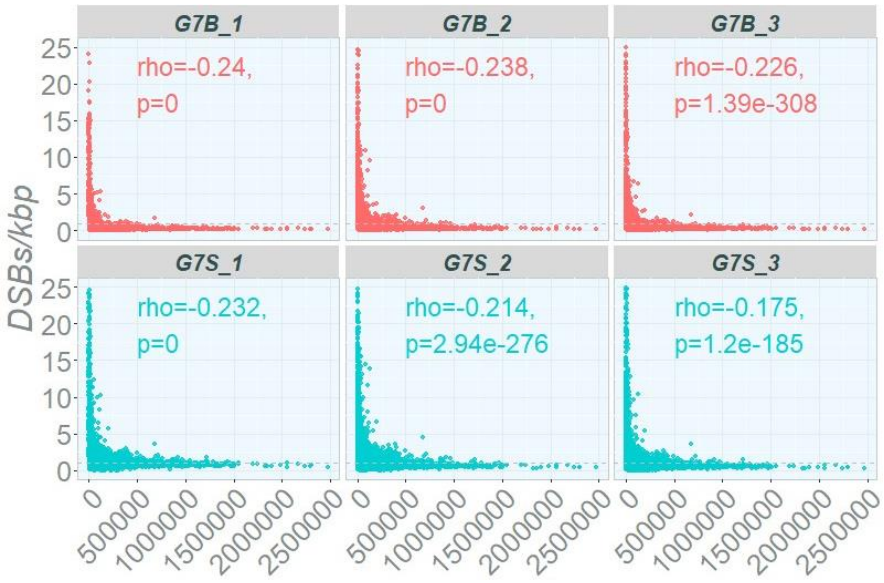
Total DSBs per gene by gene length. X-axis: displaying gene length, y-axis displaying total DSBs per gene. Differentiated cells displayed in red. GSCs displayed in turquoise. Spearman rank correlation tests performed per repeat. Rho and p values per repeat displayed. (a) E2 differentiated cells and GSCs, repeats 1 and 2. Differentiated cells denoted as follows; E2B_1: repeat 1, E2B_2: repeat 2. GSCs denoted as follows; E2S_1: repeat 1, E2S_2: repeat 2. (b) G7 differentiated cells and GSCs repeats 1-3. Differentiated cells denoted as follows; G7B_1: repeat 1, G7B_2: repeat 2, G7B_3: repeat 3. GSCs denoted as follows; G7S_1: repeat 1, G7S_2: repeat 2, G7S_3: repeat 3. (c) R10 differentiated cells repeats 1-3. Differentiated cells denoted as follows; R10B_1: repeat 1, R10B_2: repeat 2, R10B_3: repeat 3. GSCs denoted as follows; R10S_1: repeat 1, R10S_2: repeat 2, R10S_3: repeat 3.

However given that longer genes may be more prone to DSBs due to increasing length alone, gene breaks were also adjusted for gene length as previously described. Once DSB frequency was adjusted for length (DSBs/kbp), the pattern of increasing DSBs in accordance with gene length was not maintained across lines (Figure 4.6). Non-parametric correlation demonstrated a weakly negative correlation (range: <-0.1 - >-0.3) after adjustment for gene length in E2, G7 and R10 lines.

**(a)**


E2B_1
rho=-0.263,
p=0
E2B_2
rho=-0.274,
p=0
E2S_1
rho=-0.181,
p=6.99e-165
E2S_2
rho=-0.166,
p=6.36e-138
DSBs/kbp
0
5
10
15
20
25
0
500000
1000000
1500000
2000000
2500000

**(b)**


G7B_1
rho=-0.24,
p=0
G7B_2
rho=-0.238,
p=0
G7B_3
rho=-0.226,
p=1.39e-308
G7S_1
rho=-0.232,
p=0
G7S_2
rho=-0.214,
p=2.94e-276
G7S_3
rho=-0.175,
p=1.2e-185
DSBs/kbp
0
5
10
15
20
25
0
500000
1000000
1500000
2000000
2500000

**(c)**

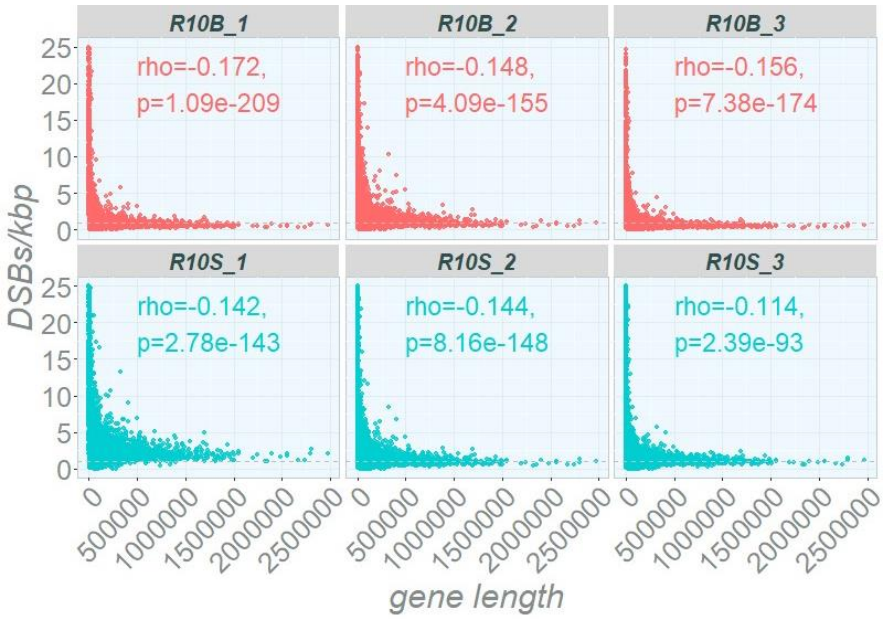
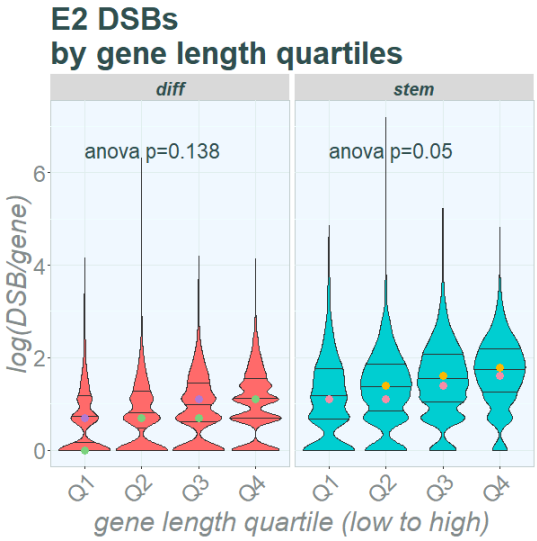

R10B_1
rho=-0.172,
p=1.09e-209
R10B_2
rho=-0.148,
p=4.09e-155
R10B_3
rho=-0.156,
p=7.38e-174
R10S_1
rho=-0.142,
p=2.78e-143
R10S_2
rho=-0.144,
p=8.16e-148
R10S_3
rho=-0.114,
p=2.39e-93
DSBs/kbp
0
5
10
15
20
25
0
500000
1000000
1500000
2000000
2500000
gene length

**Figure 4.6. Gene DSB density adjusted to gene length (DSBs/kbp) in GBM cells**

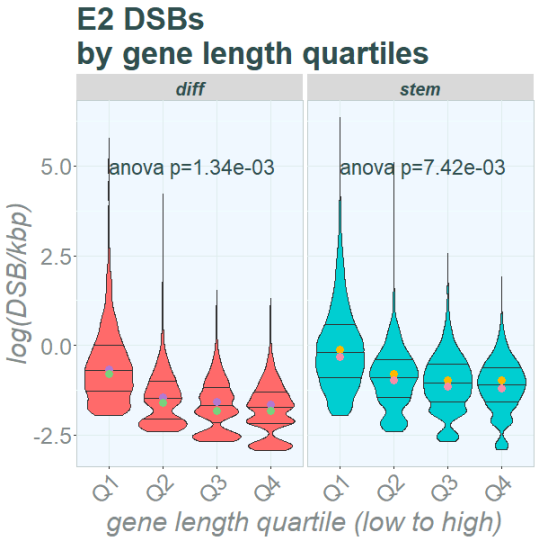
Adjusted DSBs per gene by gene length (DSB/kbp). X-axis: displaying gene length, y-axis displaying adjusted gene DSBs as DSBs/kbp. Differentiated cells displayed in red. GSCs displayed in turquoise. Spearman rank correlation tests performed per repeat. Rho and p values per repeat displayed. (a) E2 differentiated cells and GSCs, repeats 1 and 2. Differentiated cells denoted as follows; E2B_1: repeat 1, E2B_2: repeat 2. GSCs denoted as follows; E2S_1: repeat 1, E2S_2: repeat 2. (b) G7 differentiated cells and GSCs repeats 1-3. Differentiated cells denoted as follows; G7B_1: repeat 1, G7B_2: repeat 2, G7B_3: repeat 3. GSCs denoted as follows; G7S_1: repeat 1, G7S_2: repeat 2, G7S_3: repeat 3. (c)R10 differentiated cells repeats 1-3. Differentiated cells denoted as follows; R10B_1: repeat 1, R10B_2: repeat 2, R10B_3: repeat 3. GSCs denoted as follows; R10S_1: repeat 1, R10S_2: repeat 2, R10S_3: repeat 3.

Given that a number of short genes with multiple DSBs could also influence DSB density after adjustment, gene length was also investigated by quartiles. Genes were divided into quartiles by gene length and DSB frequencies plotted with and without adjustment (Figure 4.7). In quantifying absolute DSBs per gene, gene length demonstrated a trend towards an increase in DSBs in the longest gene quartiles. After adjustment for gene length (DSBs/kbp), this pattern was reversed with a reduction of DSB density in the longest genes. This was consistent across both GSCs and differentiated cell types in E2, G7 and R10 lines. Across cell lines after gene length adjustment, testing of medians displayed significance, indicating a negative correlation of DSB density with gene length.

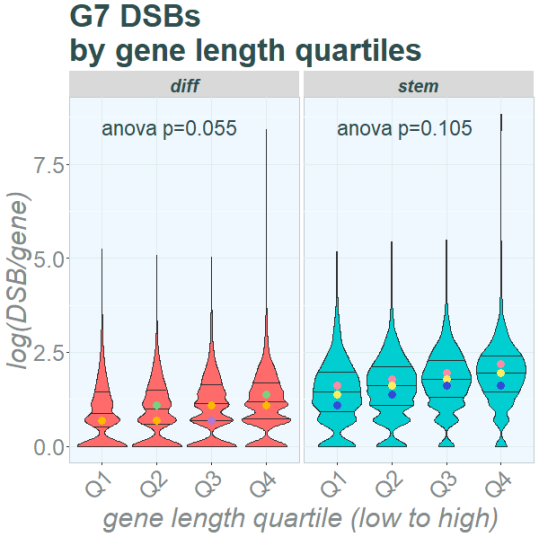
(a)


E2 DSBs
by gene length quartiles
diff
stem
anova p=0.138
anova p=0.05
log(DSB/gene)
0
2
4
6
Q1
Q2
Q3
Q4
gene length quartile (low to high)

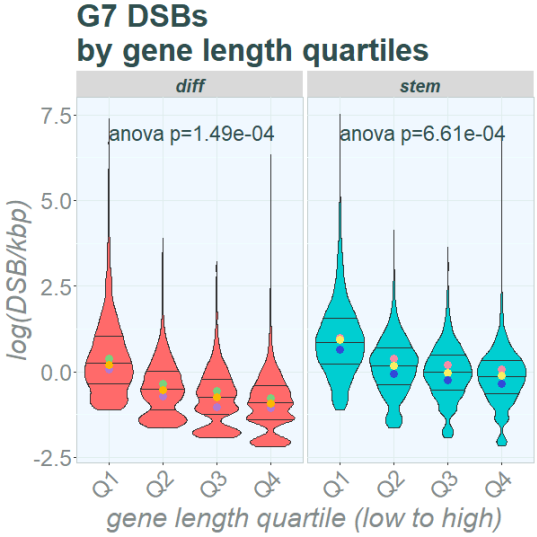
(b)


E2 DSBs
by gene length quartiles
diff
stem
anova p=1.34e-03
anova p=7.42e-03
log(DSB/kbp)
-2.5
0.0
2.5
5.0
Q1
Q2
Q3
Q4
gene length quartile (low to high)

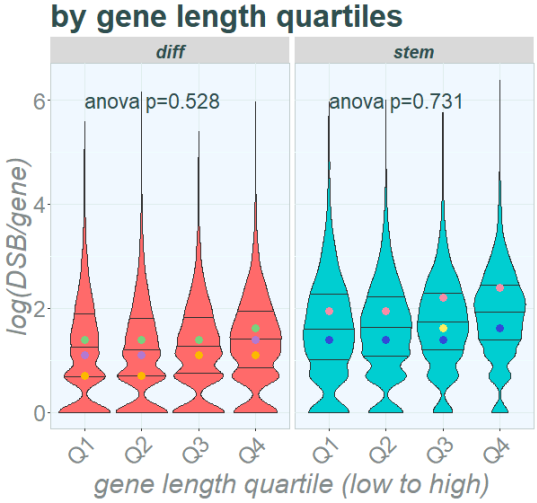
(c)


G7 DSBs
by gene length quartiles
diff
stem
anova p=0.055
anova p=0.105
log(DSB/gene)
0.0
2.5
5.0
7.5
Q1
Q2
Q3
Q4
gene length quartile (low to high)

(d)


G7 DSBs
by gene length quartiles
diff
stem
anova p=1.49e-04
anova p=6.61e-04
log(DSB/kbp)
-2.5
0.0
2.5
5.0
7.5
Q1
Q2
Q3
Q4
gene length quartile (low to high)

(e)


R10 DSBs
by gene length quartiles
diff
stem
anova p=0.528
anova p=0.731
log(DSB/gene)
0
2
4
6
Q1
Q2
Q3
Q4
gene length quartile (low to high)

(f)

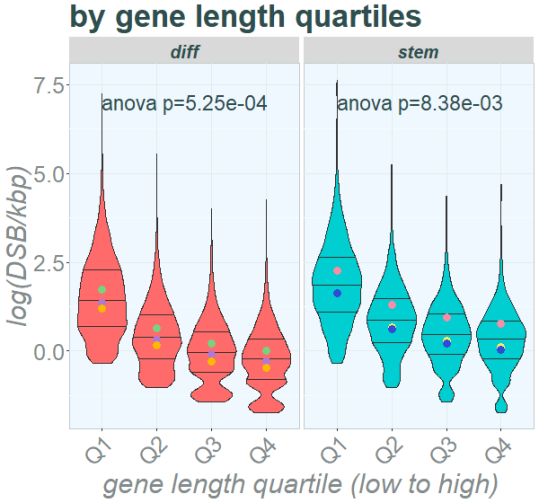

R10 DSBs
by gene length quartiles
diff
stem
anova p=5.25e-04
anova p=8.38e-03
log(DSB/kbp)
0.0
2.5
5.0
7.5
Q1
Q2
Q3
Q4
gene length quartile (low to high)

**Figure 4.7: Absolute DSB frequency and DSB density for GBM cells by gene length quartiles**

Total DSBs (DSBs/gene) and gene length-adjusted DSBs (DSBs/kbp) by gene length quartiles. Genes with DSBs were divided into quartiles and medians from DSBs per repeat calculated. Violin plot shows all genes across repeats with interquartile ranges displayed as horizontal lines. Medians represented by overlayed dots. X-axis displays gene length quartiles in order of shortest to longest (Q1: shortest genes, Q4: longest genes). Y- axis displays DSB frequency per gene. Y values log-transformed for visualisation. Differentiated cells ("diff") displayed in red. GSCs ("stem") displayed in turquoise. Analysis of variance (ANOVA) performed across medians of repeats per each group: E2 GSCs, E2 differentiated cells, G7 GSCs, G7 differentiated cells, R10 GSCs and R10 differentiated cells. Results for total DSBs displayed on the left, results for DSBs adjusted for gene length displayed on the right. (a) E2 GSCs and differentiated cells total DSBs by gene length quartile. (b) E2 GSCs and differentiated cells adjusted gene DSBs by gene length quartile. (c) G7 GSCs and differentiated cells total DSBs by gene length quartile. (d) G7 GSCs and differentiated cells adjusted gene DSBs by gene length quartile. (e) R10 GSCs and differentiated cells total DSBs by gene length quartile. (f) R10 GSCs and differentiated cells adjusted gene DSBs by gene length quartile.

Genes belonging to the "long neural gene" group outlined by Wei et al. (Wei et al., 2016) were also specifically investigated to assess DSB frequency in long neural genes alone. This allowed for the possibility that, whilst DSB density did not appear to correlate positively with gene length, long neural genes could be uniquely prone to DSBs, separate to other longer genes. Absolute DSBs were mapped in long neural genes and across non-long neural genes ("other gene sites"), showing a significant increase of DSBs in long neural genes in E2 differentiated cells ($p=0.046$), G7 differentiated cells ($p=0.011$) and GSCs ($p=0.011$) and R10 differentiated cells ($p=0.023$). Overall, all cells showed a trend towards an increase in DSBs in long neural genes compared to other gene sites. However, when both groups were adjusted for gene length, there was not a detectable significant difference between long neural genes and other groups.

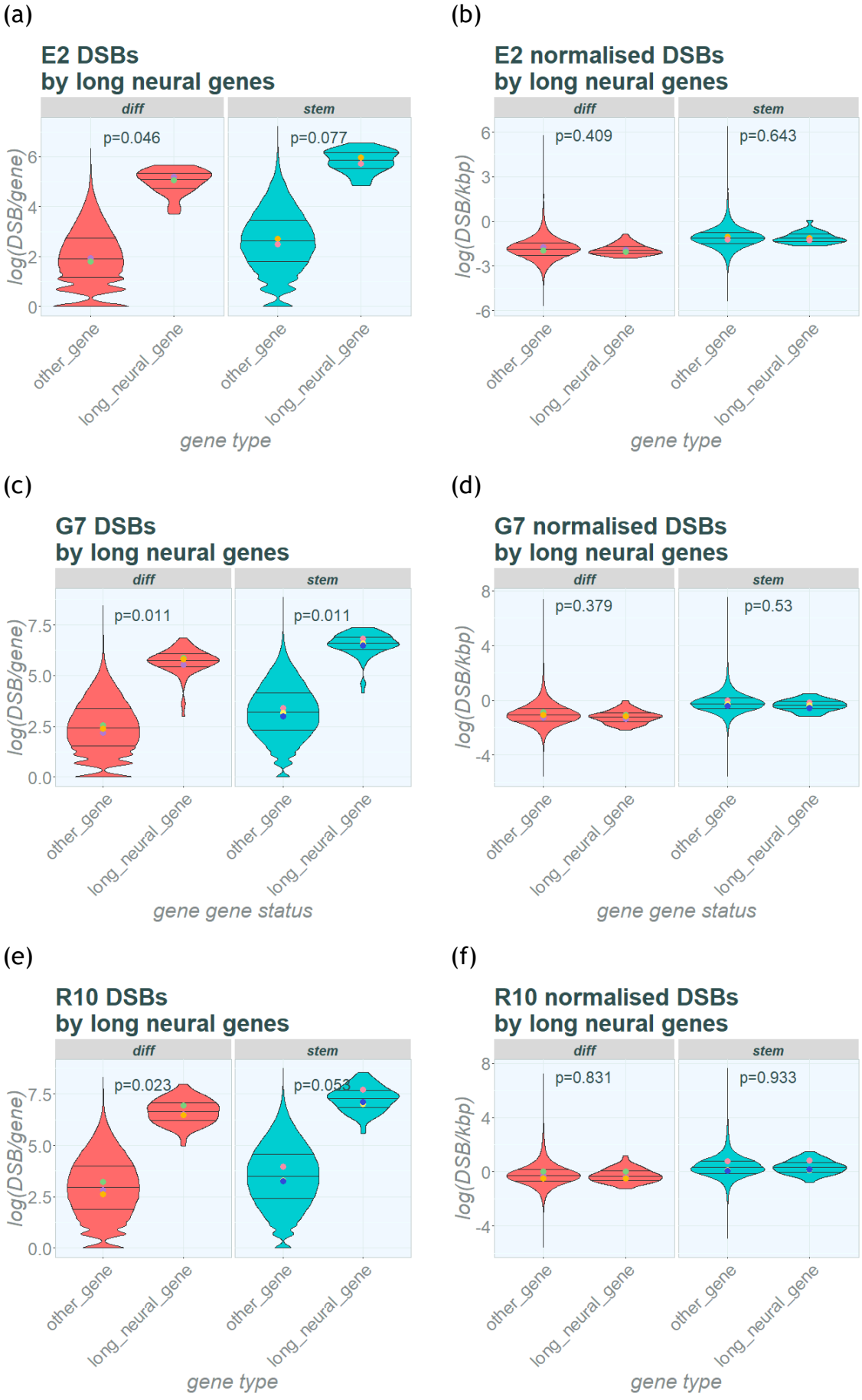



**Figure 4.8: Absolute DSB frequency and DSB density in long neural genes in GBM cells**

Total DSBs and DSBs adjusted for gene length in long neural genes and non-long neural genes ("other_gene"). Violin plots represent DSBs across all genes per repeat. Horizontal lines on violin plots represent interquartile ranges. Results log-transformed for visualisation purposes. Median results reported per repeat as overlayed dots on violin plots. Statistical testing by t-test of medians. P-values reported as graph annotations. Differentiated cells ("diff") results represented in red. GSC ("stem") results represented in turquoise. Results for total DSBs in other genes vs long neural genes displayed on the left, results for DSBs adjusted for gene length for other genes vs long neural genes displayed on the right. (a) E2 GSCs and differentiated cells total DSBs. (b) E2 GSCs and differentiated cells gene DSB density by gene length. (c) G7 GSCs and differentiated cells total DSBs. (d) G7 GSCs and differentiated cells gene DSB density by gene length. (e) R10 GSCs and differentiated cells total DSBs. (f) R10 GSCs and differentiated cells gene DSB density by gene length.

Finally, to give an overview of location of long neural genes in relation to gene length-adjusted DSBs, the mean DSB density across replicates was plotted against gene length (Figure 4.9). G7 is displayed as a representative example of GSC and differentiated cells as well as E2 and R10 cells (additional figures for G7 differentiated cells and E2 and R10 cells available in supplementary figures appendix). Long neural genes are among the longest genes in the human genome which is also visualised here (Zylka et al., 2015). As indicated in Figure 4.8, DSBs in long neural genes did not appear to demonstrate a higher than average density of DSBs when adjusted for gene length, despite showing a higher DSBs density in long neural genes overall.

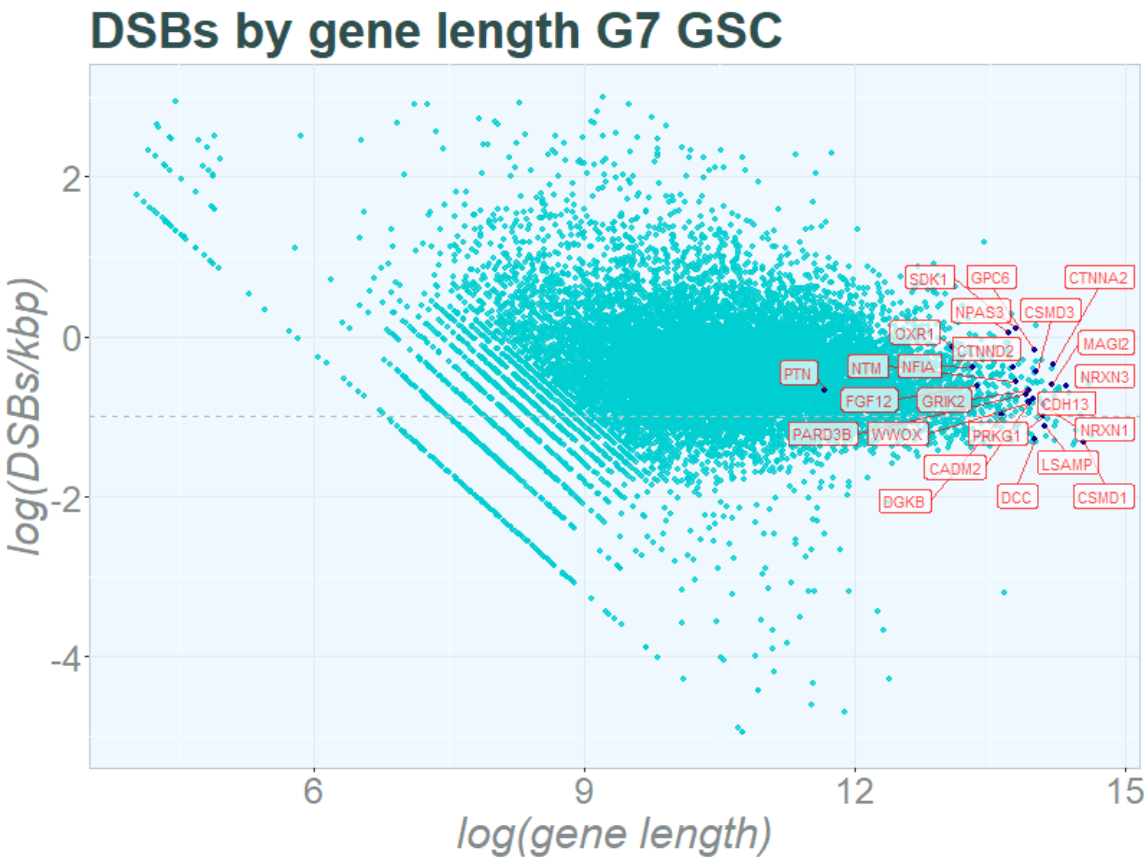



**Figure 4.9: DSBs per gene adjusted to gene length with annotated long neural genes: G7 GSCs**

Gene DSBs adjusted to gene length (DSBs/kbp) against length of gene. Mean DSB density across repeats 1-3 reported in G7 GSCs. X-axis displays log-transformed gene length, y-axis displays log-transformed gene length-adjusted DSBs (DSBs/kbp). Log-transformation performed for visualisation purposes. Turquoise dots represent individual genes. Long neural genes represented as dark blue dots and are annotated by gene name in red. Remaining graphs for E2, G7 and R10 are available in supplementary figures.

## 4.4 Conclusions

This chapter has investigated DSBs in genes with the highest DSB density in both differentiated and GSC lines and additionally in neural cells. The association of DSB density and frequency with gene length has also been explored, with a focus on investigating DSB frequency in long neural genes as locations for replication-transcription collisions. Regarding gene length, the total DSB frequency per gene and the adjusted DSB density per gene has been investigated.

### 4.4.1 Genes with high density DSBs

Having observed DSBs within 50 kbp bin regions in the previous chapter, it was pertinent to note that most of these breaks occurred within genic regions. After accounting for gene length, *MALAT1/TALAM1* remained the most frequently broken gene across GSCs and their differentiated progeny in E2, G7 and R10 cell lines. Other genes that appeared in multiple lines across the top 10 most broken genes were *PCBP1* in E2, G7 and R10, a protein coding gene that acts in chaperoning iron and is associated with autophagy (Hacioglu et al., 2023). Three genes that share characteristics of lncRNA genes; *RMRP*, a long intergenic non-coding RNA (lincRNA), *ENSG00000270640*, a sense intronic protein, and *ENSG0000023496* an antisense protein were also represented in the top ten highest DSB density for genes in E2 GSCs and G7 GSCs.

Given that there were several shared genes within the highest DSB frequencies across differentiated cells and GSCs and across cell lines this raised the important question as to whether these genes were specific to these GBM cell lines. As potential originators of GSCs, neural cell lines were also investigated for genes with high DSB density. BLISS data generated from neural cells demonstrated that *MALAT1/TALAM1* was a highly broken gene in NPC and NEU lines and was within the 100 genes of highest DSB density in NES lines. Whilst the MCF7 data and K562 has not been displayed here specifically for gene DSBs, chapter 3 also indicated there was a shared peak at chromosome 11q, correlating with the location of *MALAT1* also. This common occurrence of *MALAT1* as a gene harbouring a high DSB density is suggestive that the *MALAT1* structure or its transcriptional activity might predispose this gene to developing endogenous DSBs. As previously mentioned, the RNA structure of MALAT1 is highly complex which may indicate that DNA in this region could be prone to secondary structures predisposing to DSB formation (McCown et al., 2019). MALAT1 has also been demonstrated to have RNA G4s which interact with proteins in the nucleolus which, again may indicate a predisposition to DNA secondary structures (Ghosh et al., 2023). Non-canonical secondary DNA structures have been described as contributing to replication fork stalling and collapse by creating obstacles to replication fork machinery progression (Sharma, 2011, Wang and Vasquez, 2006). Our data shows that *MALAT1* is a hotspot for DSBs across multiple cell types. Since MALAT1 RNA is prone to

developing non-canonical structures, this may indicate that non-canonical DNA structures such as G4s are potential contributors to DSB density.

### 4.4.2 Gene length influencing DSBs

As previously described, genes with lengthy transcripts pose the risk of replication-transcription conflicts and negative supercoiling which may require intervention to relieve torsional stress via topoisomerases (McKinnon, 2016, Thongthip et al., 2022). In addition, late replicating regions that can be found in long neural genes have also been identified as high-risk regions for DSB damage (Wilson et al., 2015). This combination of late replicating regions alongside active transcription makes long neural genes a prime location at risk of RS and subsequent DSBs (Hamperl et al., 2017). These sites have been highlighted as likely hotspots of replication-transcription collisions (Bermejo et al., 2012). This made gene length particularly pertinent for investigation in this thesis. The presence of elevated levels of RS in GSCs relative to differentiated progeny has previously been established within these lines (Carruthers et al., 2018). Interestingly, rather than a detrimental effect on cell survival, elevated RS levels have inferred radioresistance in GSCs.

Gene length has been cited as an important factor in genomic fragility, with long neural genes in particular being associated with recurrent DSB clusters in previous studies (Wei et al., 2016). The reasons hypothesised for this fragility have been that very long genes and, in particular, long neural genes require more than one cell cycle to be transcribed. Therefore, these genes become potential hotspots of RS and replication-transcription collisions (Wei et al., 2016). Data presented in this chapter indicated that absolute DSB frequency was higher in the longest genes, however this effect was not maintained when adjusted for gene length. Hence, whilst the longer genes and long neural genes had more DSBs overall, they did not show a greater than average DSB density when accounting for gene length here. Furthermore, there could even be some evidence of longer genes showing some level of protection from DSBs, though this would not appear to extend to long neural genes specifically. It is important to note that whilst Wei et al (Wei et al., 2016) demonstrated recurrent DSBs occurring within long neural genes, these recurrent cluster sites were identified by exposing the cells to aphidicolin in order to induce RS. Additionally, Wei et al

identified aphidicolin-induced recurrent DSB cluster sites first and following this correlated cluster regions with long neural genes rather than investigating genes and gene length as a whole. Therefore, although our data indicated that gene length did not increase endogenous DSB density at baseline, it does not rule out that aphidicolin-induced DSB clusters might have occurred in our cell lines at these genes of interest. Curiously, E2, G7 and R10 GSCs have been shown to have higher levels of RS at baseline compared to differentiated cells (Carruthers et al., 2018). Indeed, RS has been reported as an important contributor to genome fragility in GBM overall (Balzano et al., 2021). Therefore, it might have been expected to see greater correlation between DSB density and gene length in GSCs which was not the case here. Carruthers et al. previously showed that these GSC populations demonstrated consistently raised levels of ATR, RPA and Chk1, with E2 GSCs also demonstrating a higher level of stalled or asymmetric replication forks compared to differentiated progeny (Carruthers et al., 2018). Whilst endogenous RS in these lines has been established, it may be that there is a unique element to using aphidicolin treatment for RS induction that particularly causes DSB clustering at long neural genes. It is also important to note that other groups have only shown that aphidicolin-induced RS contributed to DSBs in neural cells days after treatment (Michel et al., 2022, Wang et al., 2020). Additionally, Michel et al. demonstrated that knockdown of *p53* in neural progenitor cells abrogated the effect of aphidicolin (Michel et al., 2022). In this study, the *p53* knockdown cells showed no difference in DSB distribution between control and aphidicolin-treated cells. Hence the reasons why GSCs may fail to demonstrate DSB clusters at long neural genes could be for a number of reasons. Whilst neural cells and GSCs share some very important similarities, it should still be highlighted that there will always be important differences in terms of function due to mutation and cell programming. Even in neural progenitor cells derived from non-cancer disease processes, we see key differences in p53 function and DDR. For example, Wang et al demonstrated that neural progenitor cells from patients with hyperproliferative autistic spectrum disorder were able to activate ATR in response to RS with only low levels of p53 activation (Wang et al., 2020). Furthermore, GSCs may simply not demonstrate long gene fragility in the same way as neural cells following mutagenesis and acquired genomic instability. Finally, whilst adjusted DSB density for gene length gives a helpful comparison to other genes, the absolute DSB frequency in genes

is likely still a key factor in gene transcription and overall contribution of DSBs. The total quantity of DSBs in long neural genes may still have important biological implications and remains an important consideration in the context of RS and replication-transcription conflicts.

### 4.4.3 Conclusions summary

- GSCs and differentiated progeny demonstrate a high concordance in the genes with the highest density of DSBs

- GSCs share genes with high DSB density across GBM cell lines: E2, G7 and R10.

- *MALAT1* is a DSB gene hotspot across all analysed GBM cell lines and in neural cell lines.

- Total DSB frequency in genes increases with gene length in GSCs, however this effect is lost when adjusted for gene length and long neural genes do not show a greater than average DSB density after DSBs are adjusted for gene length.

# Chapter 5 Investigating DSBs in gene bodies and annotated genomic regions

## 5.1 Introduction

The previous chapters looked at an overview of the breakome of our cell lines of interest at a chromosomal level, explored the relationships with gene length and identified genomic sites demonstrating a high frequency of DSB generation. Given that DSBs seem to show predilection for genic sites we wished to look into this at a higher resolution, taking into account genic architecture. This chapter investigated locations of interest within genes at transcription start and termination sites (TSS and TTS respectively) and across gene bodies. In addition to this, DSBs were annotated for regions of interest within the whole genome to identify locations of high DSB density.

### 5.1.1 Gene bodies

Given the clustering of DSBs across genes, DSB density across gene bodies was investigated. In particular, having investigated long genes, TTS sites were of particular interest to assess the DSB density across these regions, especially given the potential influence of endogenous physiologically occurring DSBs secondary to topoisomerases. Topoisomerases have been cited as important in the maintenance of homeostasis in neural cells and in maintaining genomic integrity in late replicating sites (McKinnon, 2016, Helmrich et al., 2011). Additionally, topoisomerase activity in GBM has also been associated with GBM treatment resistance to RS-inducing drugs (Kenig et al., 2016).

### 5.1.2 Annotated genomic sites

Understanding that genes are complex codes made up of several components highlights the opportunity to consider sites within genes as significant regions for DSB occurrence with possible implications for gene function. Reviewing DSBs across gene locations and other genomic elements across GSCs and differentiated cells allowed for comparison between these populations. Additionally, describing DSBs in GBM gene locations has the opportunity to

provide insight into how elements of DSB distribution between GBM and non-GBM lines compare.

### 5.1.3 Aims

- Investigate DSB density across TSS, TTS and gene bodies in GSC and differentiated GBM cells (E2, G7 and R10) and compare with DSB density across neural cell lines and commercial cell lines.

- Quantify the total DSBs at annotated sites of interest within genes in GBM lines compared to expected DSB frequencies across genes with reference to neural cells and commercial cancer cell lines

## 5.2 Materials and Methods

### 5.2.1 DSBs across gene bodies, TSS and TTS regions

To obtain the pattern of DSBs across genes, the mean DSBs were calculated using the ChIPSeeker package (Yu et al., 2015). The plotPeakProf2() function was used to map mean DSBs across the whole gene body per cell line. In brief, the package allowed for calculation of mean DSBs across all genes using 800 bins per gene (Ramírez et al., 2016). Mean DSBs across genes were adjusted to gene length to allow comparison across differing gene lengths. Mean DSBs were represented across gene bodies and gene length was displayed as a percentage from 0% indicating TSS to 100% indicating TTS. Mean DSBs outside of gene bodies was also plotted for comparison where mean DSBs prior to and following TSS and TTS regions were calculated. For mean DSBs outside of gene bodies, the length of DNA used for this was 20% of the corresponding gene length. The mean DSBs across gene bodies +/-20% was then plotted as a line graph.

Mean DSBs were also calculated across TSS and TTS length specific regions using ChIPSeeker plotPeakProf2(). For each gene, 3000 bp prior to and following the TSS or TTS was identified. These regions were divided into 800 bins and the mean DSBs per bin was calculated per gene across all gene greater than 800 bp long. The mean DSBs per bin was then calculated and plotted as a line graph across TSS and TTS.

For calculated mean DSBs across TSS and gene bodies, distance to transcription start site was calculated using nearestTSS() function from EdgeR (Robinson et al., 2010) for each DSB site. Genes longer than 1000 bp were excluded from analysis. DSBs within 500 bp before or after a TSS were classed as TSS DSBs and DSBs occurring within genes but not within 500 bp of TSS were classed as gene body DSBs.

### 5.2.2 DSBs and genomic sites of interest

For annotating genomic locations DSBs were annotated using the “annotatR" package (Cavalcante and Sartor, 2017). Each DSB was annotated with a location as follows: regions within 1 to 5 kb of a TSS, promotors less than 1 kb from TSS, 5’ untranslated regions (UTRs), exons, intron/exon boundaries exons, intron/exon boundaries (200 bp upstream/downstream of boundaries between introns and exons), introns, 3’ UTRs and intergenic regions. Given that DSBs may have overlapped with more than one annotation type, the annotations were given the following site priority: Promoters less than 1 kb from TSS, 1 to 5kb promoters, 5’ UTRs, 3’ UTRs, exons, introns, intron/exon boundaries, and intergenic sites. The actual DSBs per region were also compared with randomly generated DSB frequency to compare actual DSB distribution with expected DSB distribution location. To generate randomly distributed locations which were equivalent to the dataset the regioneR::randomizeRegions package (Gel et al., 2016) was used to generate a random equivalent distribution of DSBs across genomic sites using BLISS bedfiles. Blacklist locations were excluded from randomly generated DSB files by masking prespecified blacklist sites (Amemiya et al., 2019). These regions had also already been excluded from the BLISS bedfiles.

## 5.3 Results

### 5.3.1 DSBs across genes

Gene bodies were explored in further detail with particular attention paid to TSS and TTS locations given the particular focus on the influence of transcription and topoisomerases on DSBs (Michel et al., 2022, Promonet et al., 2020).

#### 5.3.1.1 DSBs across gene bodies demonstrate conserved DSB patterns within GBM lines

Mean DSBs were mapped across gene bodies for GBM GSC and differentiated populations (Figure 5.1, Table 5.1). This allowed for an overview of the pattern of DSBs across TSS, gene bodies and TTS regions. Additionally, +/-20% length outside of gene bodies was also included as a comparator. To compare mean DSBs across TSS and gene bodies, these were displayed in Table 5.1 and TSS/gene body mean DSB ratios calculated. Across all three GBM lines there was a consistent increase in mean DSBs immediately prior to the TTS relative to gene body DSBs. The ratio of mean DSBs across TSS and gene bodies varied significantly across GBM lines ($p=0.0013$).

In E2 GSCs and differentiated cells, there was no distinguishable visible pattern across TSS. This was consistent across the two repeats and across GSCs and differentiated cells. At TTS, there was a clear and consistent peak of mean DSBs visible across GSC and differentiated cells.

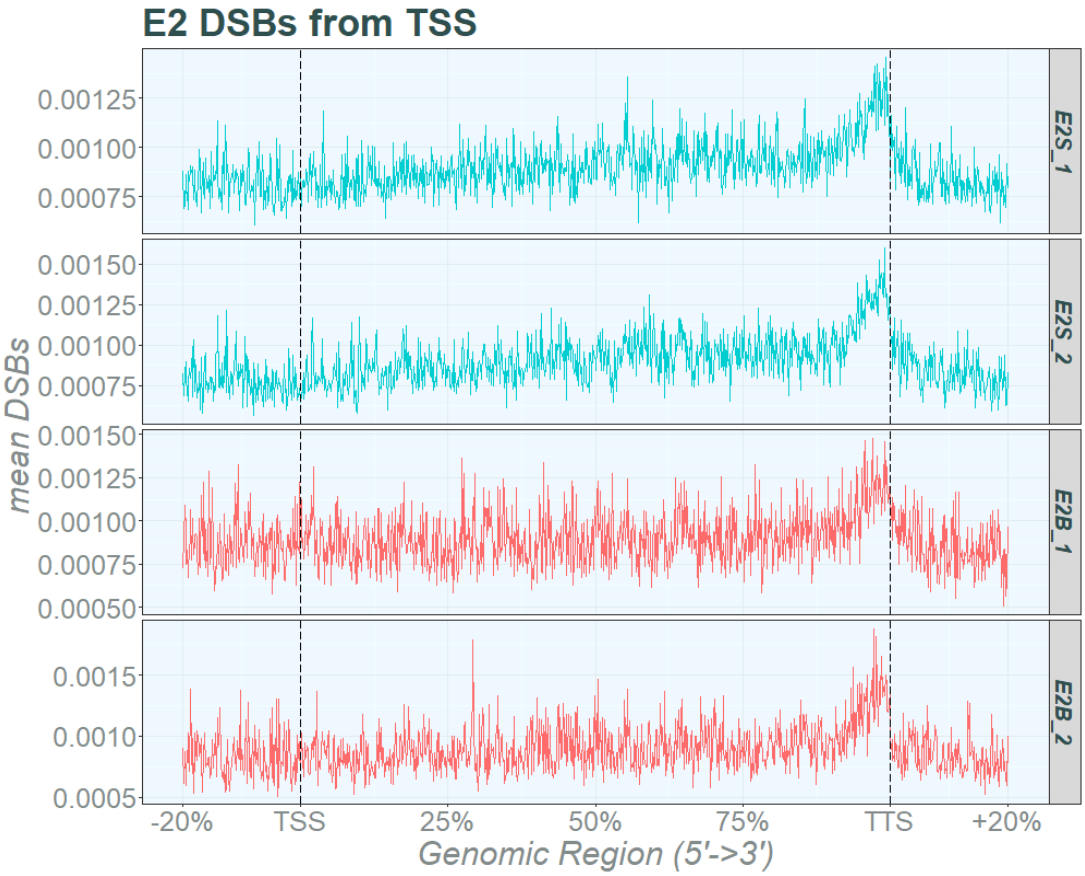
Regarding G7 cells, the mean DSB pattern was also consistent across GSCs and differentiated cells (Figure 5.1). In contrast to E2 cells, G7 cells demonstrated a higher TSS/gene body ratio (Table 5.1). Across gene bodies there appeared to be an overall increase in mean DSBs compared to sites outside of the gene bodies. There was an increase in mean DSBs immediately prior to TTS, similar to E2 cells.

Finally, the mean DSBs across R10 cells gene bodies were investigated (Figure 5.1). Mean DSBs across R10 GSCs versus differentiated cells were also highly similar in pattern to each other with TSS/gene body ratios of between 2.8 and 3.5; the lowest of the ratios compared to both E2 and G7 cells. At TTS regions R10 cells demonstrated a peak in DSBs which was highly similar in pattern to both E2 and G7 TTS regions.

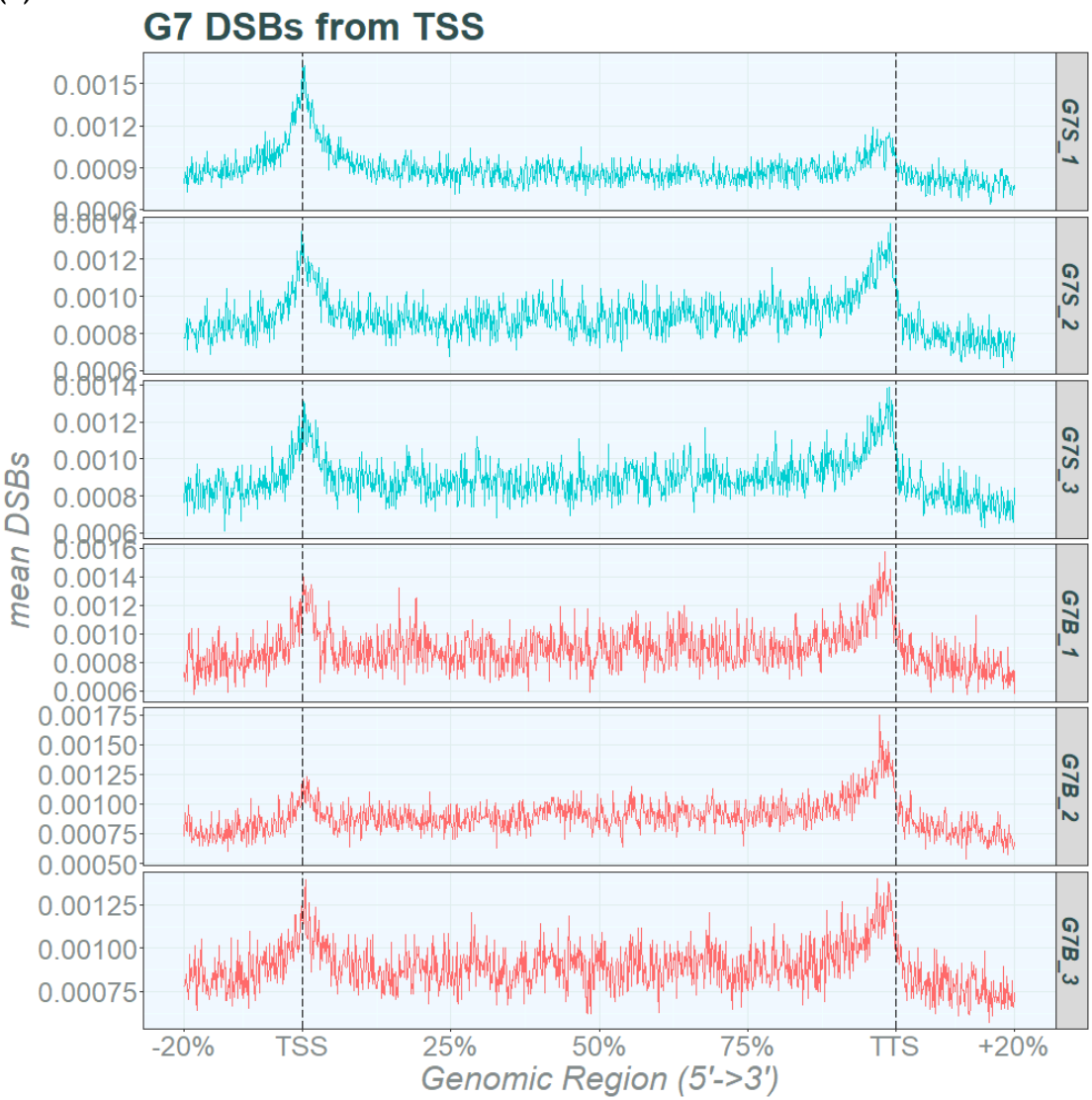
The differences in TSS DSB patterns indicated some variation in DSB distribution between the three cell lines. However there appeared to be a shared similar pattern of DSBs occurring at TTS regions. Overall, these findings indicated that TSS and TTS locations were important sites of interest regarding DSBs, apparent

differences across TSS locations but a seemingly consistent accumulation of DSBs at TTS locations.

(a)


E2 DSBs from TSS
mean DSBs
0.00125
0.00100
0.00075
0.00150
0.00125
0.00100
0.00075
0.00150
0.00125
0.00100
0.00075
0.00050
0.0015
0.0010
0.0005
E2S_1
E2S_2
E2B_1
E2B_2
-20%
TSS
25%
50%
75%
TTS
+20%
Genomic Region (5'->3')

**(b)**


G7 DSBs from TSS
mean DSBs
0.0015
0.0012
0.0009
0.0006
0.0014
0.0012
0.0010
0.0008
0.0006
0.0014
0.0012
0.0010
0.0008
0.0006
0.0016
0.0014
0.0012
0.0010
0.0008
0.0006
0.00175
0.00150
0.00125
0.00100
0.00075
0.00050
0.00125
0.00100
0.00075
G7S_1
G7S_2
G7S_3
G7B_1
G7B_2
G7B_3
-20%
TSS
25%
50%
75%
TTS
+20%
Genomic Region (5'->3')

(c)

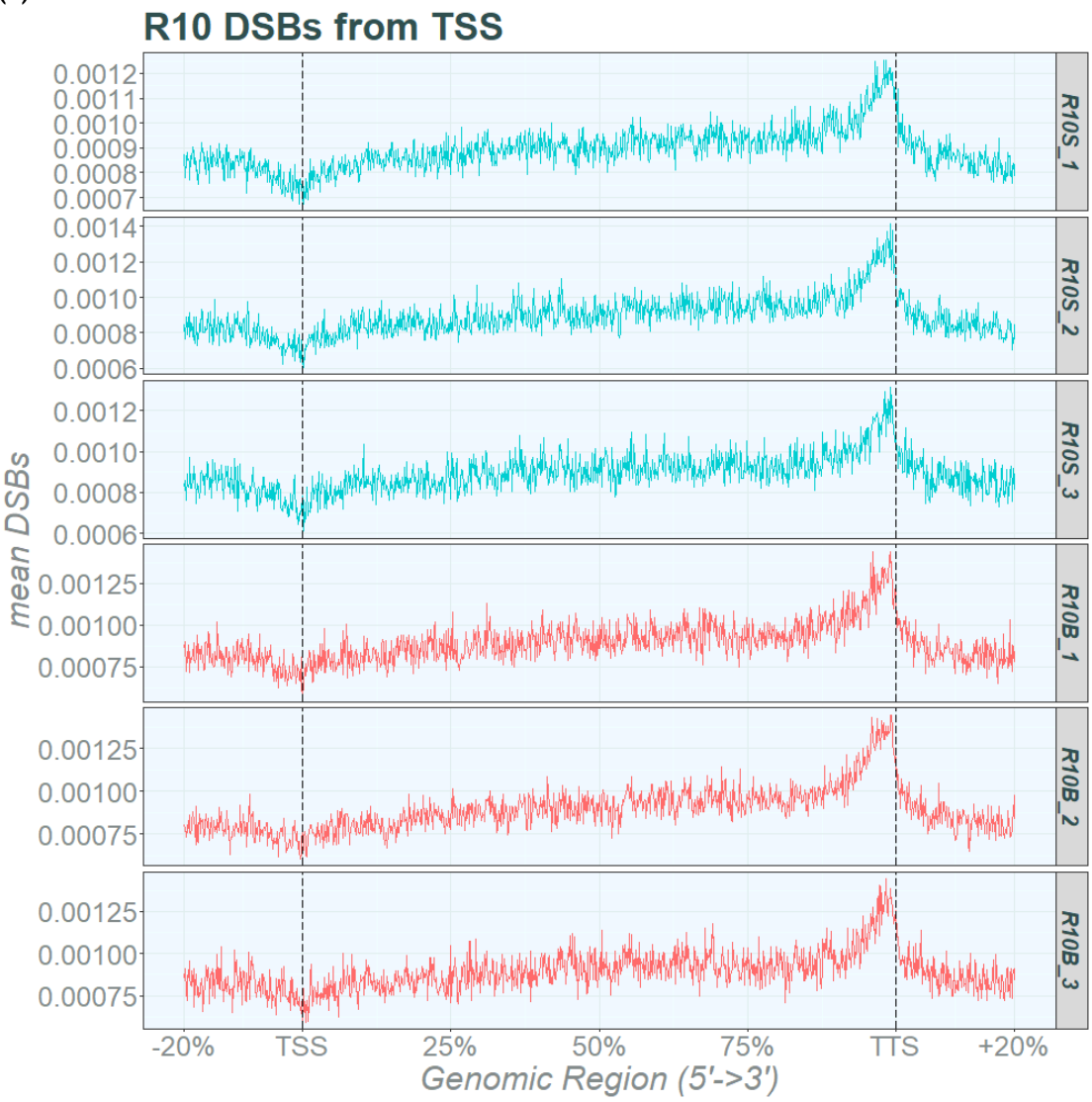



**Figure 5.1. Mean DSBs across GBM gene bodies: E2, G7 and R10**

Mean DSBs plotted along all gene bodies +/-20% of gene length. GSCs in turquoise and differentiated cells in red. Individual repeats displayed. TSS and TTS are displayed as dashed lines. DSBs across gene bodies are displayed as percentage of length in a 5' to 3' direction. (a) E2 GSCs, repeats 1-2 (E2S_1, E2S_2) and E2 differentiated cells, repeats 1-2 (E2B_1, E2B_2). (b) G7 GSCs, repeats 1-3 (G7S_1, G7S_2, G7S_3) and G7 differentiated cells, repeats 1-3 (G7B_1, G7B_2, G7B_3). (c) R10 GSCs, repeats 1-3 (R10S_1, R10S_2, R10S_3) and R10 differentiated cells, repeats 1-3 (R10B_1, R10B_2, R10B_3).

**Error! Reference source not found.** displays TSS and gene body mean DSBs and TSS/gene body DSB ratios. The TSS/gene body mean DSB ratio indicated a relative consistency between GBM repeats in E2, G7 and R10 but differences between lines. R10 and G7 demonstrated the greatest difference between lines (Kruskal-Wallis p=0.0065). Differences between E2 and G7 TSS/gene body ratios and between E2 and R10 were still significant (Wilcoxin Signed Rank p=0.0095

and p=0.0095 respectively). Across GSCs and differentiated cells GSCs tended to have a lower mean DSB TSS/gene body ratio between paired repeats compared to differentiated cells.

**Table 5.1 Mean DSBs at TSS and gene bodies in GBM cell lines**

| | Mean DSBs across TSS | Mean DSBs across gene bodies | TSS/gene body ratio | |
|---|---|---|---|---|
| **E2 GSC rep 1** | 0.000558 | 0.000131 | 4.269 | |
| **E2 diff rep 1** | 0.000353 | 0.000067 | 5.269 | |
| **E2 GSC rep 2** | 0.000457 | 0.000109 | 4.194 | |
| **E2 diff rep 2** | 0.000282 | 0.000055 | 5.130 | |
| **G7 GSC rep 1** | 0.002258 | 0.000322 | 7.020 | |
| **G7 diff rep 1** | 0.000730 | 0.000101 | 7.241 | |
| **G7 GSC rep 2** | 0.001772 | 0.000255 | 6.947 | |
| **G7 diff rep 2** | 0.000990 | 0.000136 | 7.256 | |
| **G7 GSC rep 3** | 0.001364 | 0.000220 | 6.194 | |
| **G7 diff rep 3** | 0.000853 | 0.000123 | 6.956 | |
| **R10 GSC rep 1** | 0.002290 | 0.000757 | 3.023 | |
| **R10 diff rep 1** | 0.000911 | 0.000265 | 3.434 | |
| **R10 GSC rep 2** | 0.001264 | 0.000400 | 3.157 | |
| **R10 diff rep 2** | 0.001221 | 0.000367 | 3.325 | |
| **R10 GSC rep 3** | 0.001170 | 0.000408 | 2.870 | |
| **R10 diff rep 3** | 0.000726 | 0.000228 | 3.181 | |
| **Kruskal-Wallis test** | p=0.0013 | **E2-G7** p=0.0095 | **E2-R10** p=0.0095 | **G7-R10** p=0.0065 |

Mean DSBs across TSS and gene bodies for GSCs and differentiated GBM lines E2, G7 and R10. TSS region captured as +/- 500bp from TSS. Gene body region captured as +500bp from TSS to gene end. Gene less thank 1000bp excluded from analysis. Mean DSB ratio of TSS DSBs/gene body DSBs also shown. Significance testing performed against TSS/gene body ratios. Kruskal-Wallis testing performed for non-parametric data (Shapiro-Wilk p=0.016). Wilcoxin Signed Rank test performed with BHP correction for multiple testing.

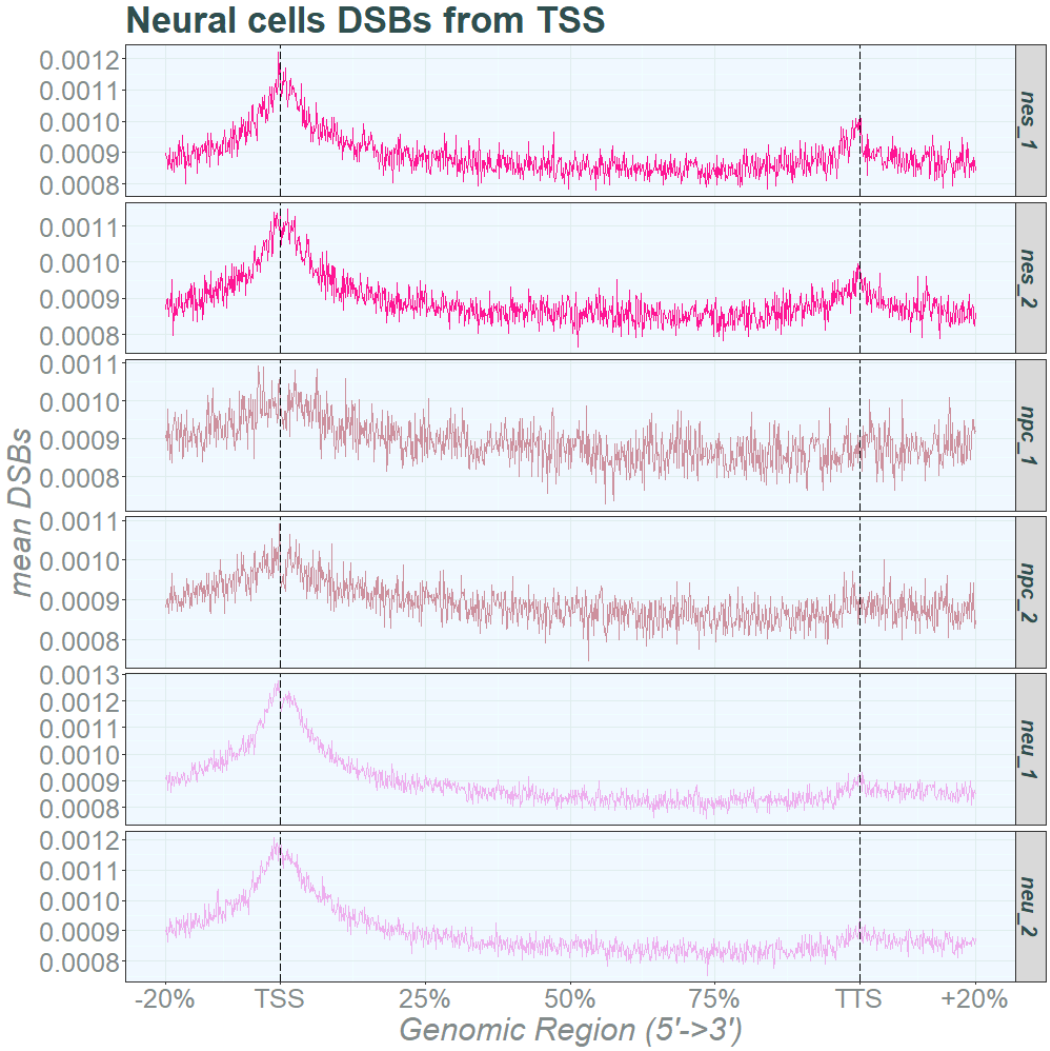
#### 5.3.1.2 DSB patterns across gene bodies in neural cells and commercial cell lines demonstrate consistently higher mean DSBs at TSS locations

Given that GSCs showed differences in DSB distribution across TSS regions as well as a consistent peak in DSBs at TTS, it was of interest to determine whether these differences were shared across other cell types. Therefore, mean DSB patterns were also looked at in the three publicly available neural cell lines NES, NPC and NEU in addition to the commercial cancer lines MCF7 and K562.

Across neural cell gene bodies, mean DSB patterns demonstrated an increase in DSBs prior to and following TSS. The mean DSBs had some smaller variations across TTS regions between neural lines where NES cells demonstrated a peak at TTS which was less apparent in NPC and NEU cells. Additionally, the peak around TTS regions in NES cells was less prominent than the peak at TSS sites and also appeared less obvious compared to GSC lines. Mean DSBs across gene bodies appeared relatively similar across NES, NPC and NEU cells. The NPC cells had a less distinct peak at TSS regions and a greater "background noise" visible overall.

Regarding commercial cancer cell lines MCF7 and K562, the mean DSBs across gene bodies both showed an increase in mean DSBs immediately prior to and following TSS regions. K562 did not show signs of a peak at TTS or any visible variation between gene body regions and regions +/-20% outside of gene bodies. Regarding MCF7, there appeared to be an increase in the mean DSBs across the gene bodies compared to +/-20% outside of gene body regions. Additionally, MCF7 cells also appeared to show a peak immediately prior to TTS regions. This indicated that DSB pattern across genes was not identical across cell lines from other tissues and, in particular, that the relative accumulation of DSBs at TTS was not universal to the commercial cell lines investigated.

(a)


Neural cells DSBs from TSS
mean DSBs
0.0012
0.0011
0.0010
0.0009
0.0008
0.0013
nes_1
nes_2
npc_1
npc_2
neu_1
neu_2
-20%
TSS
25%
50%
75%
TTS
+20%
Genomic Region (5'->3')

(b)

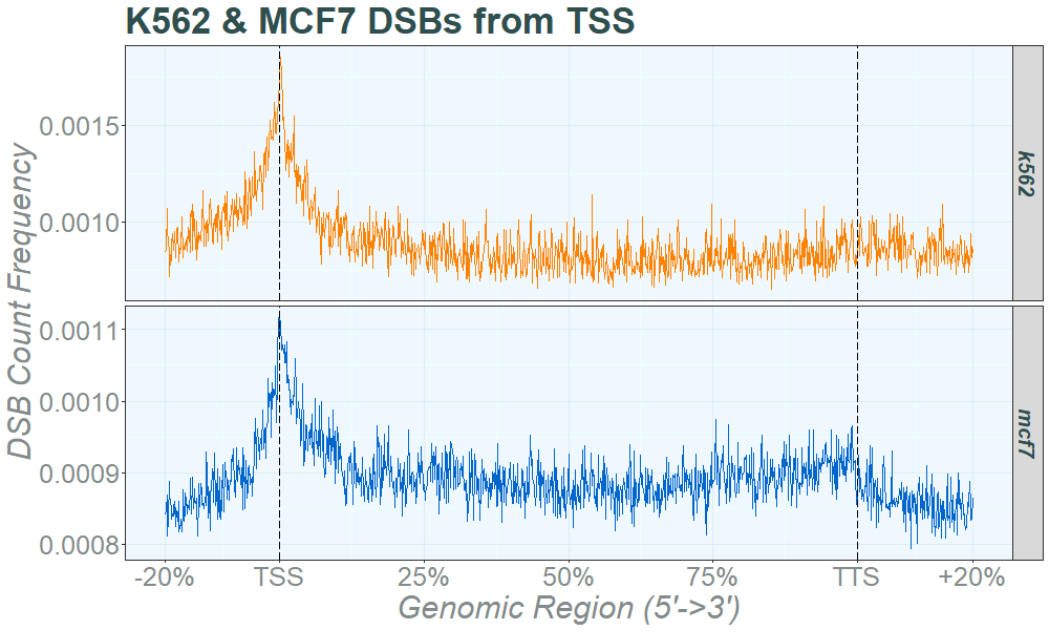

K562 & MCF7 DSBs from TSS
DSB Count Frequency
0.0015
0.0010
0.0011
0.0009
0.0008
k562
mcf7
-20%
TSS
25%
50%
75%
TTS
+20%
Genomic Region (5'->3')

**Figure 5.2. Mean DSBs across neural cell and commercial cancer cell line gene bodies: NES, NPC, NEU and K562 & MCF7**

Mean DSBs plotted along all gene bodies +/-20% of gene. Neural cells plotted in pink, K562 cells plotted in orange and MCF7 cells plotted in blue. Neural cells display individual repeat data. TSS and TTS are displayed as dashed lines. DSBs across gene bodies are displayed as percentage of length in a 5' to 3' direction. (a) Neural cells displayed in order top to bottom: NES (neuroepithelial stem cells in hot pink) repeats 1-2 nes_1 and nes_2, NPC (neural progenitor cells in dark pink) repeats 1-2 npc_1 and npc_2, NEU (post-mitotic neural cells in light pink) repeats 1-2 neu_1 and neu_2. (b) Commercial cancer lines displayed: K562 (erythroleukaemia in orange) single sample k562, MCF7 (breast cancer in blue) single sample mcf7.

#### 5.3.1.3 DSB patterns at TSS are not uniform across GBM lines

Having identified that both TSS and TTS locations displayed differences in DSB patterns in GBM lines, TSS and TTS locations were mapped specifically to examine in further detail. For this, rather than mapping mean DSBs across all gene bodies, a distance of 3000 bp prior to and following TSS and TTS regions was mapped using mean DSBs. This allowed more specific review of these regions individually and also took into account absolute distance rather than distance as a measure of percentage of total gene length.

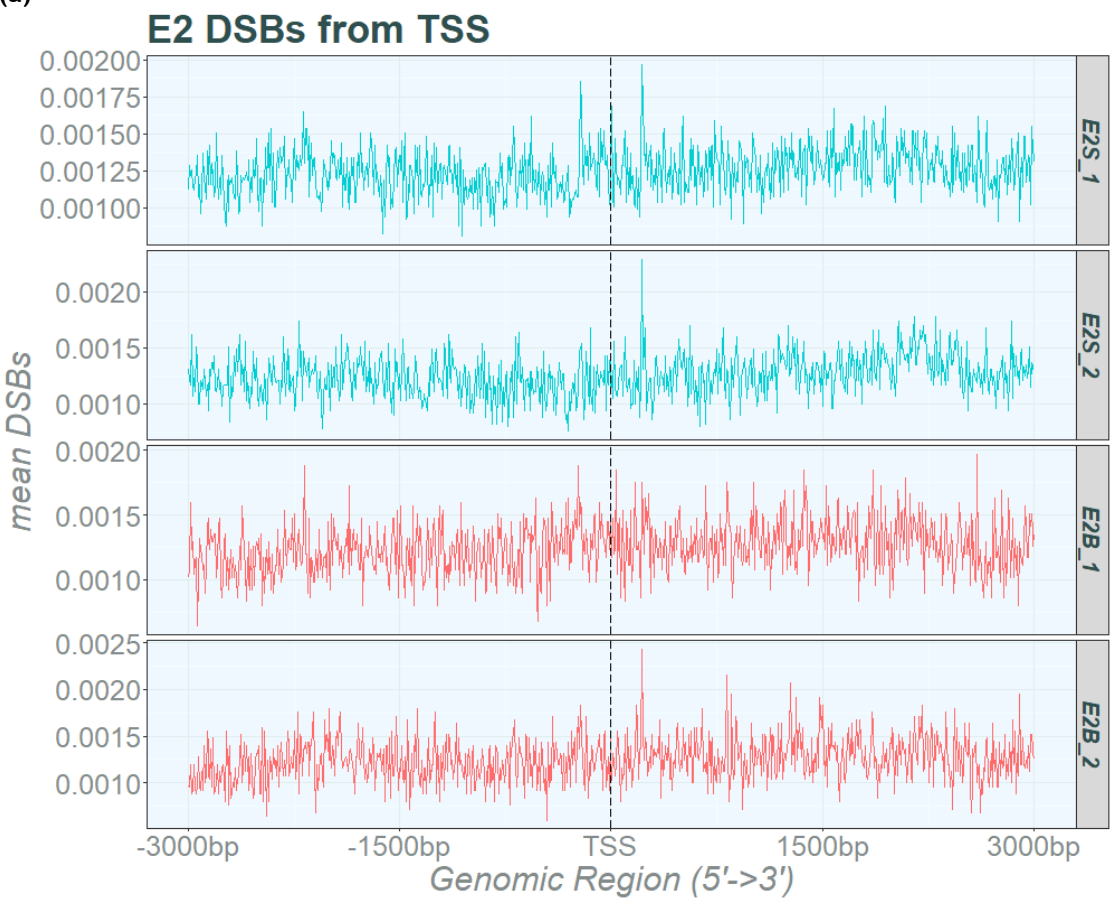
The mean DSBs across TSS for GBM cell lines is displayed in Figure 5.3. As previously visualised in Figure 5.1, E2 GSCs and differentiated cells demonstrated a relatively consistent mean DSB pattern across the TSS which did not change in pattern across repeats or across GSCs and differentiated cells.

Regarding G7 lines, there remained an increase in mean DSBs across TSS in GSCs and differentiated lines, however GSCs appeared to develop a small dip within 500 bp of the TSS. Though, even in the dip immediately prior to and following the TSS, the mean DSB frequency remained higher than sites -1500 bp from TSS. This was less clearly present in differentiated cells.

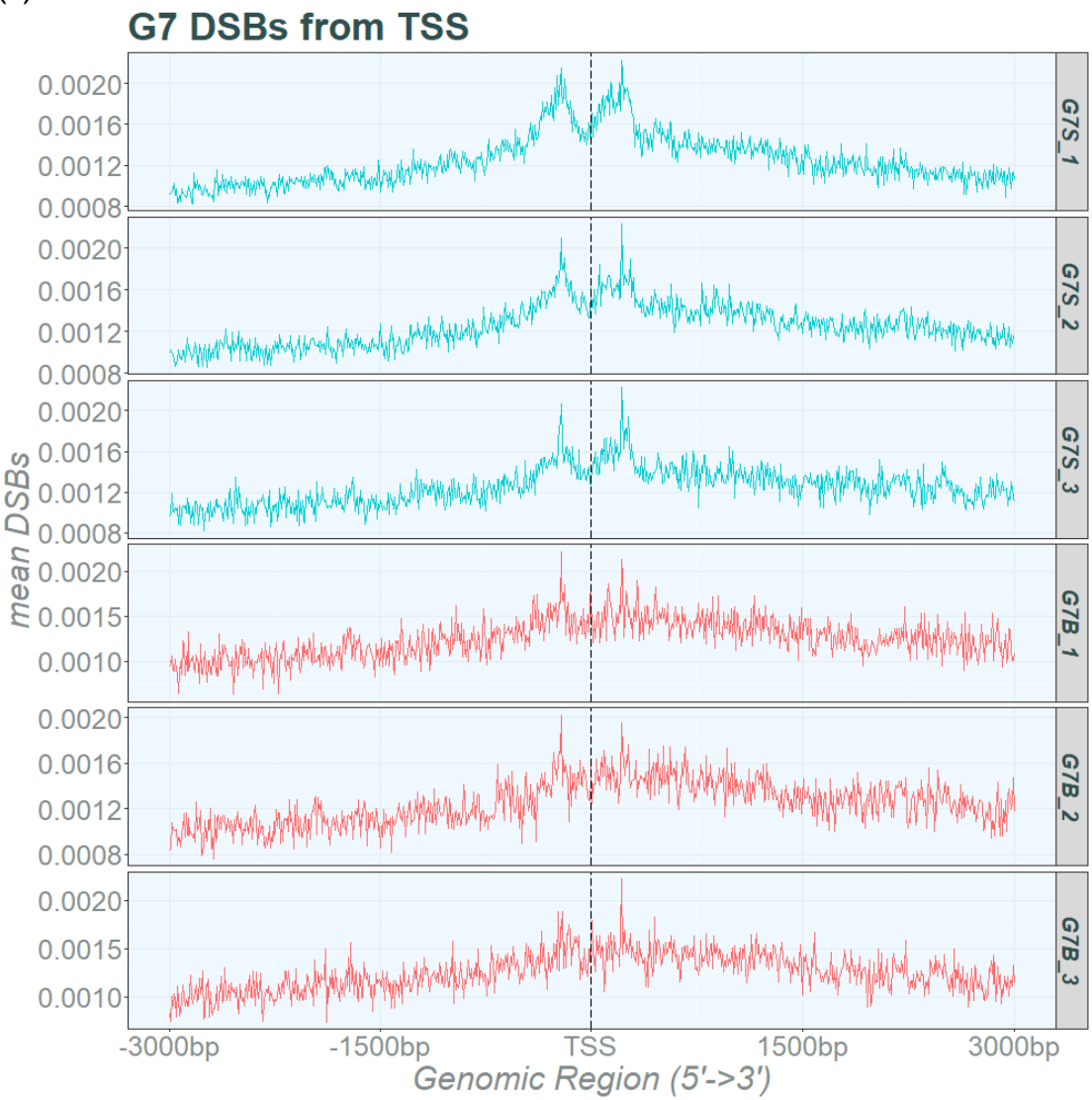
Finally, R10 cells continued to show a dip in mean DSBs at TSS. This was also consistent with Figure 5.1 when reviewed in terms of absolute distances from TSS rather than gene length percentage. The dip in mean DSBs was maintained across repeats and also across R10 GSC and differentiated lines.

Taken together, the differences across GBM lines across the gene bodies at TSS was maintained when looking at distance from TSS. Patterns of mean DSBs across TSS were conserved across repeats and between GSCs and differentiated cells.

**(a)**



**(b)**


G7 DSBs from TSS
mean DSBs
0.0020
0.0016
0.0012
0.0008
0.0015
0.0010
G7S_1
G7S_2
G7S_3
G7B_1
G7B_2
G7B_3
-3000bp
-1500bp
TSS
1500bp
3000bp
Genomic Region (5'->3')

(c)

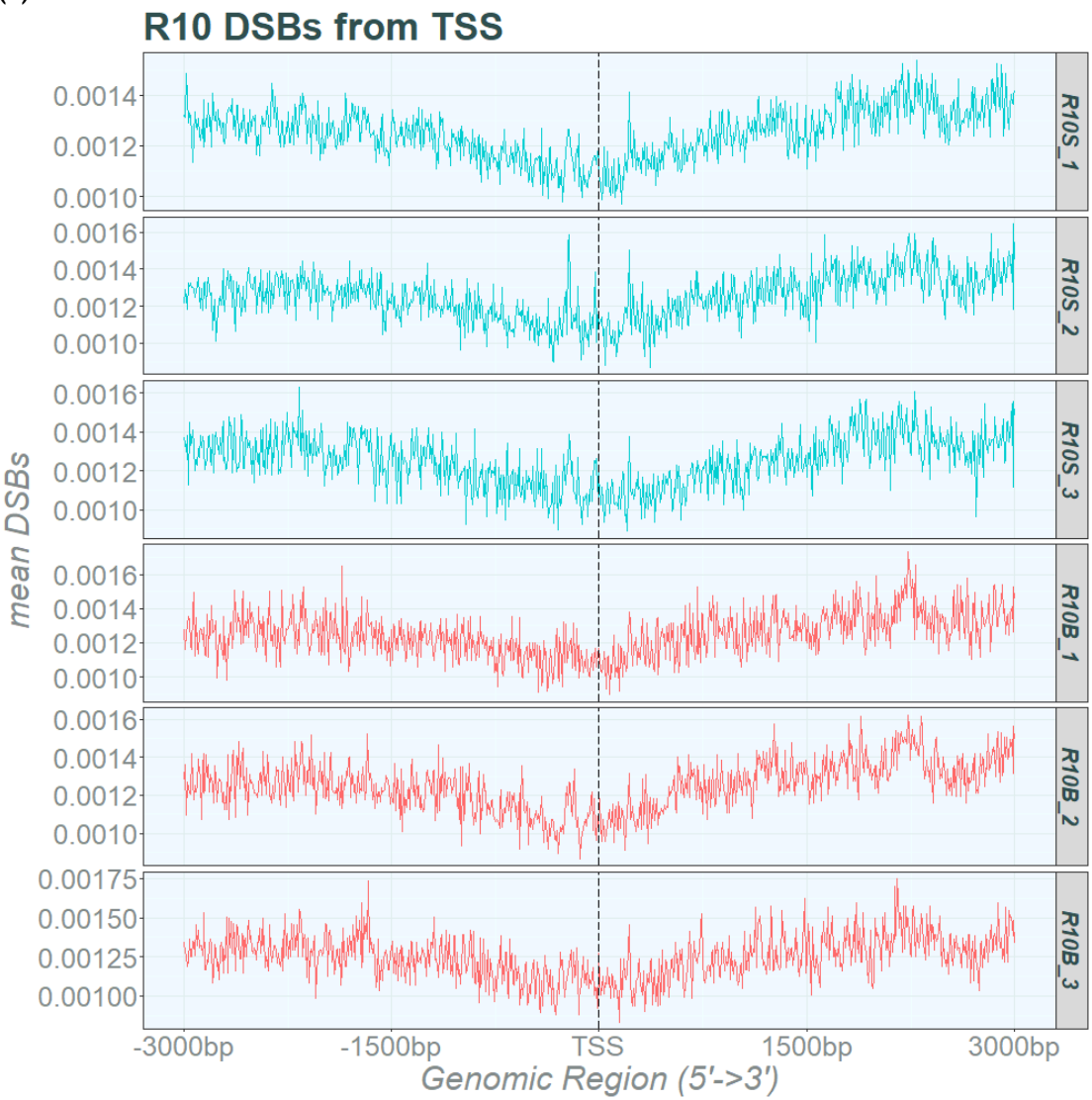



**Figure 5.3. Mean DSB frequency across TSS +/-3000 bp in GBM lines E2, G7 and R10**

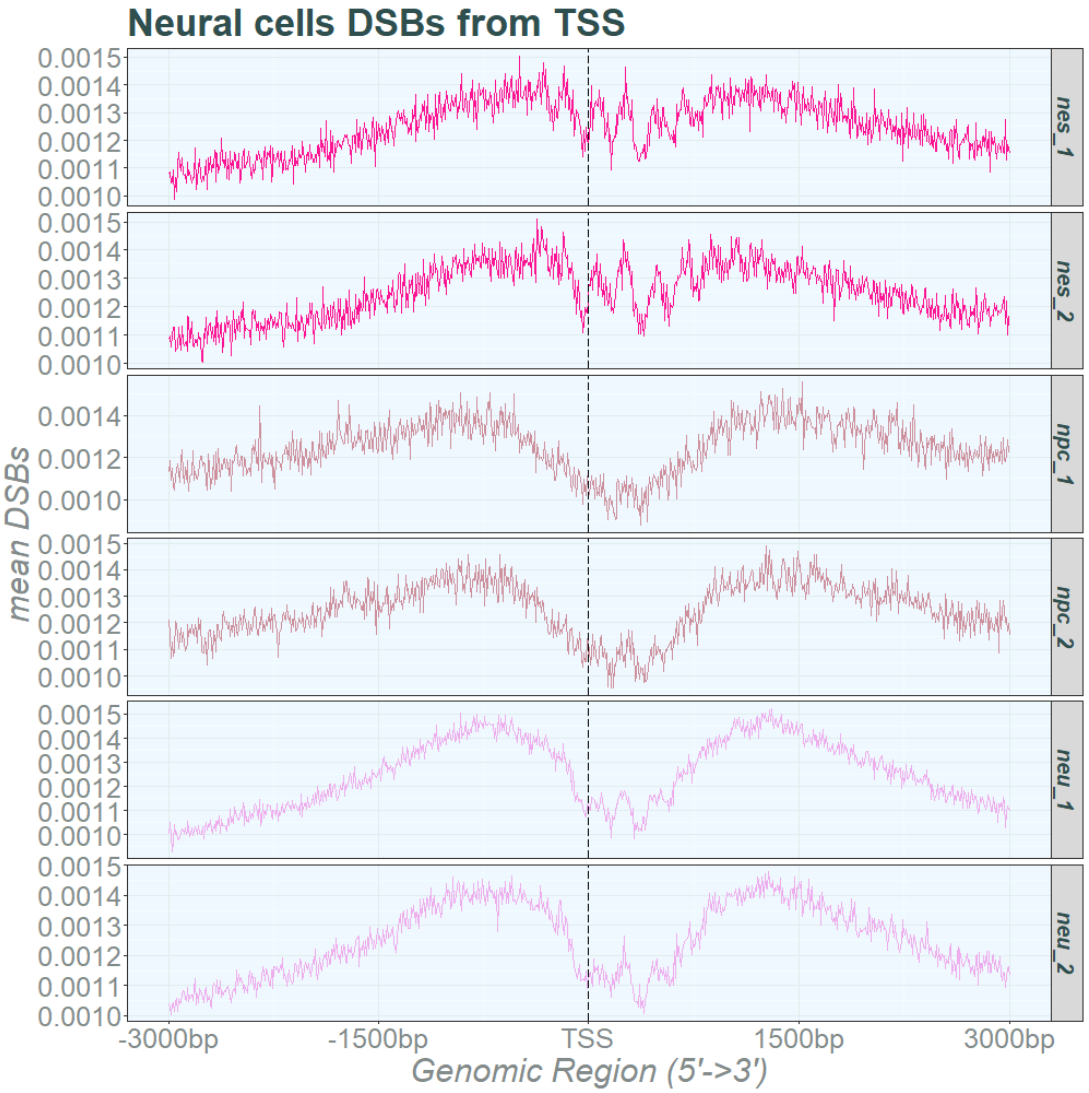
Mean DSB frequency across GBM TSS in E2, G7 and R10 lines. GSCs in turquoise and differentiated cells in red. Individual repeats displayed. TSS are displayed as dashed central line. Region displays 3000 bp prior to and following TSS in a 5' to 3' direction. (a) E2 GSCs, repeats 1-2 (E2S_1, E2S_2) and E2 differentiated cells, repeats 1-2 (E2B_1, E2B_2). (b) G7 GSCs, repeats 1-3 (G7S_1, G7S_2, G7S_3) and G7 differentiated cells, repeats 1-3 (G7B_1, G7B_2, G7B_3). (c) R10 GSCs, repeats 1-3 (R10S_1, R10S_2, R10S_3) and R10 differentiated cells, repeats 1-3 (R10B_1, R10B_2, R10B_3).

The mean DSB frequency across TSS in neural cells and commercial cell lines was also mapped for comparison. For neural cells, whilst there was an apparent increase in mean DSB frequency across TSS in Figure 5.2, all three neural cells NES, NPC and NEU, showed a more nuanced pattern when looking at absolute distance from TSS. Whilst NPC and NEU lines showed an overall increase in mean DSB at TSS there was also a clear dip in DSBs immediately prior to and following

the TSS. This pattern was maintained across repeats. NES samples had a similar pattern to NPC and NEU lines though the dip was less pronounced overall. This pattern in mean DSB frequency had some similarities to G7 GSCs, though was clearly more pronounced in neural cell lines.

With regards to the commercial cancer lines K562 and MCF7, both continued to demonstrate an overall increase in mean DSBs along TSS. The K562 line did appear to show a small decrease in mean DSBs immediately before and after TSS, though conclusions were limited given this was a single sample. However, this short mean DSB frequency decrease was reminiscent in appearance to G7 GSC TSS patterns. Altogether, the commercial cancer cell lines showed an overall peak across TSS. Interestingly, at this level of detail, neural cells were able to demonstrate a more complex DSB pattern at TSS, with an overall regional increase but also a clear dip immediately prior to and following TSS. This pattern appeared unique to these neural cells and was different even compared to R10 which also demonstrated a dip in mean DSB frequency at TSS. Both MCF7 and K562 matched G7 TSS best in terms of pattern of DSBs across TSS locations.

(a)


Neural cells DSBs from TSS
mean DSBs
nes_1
nes_2
npc_1
npc_2
neu_1
neu_2
-3000bp
-1500bp
TSS
1500bp
3000bp
Genomic Region (5'->3')

(b)

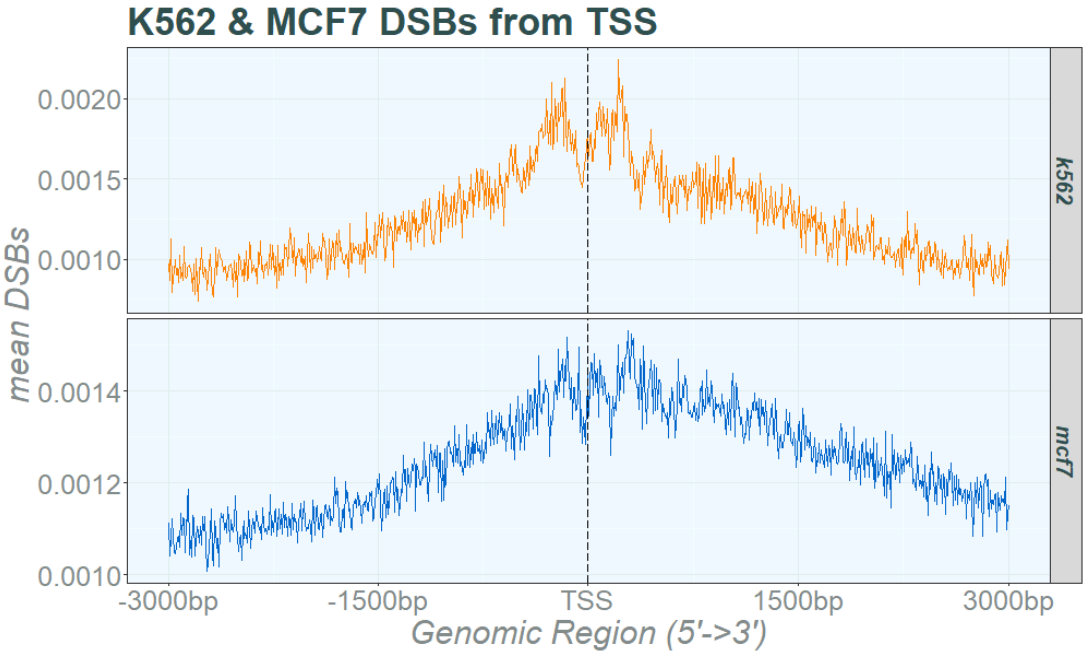

K562 & MCF7 DSBs from TSS
mean DSBs
k562
mcf7
-3000bp
-1500bp
TSS
1500bp
3000bp
Genomic Region (5'->3')

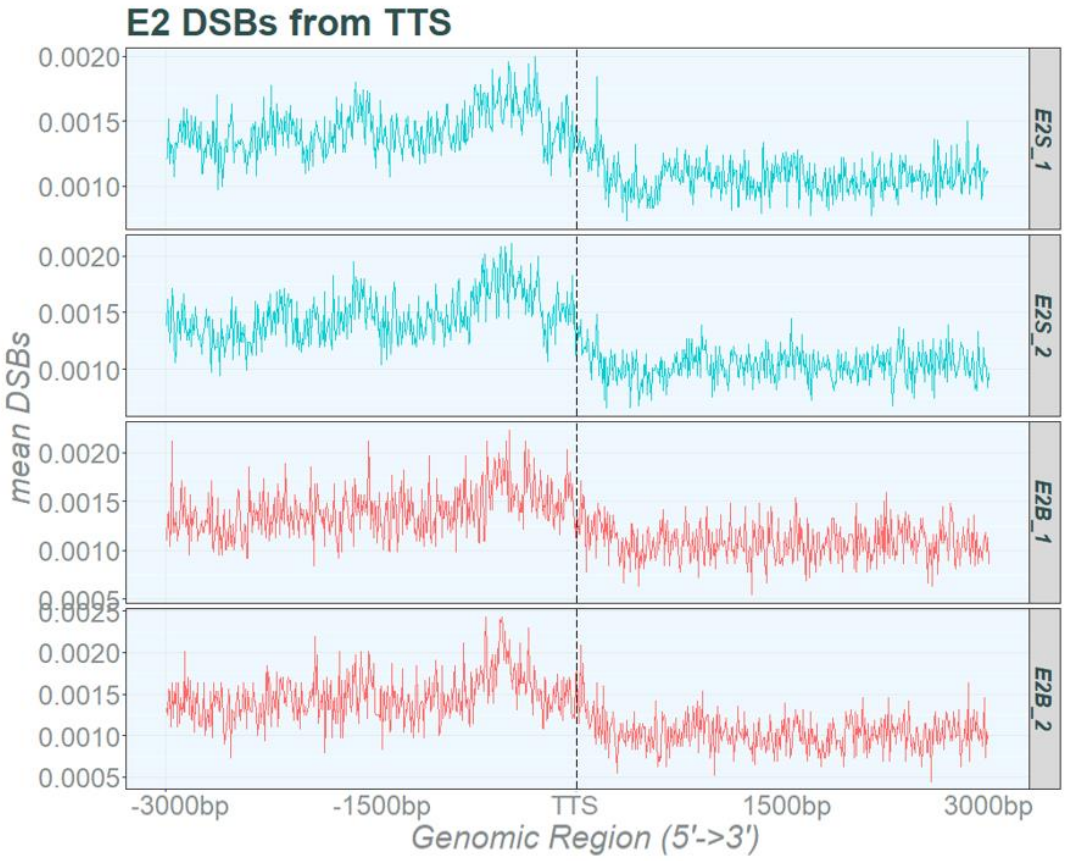
**Figure 5.4. Mean DSB frequency across TSS +/-3000 bp in neural cell lines NES, NPC and NEU and commercial cancer lines K562 and MCF7**

Mean DSBs across neural cells and commercial cells TSS +/- 3000 bp. Neural cells plotted in pink, K562 cells plotted in orange and MCF7 cells plotted in blue. Neural cells display individual repeat data. TSS are displayed as dashed central line. Region displays 3000 bp prior to and following TSS in a 5' to 3' direction. (a) Neural cells displayed in order top to bottom: NES (neuroepithelial stem cells in hot pink) repeats 1-2 nes_1 and nes_2, NPC (neural progenitor cells in dark pink) repeats 1-2 npc_1 and npc_2, NEU (post-mitotic neural cells in light pink) repeats 1-2 neu_1 and neu_2. (b) Commercial cancer lines displayed: K562 (erythroleukaemia in orange) single sample k562, MCF7 (breast cancer in blue) single sample mcf7.

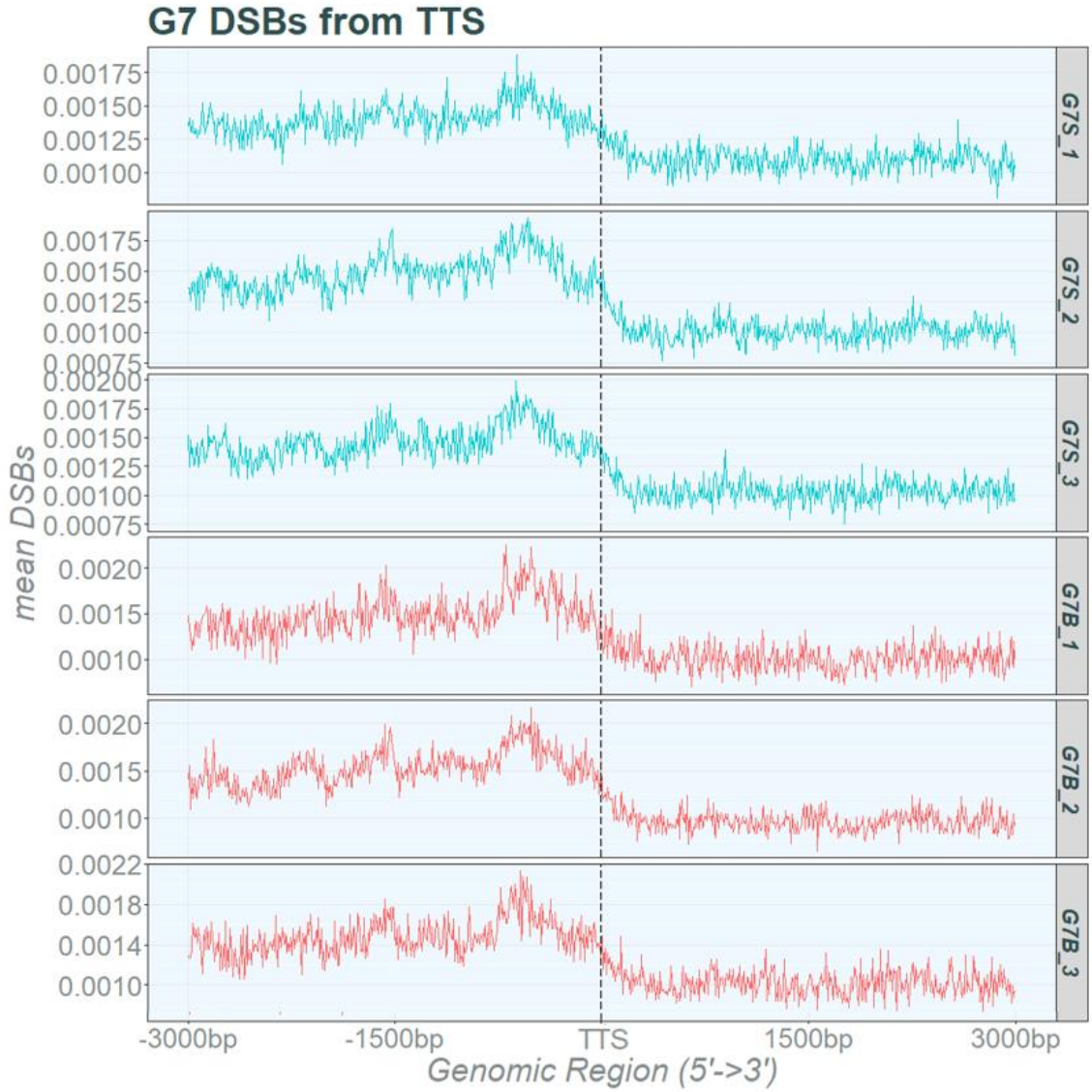
#### 5.3.1.4 DSB patterns at TTS demonstrate consistently higher mean DSB frequency across all GBM lines

Having identified a shared peak of mean DSB frequency around TTS in GBM lines in Figure 5.1 these were also investigated using absolute distance from TTS in Figure 5.5. Across GSCs and differentiated lines, there was a consistency in mean DSB frequency pattern at TTS. E2 GSCs and differentiated lines showed a peak in mean DSB frequency around 1000 bp before the TTS which then dropped off, crossing the TTS and dropped to below the mean starting DSB frequency (-3000 bp from TTS) following the TTS. In G7 this pattern was repeated where mean DSB frequency peaked at 1000 bp before the TTS and then had dropped to below the mean starting DSB frequency by 500 bp following TTS. Again, GSC and differentiated cell lines were consistent in pattern and across repeats. Finally, R10 again showed a similar pattern of increase in mean DSB frequency at 1000 bp prior to TTS, reaching a nadir at 500 bp following TTS. In R10, there was also a small dip followed by a small peak at <500 bp before the TTS which was consistent across the repeats and across GSC and differentiated lines. This was also partially visible across E2 and G7 lines, however the background noise DSBs made this more difficult to elucidate.

(a)


E2 DSBs from TTS
mean DSBs
0.0020
0.0015
0.0010
0.0005
E2S_1
E2S_2
E2B_1
E2B_2
-3000bp
-1500bp
TTS
1500bp
3000bp
Genomic Region (5'->3')

(b)


G7 DSBs from TTS
mean DSBs
G7S_1
G7S_2
G7S_3
G7B_1
G7B_2
G7B_3
-3000bp
-1500bp
TTS
1500bp
3000bp
Genomic Region (5'->3')

(c)

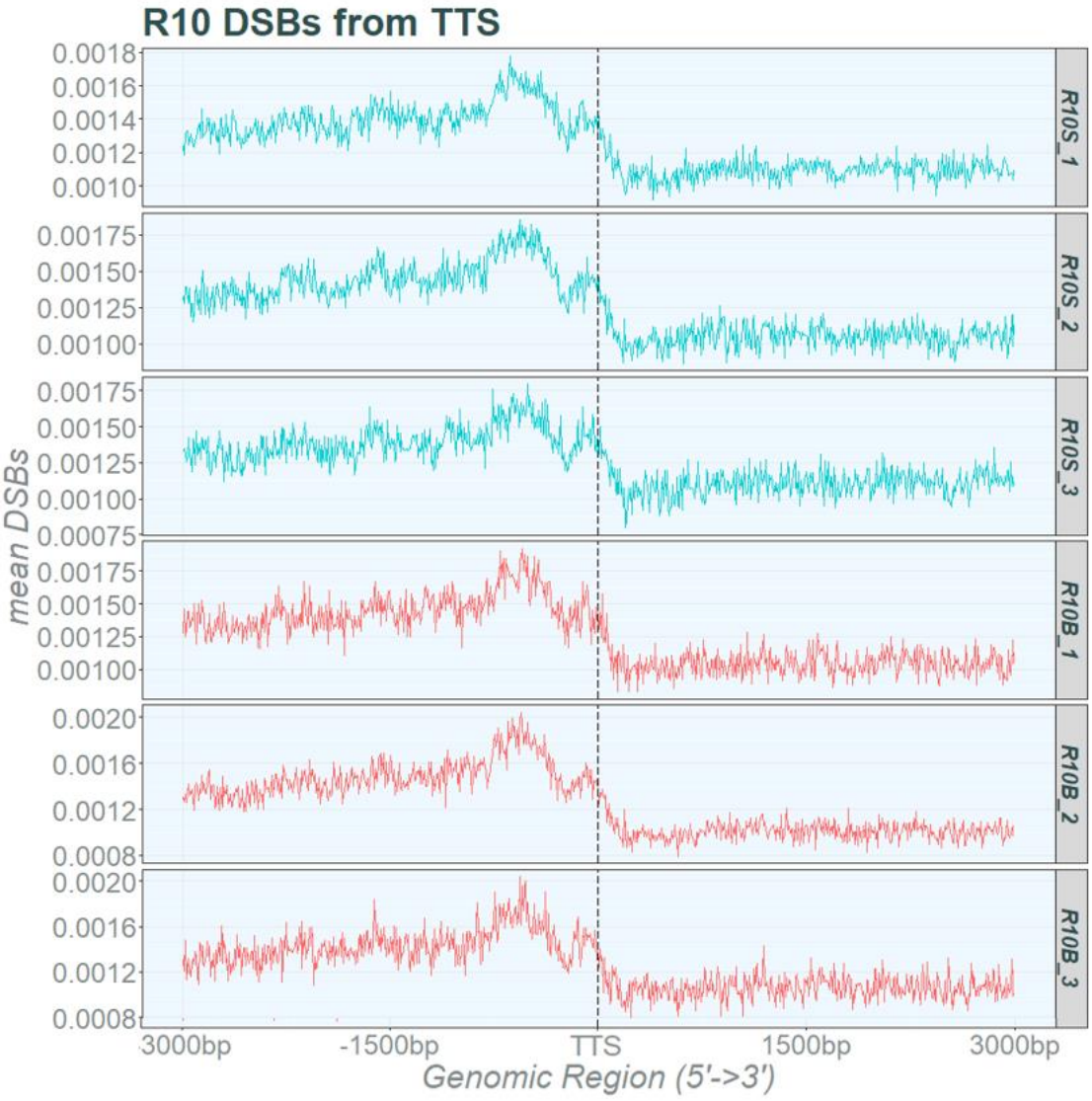



**Figure 5.5. Mean DSB frequency across TTS +/-3000 bp in GBM lines E2, G7 and R10**

Mean DSB frequency across GBM TTS in E2, G7 and R10 lines. GSCs in turquoise and differentiated cells in red. Individual repeats displayed. TTS displayed as dashed central line. Region displays 3000 bp prior to and following TTS in a 5' to 3' direction. (a) E2 GSCs, repeats 1-2 (E2S_1, E2S_2) and E2 differentiated cells, repeats 1-2 (E2B_1, E2B_2). (b) G7 GSCs, repeats 1-3 (G7S_1, G7S_2, G7S_3) and G7 differentiated cells, repeats 1-3 (G7B_1, G7B_2, G7B_3). (c) R10 GSCs, repeats 1-3 (R10S_1, R10S_2, R10S_3) and R10 differentiated cells, repeats 1-3 (R10B_1, R10B_2, R10B_3).

The mean DSB frequency across TTS was also investigated in neural cells and commercial cancer cell lines (Figure 5.6). Neural cell lines also showed a small but less pronounced peak at TTS regions. NES repeats (nes_1 and nes_2) displayed an increase in mean DSBs from 1000 bp before the TTS, peaking immediately prior to TTS and resolving within 1000 bp following TTS. The mean

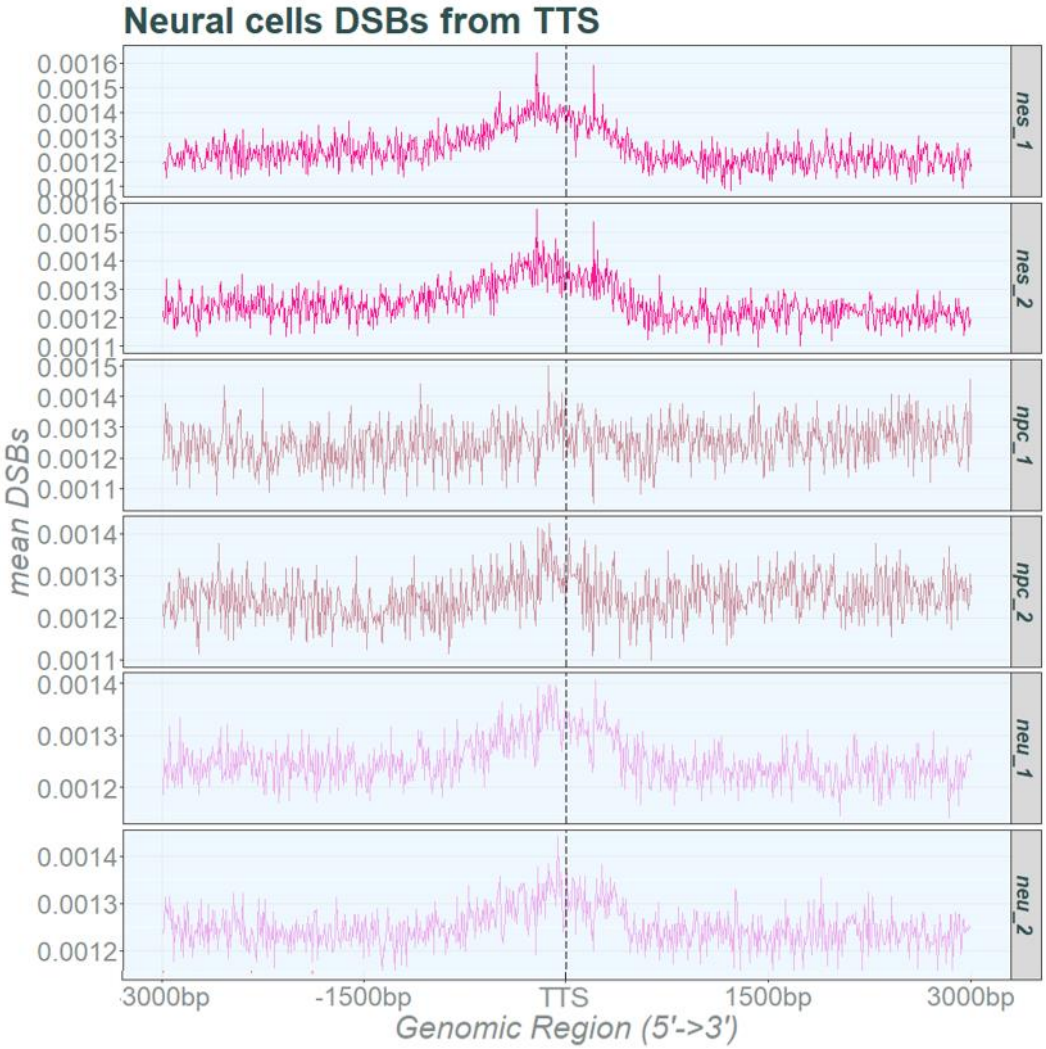
DSBs across TTS in NPC cells (npc_1 and npc_2) had a greater background variation, making mean DSB frequency patterns across TTS less distinct. The npc_1 repeat did not show a consistent increase of mean DSB frequency across TTS, however npc_2 appeared to show a small increase of mean DSB frequency at -500 bp before TTS, with a resolution to baseline <+500 bp after TTS. Finally, regarding neural cells, NEU lines, neu_1 and neu_2, both showed an increase in mean DSB frequency prior to TTS sites. This also resolved within +1000 bp following TTS.

The commercial cancer cell lines, K562 and MCF7, were plotted individually for mean DSB frequency across TTS. The mean DSB frequency of K562 did not display any obvious pattern of change across TTS regions. Mean DSB frequency across K562 varied from between 0.0010 to 0.0016 DSBs per base pair but broadly remained around 0.0012 per base pair across TTS. The mean DSB frequency prior to TTS within genes was approximately similar to mean DSB frequency following TTS outside of gene boundaries.

Regarding MCF7 cells, these demonstrated higher mean DSB frequency prior to TTS with a decrease within the -500 bp prior to TTS. Rather than an increase in mean DSB frequency prior to TTS, MCF7 cells showed a drop in baseline from before TTS to after TTS. Mean DSB frequency at -3000 bp to -500 bp prior to TTS was broadly higher (0.0013 per base pair) than 0 to +3000 bp after TTS (0.0012 per base pair).

Across all the cell lines, mean DSB frequency patterns of neural cells at TTS were most similar to GBM lines, showing an increase in mean DSB frequency before TTS, with a decrease following TTS. The difference between mean DSB frequency at -3000 bp before TTS and +3000 bp after TTS varied; GBM lines showed a decrease outside of gene boundaries, whereas neural cells were broadly similar at this level. Altogether, whilst GBM cells and commercial cell lines all demonstrated some evidence of a peak of DSBs at TTS, this was consistently visually clearer in GBM lines highlighting that DSBs in GBM may be disproportionately higher at gene end sites than in other cell lines.

(a)


Neural cells DSBs from TTS
mean DSBs
nes_1
nes_2
npc_1
npc_2
neu_1
neu_2
3000bp
-1500bp
TTS
1500bp
3000bp
Genomic Region (5'->3')

(b)

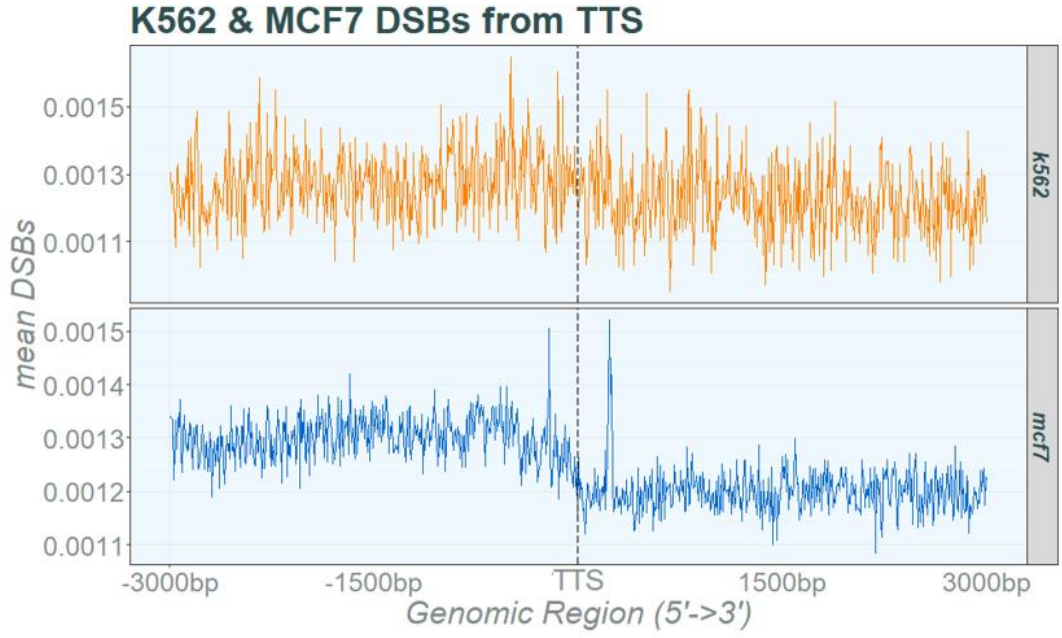

K562 & MCF7 DSBs from TTS
mean DSBs
k562
mcf7
-3000bp
-1500bp
TTS
1500bp
3000bp
Genomic Region (5'->3')

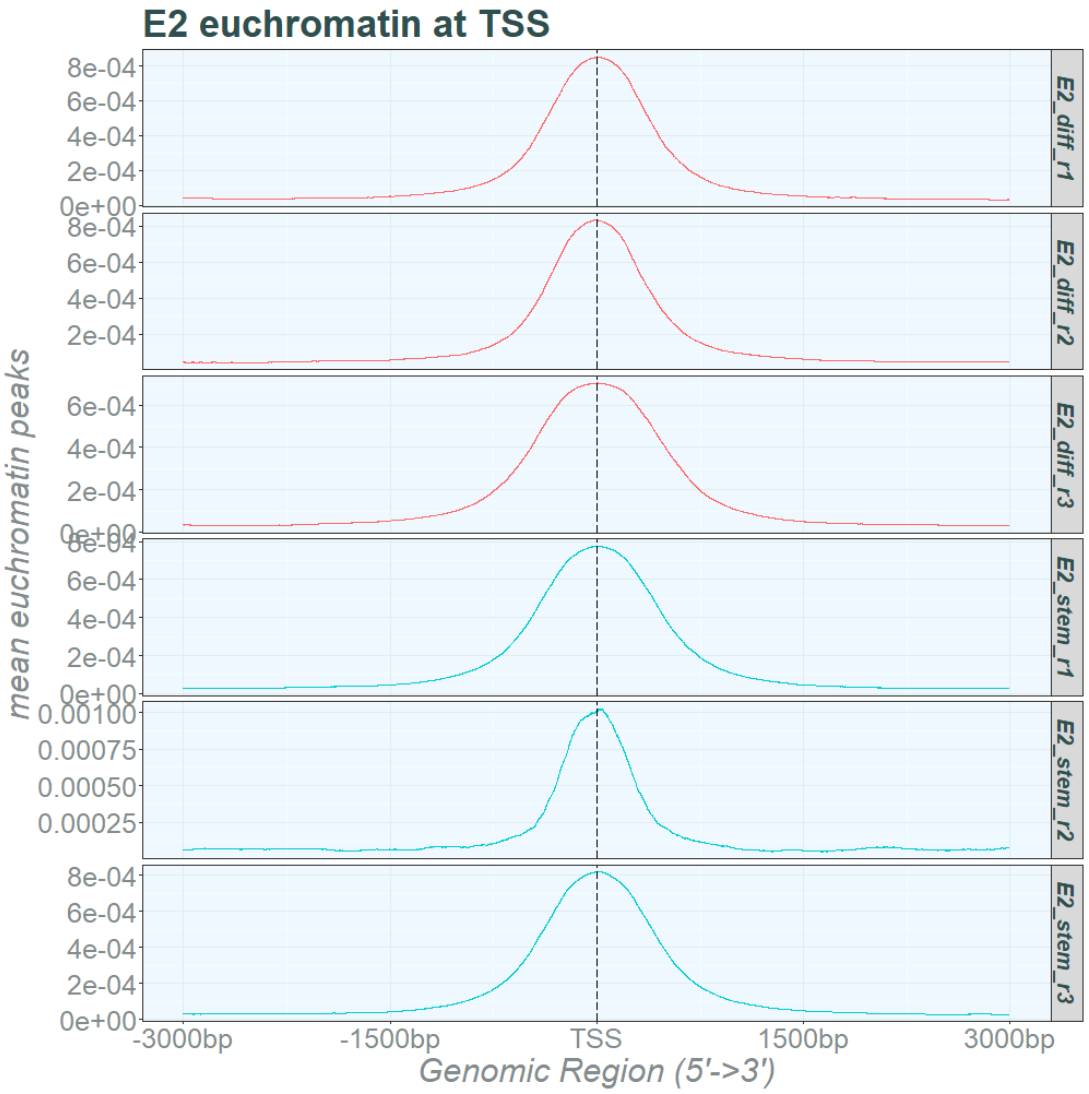
**Figure 5.6. Mean DSB frequency across TTS +/-3000 bp in neural cell lines NES, NPC and NEU and commercial cancer lines K562 and MCF7**

Mean DSB frequency across neural cells and commercial cells TTS. Neural cells plotted in pink, K562 cells plotted in orange and MCF7 cells plotted in blue. Individual repeats for neural cells displayed. TTS are displayed as dashed central line. Region displays 3000 bp prior to and following TTS in a 5' to 3' direction. (a) Neural cells displayed in order top to bottom: NES (neuroepithelial stem cells in hot pink) repeats 1-2 nes_1 and nes_2, NPC (neural progenitor cells in dark pink) repeats 1-2 npc_1 and npc_2, NEU (post-mitotic neural cells in light pink) repeats 1-2 neu_1 and neu_2. (b) Commercial cancer lines displayed: K562 (erythroleukaemia in orange) single sample k562, MCF7 (breast cancer in blue) single sample mcf7.

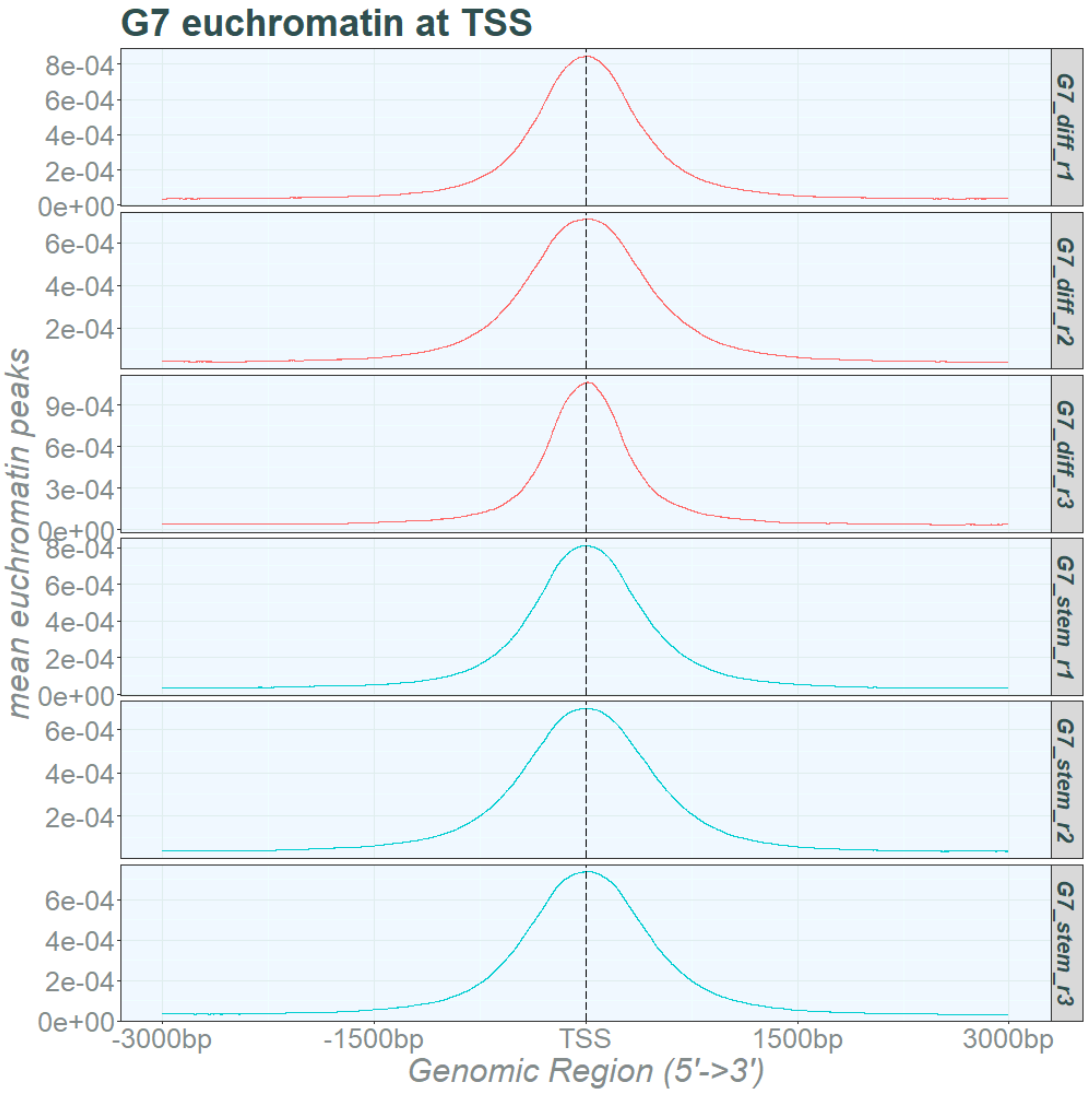
#### 5.3.1.5 Euchromatin enrichment across TSS and TTS locations is uniform across GBM lines

Having mapped DSBs across TSS and TTS regions in GBM lines, euchromatin profiles across TSS and TTS sites were also profiled for reference and to determine patterns of euchromatin enrichment across these sites. Figure 5.7 displays mean euchromatin peaks across TSS for GSC and differentiated cell GBM lines E2, G7 and R10. Euchromatin is well known to increase in abundance across TSS due to the need for highly accessible sites at these locations for transcriptional activity. This was reflected in Figure 5.7. Across both differentiated and GSC lines there was a clear increase in mean euchromatin peak frequency across TSS, starting from -1000 bp prior to TSS, peaking at TSS and then decreasing back to the baseline at +1000 bp after TSS. This pattern was present in all three GBM lines E2, G7 and R10 and was in line with ATAC-seq data in many other cell types. Unlike mean DSB frequency across TSS, there were no differences between mean euchromatin pattern across TSS in GBM cell lines. Though DSB frequency was highly variable across TSS in GBM lines, this data indicated that euchromatin enrichment patterns across TSS were largely uniform and did not necessarily follow DSB patterns at TSS.

(a)


E2 euchromatin at TSS
mean euchromatin peaks
E2_diff_r1
E2_diff_r2
E2_diff_r3
E2_stem_r1
E2_stem_r2
E2_stem_r3
8e-04
6e-04
4e-04
2e-04
0e+00
0.00100
0.00075
0.00050
0.00025
-3000bp
-1500bp
TSS
1500bp
3000bp
Genomic Region (5'->3')

(b)


G7 euchromatin at TSS
mean euchromatin peaks
8e-04
6e-04
4e-04
2e-04
0e+00
9e-04
3e-04
G7_diff_r1
G7_diff_r2
G7_diff_r3
G7_stem_r1
G7_stem_r2
G7_stem_r3
-3000bp
-1500bp
TSS
1500bp
3000bp
Genomic Region (5'->3')

(c)

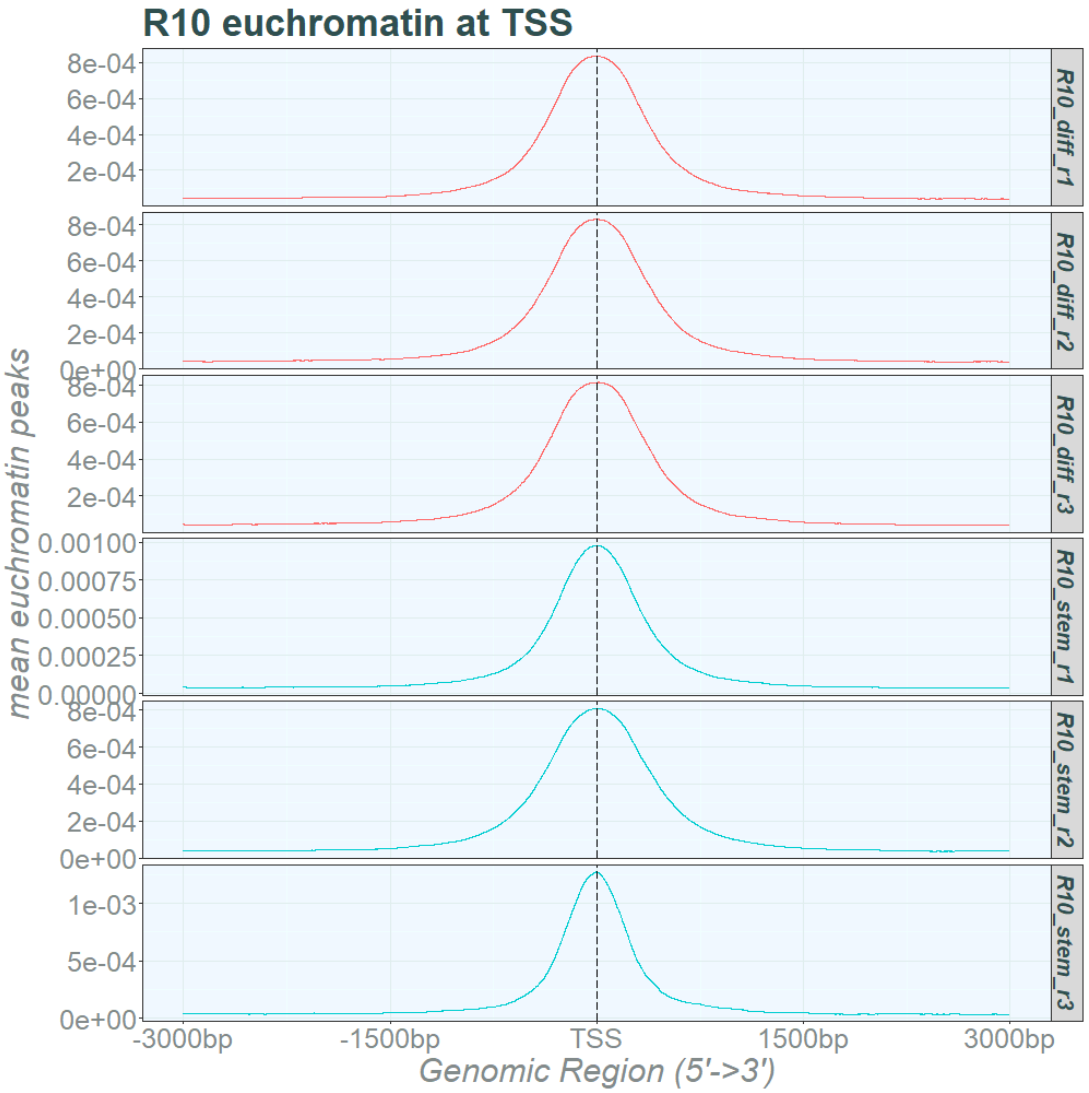


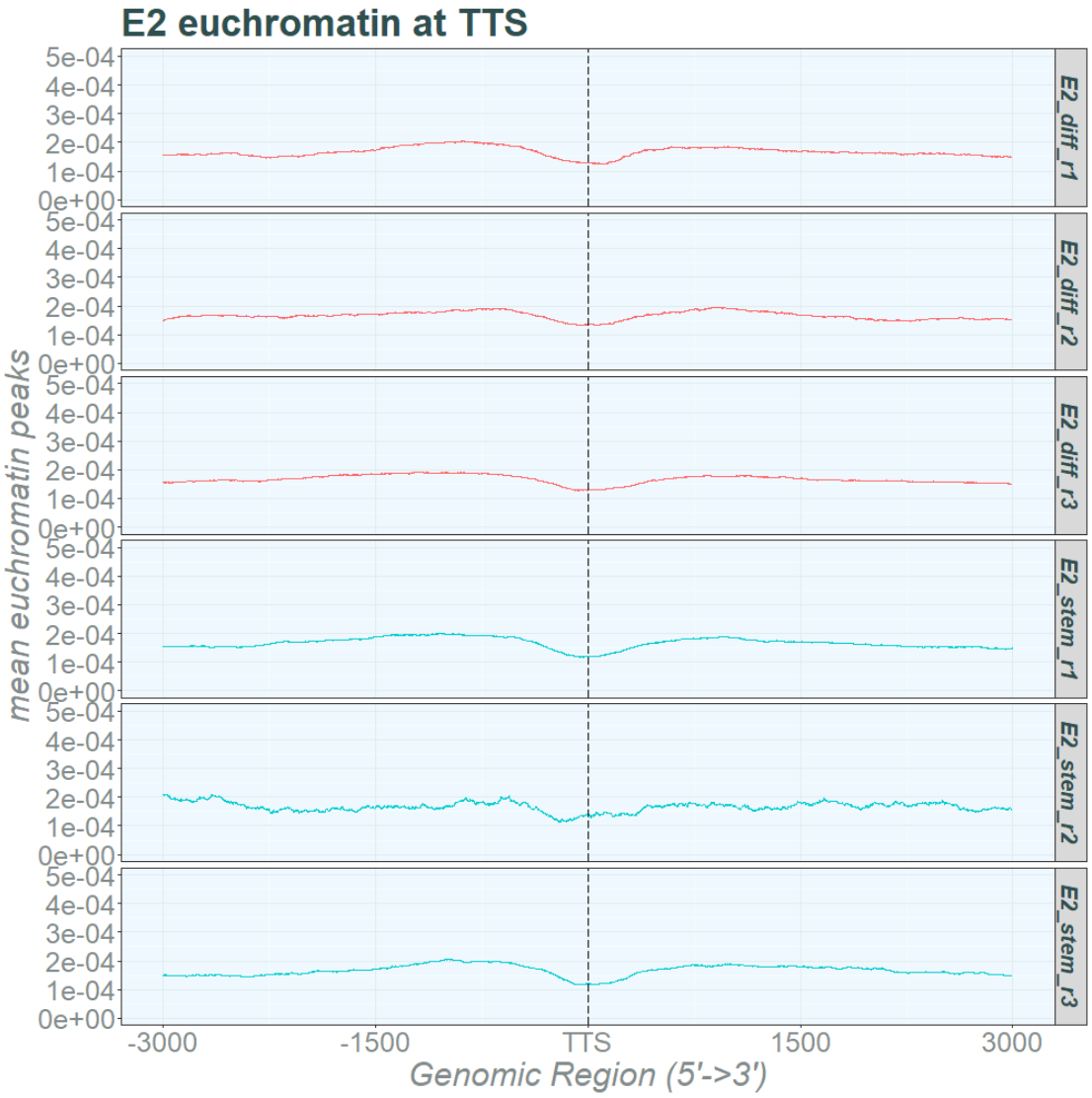

**Figure 5.7. Mean euchromatin peak frequency across TSS for GBM lines E2, G7 and R10**

Mean euchromatin peak frequency across GBM TSS in E2, G7 and R10 lines. differentiated cells in red (top) and GSCs in turquoise (bottom). Individual repeats displayed. TSS are displayed as dashed central line. Region displays 3000 bp prior to and following TSS in a 5' to 3' direction. (a) and E2 differentiated cells, repeats 1-3 (E2_diff_r1, E2_ diff_r2, E2_ diff_r3) and E2 GSCs, repeats 1-3 (E2_stem_r1, E2_stem_r2, E2_stem_r3). (b) G7 differentiated cells, repeats 1-3 (G7_diff_r1, G7_ diff_r2, G7_ diff_r3) and G7 GSCs, repeats 1-3 (G7_stem_r1, G7_ stem _r2, G7_ stem _r3). (c) R10 differentiated cells, repeats 1-3 (R10 _diff_r1, R10 _ diff_r2, R10 _ diff_r3) and R10 GSCs, repeats 1-3 (R10 _stem_r1, R10 _ stem _r2, R10 _ stem _r3).

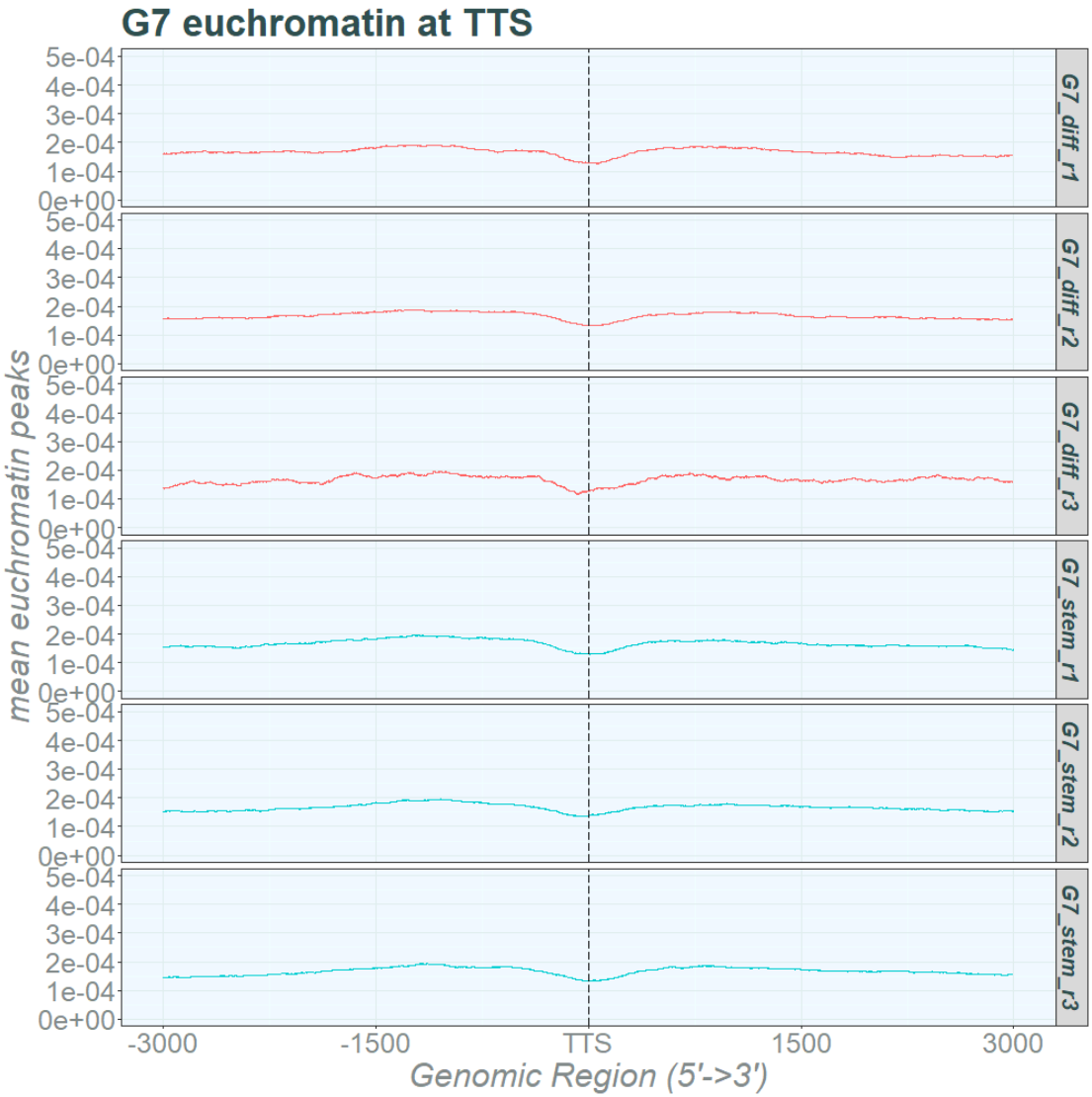
The pattern of euchromatin enrichment across TTS was also studied (Figure 5.8). Conversely to TSS, TTS did not demonstrate the same pattern of mean euchromatin peak frequency as displayed at TSS. Rather, mean euchromatin peak frequency across TTS showed a subtle decrease in frequency around TTS

across all GSCs and differentiated cells. Where DSBs at TSS demonstrated a consistent increase, the mean euchromatin peak frequency conversely demonstrated a small consistent decrease. This consistent flattening and small decrease in euchromatin peaks across TTS locations indicated that DSBs did not appear to consistently follow euchromatin enrichment patterns at TTS. Taken together with the mean DSB patterns, euchromatin did not directly reflect DSB frequency across either TSS or TTS locations, indicating that DSBs were not dependent on euchromatin enrichment.

(a)



(b)


G7 euchromatin at TTS
mean euchromatin peaks
5e-04
4e-04
3e-04
2e-04
1e-04
0e+00
G7_diff_r1
G7_diff_r2
G7_diff_r3
G7_stem_r1
G7_stem_r2
G7_stem_r3
-3000
-1500
TTS
1500
3000
Genomic Region (5'->3')

(c)

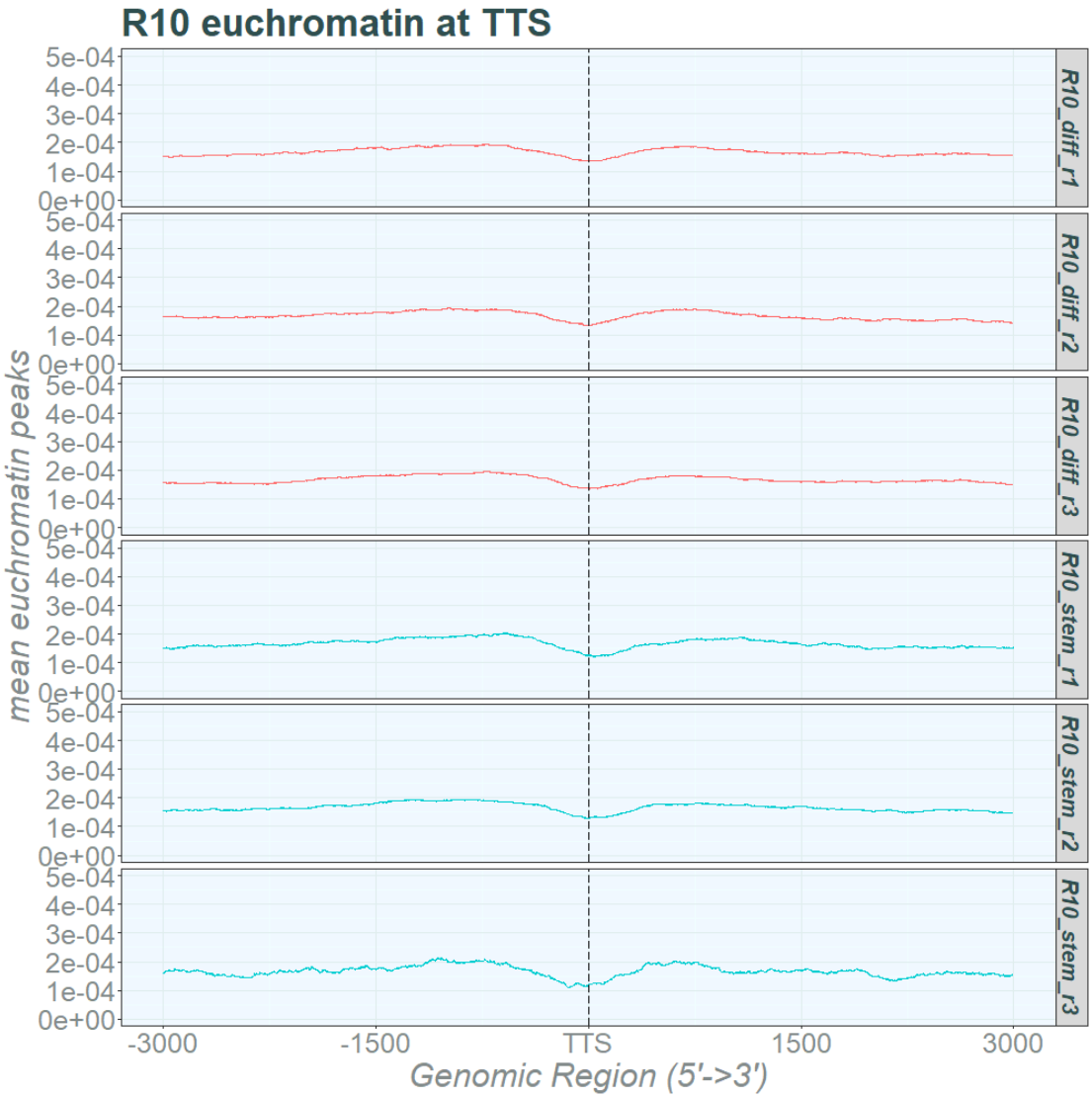



**Figure 5.8. Mean euchromatin peak frequency across TTS for GBM lines E2, G7 and R10**

Mean euchromatin peak frequency across GBM TSS in E2, G7 and R10 lines. differentiated cells in red (top) and GSCs in turquoise (bottom). Individual repeats displayed. TTS are displayed as dashed central line. Region displays 3000 bp prior to and following TTS in a 5' to 3' direction. (a) E2 differentiated cells, repeats 1-3 (E2_diff_r1, E2_ diff_r2, E2_ diff_r3) and E2 GSCs, repeats 1-3 (E2_stem_r1, E2_stem_r2, E2_stem_r3). (b) G7 differentiated cells, repeats 1-3 (G7_diff_r1, G7_ diff_r2, G7_ diff_r3) and G7 GSCs, repeats 1-3 (G7_stem_r1, G7_ stem _r2, G7_ stem _r3). (c) R10 differentiated cells, repeats 1-3 (R10 _diff_r1, R10 _ diff_r2, R10 _ diff_r3) and R10 GSCs, repeats 1-3 (R10 _stem_r1, R10 _ stem _r2, R10 _ stem _r3).

## 5.3.2 DSBs and genomic sites of interest

Having identified that there were differences across mean DSB frequency patterns across TSS and TTS in GBM lines compared to neural cells and commercial cancer lines, these were investigated in more detail. For this, DSBs

were annotated for their position across the genome. DSBs were annotated under the following locations:1-5 kbp from TSS, promotor <1 kbp from TSS, 5' UTRs, exons, intron/exon boundaries, introns, 3' UTRs and finally, intergenic regions and then total DSBs per annotated site calculated. A randomly generated DSB map based on the DSB totals and genome locations was calculated for comparison with actual DSB frequency per annotated region.

#### 5.3.2.1 GBM lines have a greater than predicted DSB frequency at genic locations

Figure 5.9 displays actual versus randomly generated DSB distribution across E2, G7 and R10 GSCs and differentiated cells for repeat 1 as a bar graph representation of DSB spread across the genome. Bar graphs for E2 repeat 2 and G7 and R10 repeats 2 and 3 are available in supplementary figures. Total DSBs across annotated sites were higher across all sites except from intergenic regions. Actual DSB frequency was higher in genic regions in both GSC and differentiated cells across all lines E2, G7 and R10. Overall, 3' UTR locations demonstrated a greater than 2-fold increase in actual DSBs compared to predicted DSBs, consistent with the peak of mean DSBs seen across TTS in GBM lines. Furthermore, GBM lines had a lower than predicted DSB frequency at intergenic locations. However, there were not clear distinctions in DSB distribution between GSCs and differentiated progeny.

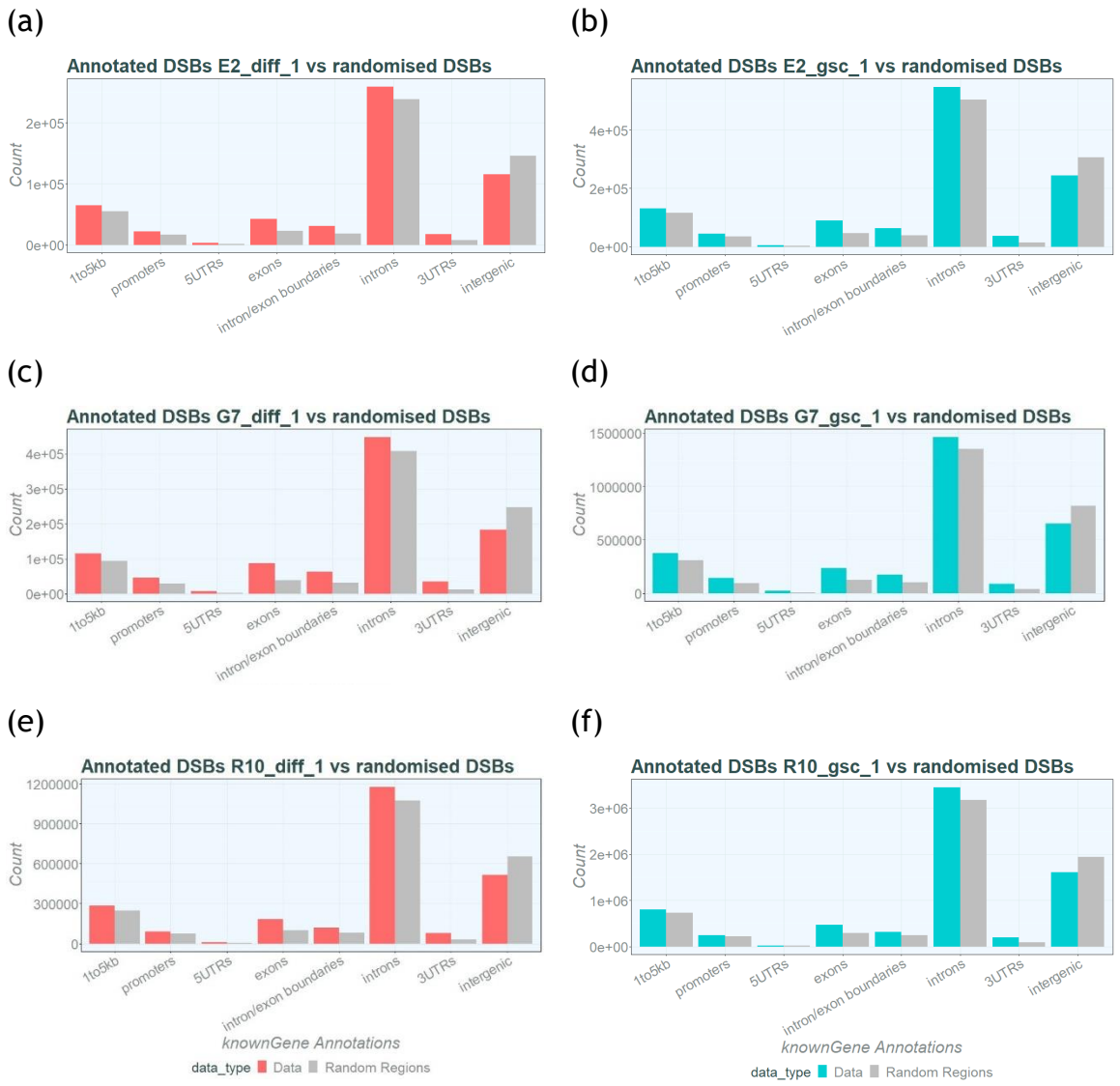



**Figure 5.9: Annotated DSB locations vs random expected DSB locations in GBM cells**

Actual versus expected DSBs per annotated genomic region in GSCs and differentiated cells replicate 1. Actual DSBs in turquoise (GSCs) or red (differentiated cells), expected DSBs in grey. Order: regions between 1-5 kbp from transcription start sites, promotors <1 kbp from transcription start site, 5' UTRs, exons, intro/exon boundaries, introns, 3' UTRs and intergenic regions. For overlapping annotation site priority: Promoters, 1 to 5 kbp promoters, 5' UTRs, 3' UTRs, exons, introns, intron/exon boundaries, and intergenic sites. (a) E2 differentiated rep 1. (b) E2 GSC rep 1. (c) G7 differentiated rep 1. (d) G7 GSC rep 1. (e) R10 differentiated rep 1. (f) R10 GSC rep 1.

**Table 5.2. Expected vs actual DSBs: mean fold change in GBM lines E2, G7 and R10**

| | Mean E2 diff | Mean E2 GSC | Mean G7 diff | Mean G7 GSC | Mean R10 diff | Mean R10 GSC |
|---|---|---|---|---|---|---|
| ***1 to 5 kbp from TSS*** | 1.168 | 1.128 | 1.270 | 1.224 | 1.139 | 1.102 |
| ***3' UTRs*** | 2.607 | 2.618 | 3.154 | 2.628 | 2.547 | 2.131 |
| ***5' UTRs*** | 1.745 | 1.746 | 2.737 | 2.591 | 1.329 | 1.234 |
| ***Exons*** | 1.973 | 1.956 | 2.463 | 2.119 | 1.856 | 1.619 |
| ***Intergenic*** | 0.781 | 0.790 | 0.711 | 0.756 | 0.789 | 0.826 |
| ***Intron/exon boundaries*** | 1.725 | 1.690 | 2.152 | 1.881 | 1.469 | 1.328 |
| ***Introns*** | 1.088 | 1.088 | 1.108 | 1.096 | 1.094 | 1.083 |
| ***Promoter <1 kbp to TSS*** | 1.297 | 1.224 | 1.609 | 1.533 | 1.177 | 1.107 |

Total DSBs per annotated site of interest: 1 to 5 kbp from TSS, 3' UTRs, 5' UTRs, exons, intergenic sites, intron/exon boundaries, introns and promoters <1 kbp from TSS. Expected DSB frequency calculated per repeat generated by creating randomly distributed DSB locations from original file. Expected DSBs normalised to a value of 1 per site of interest. Actual DSB distribution represented as a fold-change of the expected values. Fold-change per site of interest reported for all repeats and a mean fold change across differentiated cells and GSCs for E2, G7 and R10 reported above.

#### 5.3.2.2 GBM lines have a relatively higher proportion of DSBs at exons and 3' UTRs than neural cells

Having identified an increase in DSB frequency compared to expected DSBs in GBM lines this was investigated in other cell types to identify differences and similarities in relative DSB frequency.

To determine any shared or divergent patterns in DSB distribution across neural cells and GSCs, annotated sites were plotted as a proportion of DSBs as shown in **Error! Reference source not found.** Both neural and GSCs demonstrated that the greatest proportion of DSBs was occurring in introns, followed by intergenic sites. GSCs showed a higher proportion of DSBs occurring in exons and 3' UTRs compared to neural cells. Neural cells demonstrated a greater proportion of DSBs occurring within intronic regions relative to GSCs, with GSCs broadly showing higher break proportions in 5' UTRs, exons and 3' UTRs compared to neural cells. Neural cells demonstrated a higher proportion of DSBs occurring within intergenic regions for GSCs.

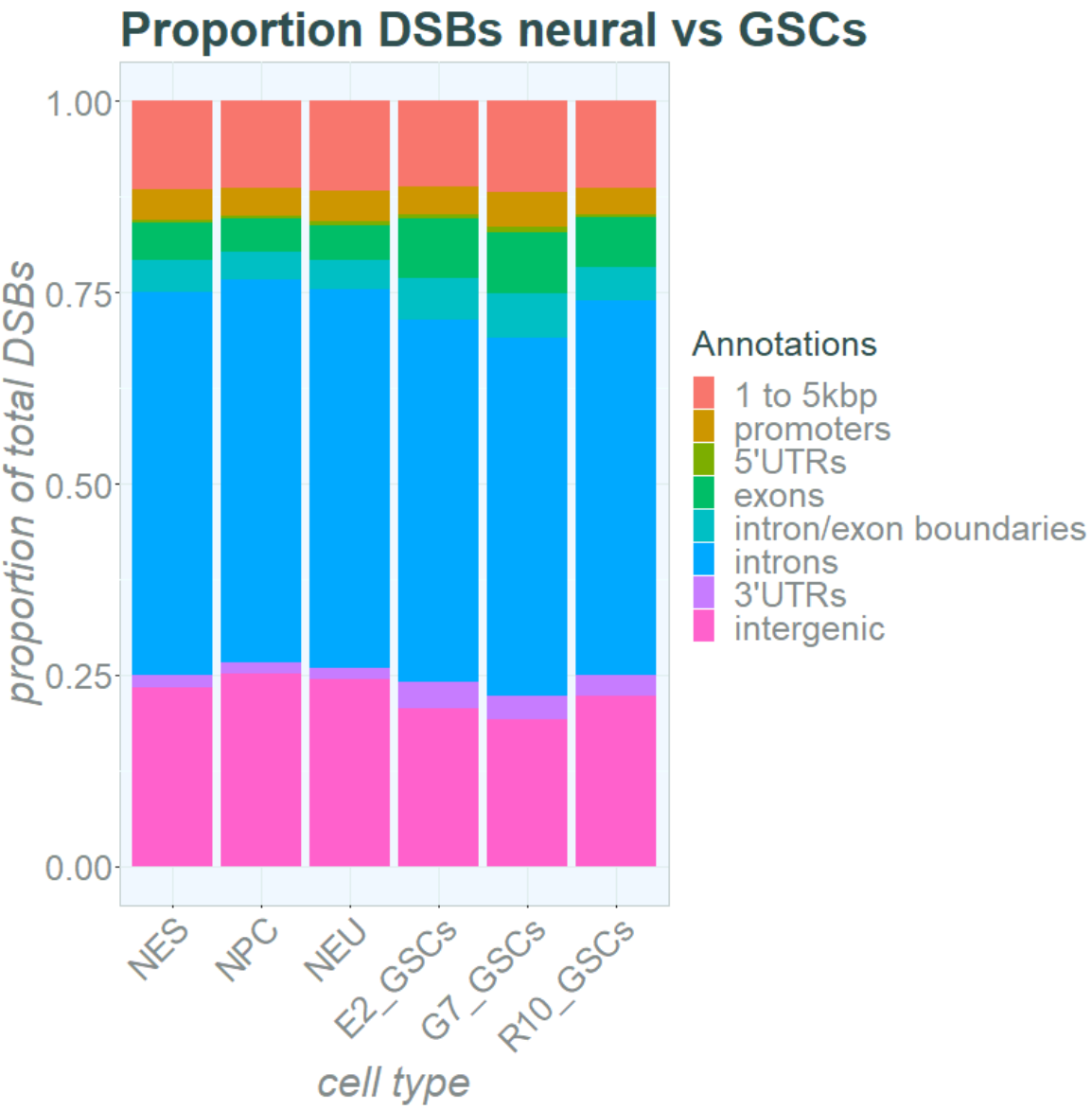



**Figure 5.10. Distribution of DSBs across neural cells and GSCs**

Distribution of genomically annotated DSBs across neural cell lines and GSC lines as a proportion of total DSBs per sample group. Columns 1-3 represent neural lines NES, NEU and NPC. Columns 4-6 represent GSCs E2S, G7S and R10S. Annotations are from top to bottom as follows: regions between 1-5 kbp from TSS, promotors <1 kbp from TSS, 5' UTRs, exons, intron/exon boundaries, introns, 3' UTRs and intergenic regions.

DSB frequency across the described annotated regions was also calculated for neural cells and commercial cancer lines. Expected DSBs were generated for reference comparison in these cells also.

As a means of comparison of DSB frequency distribution across annotated sites of interest, the DSB frequency across GSCs, neural cells and commercial cancer cell lines were normalised to individual expected DSB frequency. Therefore, the DSB frequency across differing cell types was represented as a fold-change of the

expected DSB distribution. The expected versus actual DSB frequency is represented in Table *5.3*. There was a global decrease in fold-change compared to expected DSBs in intergenic sites across GSC lines, commercial cell lines and neural cells. The GSCs showed the greatest increase in fold-change in 3' UTRs (2.513) and at exons (1.947) compared to commercial cancer cells and neural cells. The commercial cancer line K562 showed the highest fold-change increase at 5' UTRs (2.610). Though there was a fold increase across the 3' UTR regions from predicted in the cancer cell lines and neural cell lines of between 1.09 and 1.49, with the cancer cell lines showing a higher fold change than neural cells. This was a considerably lower fold change than was seen in GSC lines. GSC lines, neural cells and commercial cancer cells all showed a decrease in fold change across intergenic sites.

**Table 5.3. Expected vs actual DSBs in K562, MCF7 and means of neural cell lines**

| | All GSCs means | K562 | MCF7 | Mean NES | Mean NPC | Mean NEU |
|---|---|---|---|---|---|---|
| ***1 to 5 kbp from TSS*** | 1.182 | 1.291 | 1.207 | 1.158 | 1.105 | 1.168 |
| ***3' UTRs*** | 2.513 | 1.401 | 1.487 | 1.294 | 1.086 | 1.1465 |
| ***5' UTRs*** | 1.908 | 2.610 | 1.763 | 1.446 | 1.158 | 1.4915 |
| ***Exons*** | 1.947 | 1.558 | 1.484 | 1.221 | 1.064 | 1.144 |
| ***Intergenic*** | 0.760 | 0.793 | 0.788 | 0.810 | 0.861 | 0.847 |
| ***Intron/exon boundaries*** | 1.665 | 1.541 | 1.439 | 1.232 | 1.042 | 1.134 |
| ***Introns*** | 1.113 | 1.092 | 1.115 | 1.115 | 1.090 | 1.094 |
| ***Promoter <1 kbp to TSS*** | 1.324 | 1.674 | 1.354 | 1.266 | 1.122 | 1.305 |

Total DSBs per annotated site of interest: 1 to 5 kbp from TSS, 3' UTRs, 5' UTRs, exons, intergenic sites, intron/exon boundaries, introns and promoters <1 kbp from TSS. Expected DSB frequency calculated per repeat generated by creating randomly distributed DSB locations from original file. Expected DSBs normalised to a value of 1 per site of interest. Actual DSB distribution represented as a fold-change of the expected values. Fold-change per site of interest reported for all repeats and a mean fold change across GSCs, NES, NPC and NEU are reported above. For cell lines MCF7 and K562, the fold change for the single repeat is displayed.

Relative DSB frequency across GSCs and neural cells was compared using fold change of the actual versus expected DSB frequency. Table *5.4* displays the mean fold-change of DSBs across neural cell lines compared to GSC cell lines. GSCs showed a significantly higher fold-change of DSBs from expected in 3' UTRs

compared to neural cells ($p<0.001$) with neural cells displaying a mean fold-change from expected of 1.176 and GSCs displaying a mean fold-change from expected to actual DSBs of 2.513. GSCs also displayed a significantly greater increase in fold-change from expected to actual compared to neural cell fold-change at DSBs in exons and at DSBs in intron/exon boundaries. Conversely, whilst both neural cells and GSCs showed a fold change reduction with a lower than expected frequency of DSBs at intergenic regions there was a significantly greater relative decrease in GSCs compared to neural cells. These findings confirmed the consistent difference seen at 3' UTRs between GBM lines and neural cells of a relatively greater proportion of DSBs occurring at gene end sites compared to neural cells.

Overall, GSCs demonstrated a consistent increase in DSBs at TTS which was greater than that of both neural cells and the two commercial cancer cell lines. Across all lines, intergenic regions had lower than predicted DSBs, with GSCs also being significantly lower than neural cells.

**Table 5.4. Expected vs actual DSBs in GSC lines**

| | All neural cells means | All GSCs means | Adjusted p-value |
|---|---|---|---|
| ***1 to 5kbp from TSS*** | 1.144 | 1.182 | 0.174 |
| ***3' UTRs*** | 1.176 | 2.513 | <0.001 |
| ***5' UTRs*** | 1.365 | 1.908 | 0.0515 |
| ***Exons*** | 1.143 | 1.947 | <0.001 |
| ***Intergenic*** | 0.840 | 0.760 | <0.001 |
| ***Intron/exon boundaries*** | 1.136 | 1.665 | 0.001 |
| ***Introns*** | 1.100 | 1.113 | 0.064 |
| ***Promoter <1kbp to TSS*** | 1.237 | 1.324 | 0.315 |

Total DSBs per annotated site of interest: 1 to 5 kbp from TSS, 3' UTRs, 5' UTRs, exons, intergenic sites, intron/exon boundaries, introns and promoters <1 kbp from TSS. Expected DSB frequency calculated per repeat generated by creating randomly distributed DSB locations from original file. Expected DSBs normalised to a value of 1 per site of interest. Actual DSB distribution represented as a fold-change of the expected values. Fold-change per site of interest reported for all repeats and a mean fold change across neural cells and GSCs is reported above. Fold-change of neural cells and fold-changes of GSCs was statistically tested. Significance testing by t-test with BHP correction.

## 5.4 Discussion and conclusions

This chapter has investigated mean DSB patterns and frequency within genes and gene bodies in GBM lines, having identified genes as key sites of DSBs within the genome. Mean DSBs across genes were mapped, demonstrating TSS and TTS as areas of interest. Sites within genes were then investigated with reference to non-GBM cell types and expected DSB frequency distribution across the genome.

### 5.4.1 DSBs across genes

Observing mean DSBs across genes in GBM GSC and differentiated cells demonstrated considerable similarities in patterns at TTS but some key differences in DSB patterns at TSS regions. The DSB pattern between E2, G7 and R10 GBM lines varied considerably across TSS regions (Figure 5.1). Even when accounting for absolute distance from TSS, these differences were maintained (Figure 5.3). Interestingly, neural cells and commercial cancer cell lines had an apparently similar pattern of mean DSBs across gene TSS regions (Figure 5.2), with a peak at the TSS across both groups, reminiscent of G7 lines. However neural cells had some differences in appearance when mapped to the absolute distance across TSS of -/+3000 bp (Figure 5.4). Curiously, neural cells seemed to show a dip immediately surrounding the TSS which was concealed within the mean DSBs mapped across the whole gene body. Interestingly, K562 cells and G7 cells, which demonstrated the peak across TSS, showed a small dip immediately surrounding the TSS. Looking at E2, it was difficult to establish any true pattern of mean DSBs across TSS. In contrast, R10 demonstrated a dip in mean DSBs across TSS regions which was present at both the gene body and TSS-specific level (Figure 5.1, Figure 5.3).

Previous studies in DSB mapping have shown an increase in DSBs at TSS in transcribed genes of neural stem/progenitor cells, consistent with the results seen in G7 GBM lines, the neural cell lines and commercial cancer cell lines (Singh et al., 2020). This difference of mean DSB pattern across GBM lines may be reflective of the heterogeneity of GBM and therefore represent differences across transcription and chromatin. The variation of TSS DSB patterns in GBM lines could also be related to alternative TSS, where the TSS for a gene can differ depending on the cell or tissue (Reyes and Huber, 2018). This alternative

TSS will not necessarily be recognised within the reference genome and therefore would not have been accounted for during annotation of TSS regions. However, whilst this could explain the lack of a clear peak at TSS in E2, it does not fully justify the dip in DSBs seen in R10. It is possible that there is an interplay between alternative TSS and nucleosome positioning. It has been observed in other cells that proximity to nucleosomes has been associated with lower DSBs following exposure to IR (Brambilla et al., 2020). Given that nucleosome positioning is dependent on transcriptional expression, it is possible that the "classical" reference genome TSS are more proximal to nucleosomes than alternative TSS in R10. Therefore, if R10 were utilising alternative TSS, a peak in DSBs may not be identifiable at the reference genome TSS. Similarly, if the reference genome TSS was more proximal to a nucleosome it is possible that there could be a decrease in DSBs.

Looking in more detail at TSS, the mean DSB patterns across absolute TSS distance showed in some cell lines a small dip immediately surrounding TSS despite an overall increase in mean DSBs around TSS (neural cells, K562 and G7). It is possible that this may reflect a technical part of the BLISS assay: the trimming step in BLISS where DSBs must be blunted in order for dsDNA tags to attach. Complex DSBs at TSS regions may be more prone to requiring trimming for blunting of DSBs, especially if there are larger overhangs or an excess of transcriptional machinery. However, it is interesting that the size of DSB dip surrounding TSS differs across cell types. This may suggest that technical aspects of the BLISS assay do not solely account for this finding. Rather, this may reflect the biological difference such as changes across cell types in gene transcription and in nucleosome positioning as discussed above.

The mean DSB frequency plotted across TTS in GBM lines displayed a consistent increase, both visible across the gene body and when mapped at the absolute distance across TTS. Interestingly, this appeared to be far more obvious in GBM lines than in any other cell type. This clear, consistent peak of mean DSBs at TTS regions in all GBM lines, suggested some correlation with DSBs at the end of gene transcription in these cells. A possible cause of this peak could be the action of topoisomerase acting on negative supercoiling to prevent torsional stress (Kenig et al., 2016). TopIIß has been cited as upregulated in GSCs and interestingly a potential mediator of GSC treatment resistance (Kenig et al., 2016).

Additionally, TopIIß has also been shown to display higher levels of expression relative to NSCs. Interestingly and in part contradiction, other topoisomerases such as TopI have been considered effective players in alleviating RS, despite the fact that RS is known to be elevated in GBM (Promonet et al., 2020, Carruthers et al., 2018). Neither neural cells nor commercial cancer cell lines demonstrated as obvious a peak of DSBs at the TTS, implying that this was a particular finding specific to these glioma lines in the first instance.

### 5.4.2 Annotated genomic sites

Given the observed mean DSBs across genes at TSS and TTS, it was of interest to investigate DSBs across other genomic sites also. Having mapped actual verses expected DSB frequency across GBM, neural cells and commercial cell lines, it became apparent that there was a lower than expected DSB frequency in intergenic sites across all cell types. There may be some considerations for this; one of which is that DSBs in heterochromatin regions may still be more difficult to map and sequence, even if not occurring in blacklisted regions. It may also be that regions of heterochromatin are also less accessed and less damaged, making DSBs simply less frequent. Several differences across actual DSB frequency were clear when comparing fold change across neural and cancer cells. Given the single repeat availability for MCF7 and K562, no statistical difference was tested between GSCs and commercial cancer cells, however these both demonstrated some fold change increases at 3' UTRs of >1.4. Though not as great as GBM lines, K562 and MCF7 still demonstrated a fold change greater than that of neural lines which had a maximum fold change increase of 1.294. Both GBM lines and commercial cancer cells appeared to show a greater fold change decrease in actual verses expected DSBs in intergenic sites compared to neural cells. Indeed, the absolute proportion of DSBs within intergenic sites was higher in neural cells compared to GSCs and showed a statistically significant difference between GSCs and neural cells (Figure 5.10, Table *5.4*). The reasons for the relative decrease in expected DSBs at intergenic regions in GSCs were not clear, however it may reflect the relative quiescence of neural cells compared to the more active GSCs. A study by Lawlor et al. (Lawlor et al., 2020) highlighted that oncogenic NSCs were capable of outcompeting healthy NSCs and inducing a quiescent state in healthy cells, allowing for preferential proliferation of oncogenic counterparts. There is likely a relative increase in activity and proliferation in

GSCs compared to neural cells. DSBs that are related to cellular activity such as transcription and replication would likely occur in genic sites as opposed to intergenic sites. Therefore, if GSCs had a greater number of DSBs occurring secondary to transcription or replication, the proportion of DSBs would likely be higher in genic sites than in intergenic sites. Conversely, if neural cells were in a relatively quiescent state, they may possess a lower proportion of DSBs in genic sites and so would have a higher proportion of DSBs occurring within intergenic locations instead.

As previously described, GBM lines showed an apparent increase in mean DSBs at TTS which appeared greater than that of both commercial cancer lines and neural cell lines. This was confirmed when comparing the expected versus actual fold change of DSBs at 3' UTRs in neural cells and GSCs, where neural cells demonstrated a mean fold change from expected of 1.176 compared to GSCs which demonstrated a mean fold change from expected of 2.513. This again appears to suggest that these GBM lines harbour an important difference of DSB distribution at TTS compared to neural and arguably commercial cancer cell lines. This indicates a possible GBM-specific tendency towards DSB accumulation at TTS which could be linked to upregulation of TopIIß (Kenig et al., 2016). Topoisomerase inhibitors have previously been demonstrated as showing strong efficacy *in vitro* in GBM, however clinical trials have failed to demonstrate similar results (Darling and Thomas, 2001). Whilst the topoisomerase inhibitor etoposide has been associated with some advantage in survival through meta-analysis, meaningful clinical benefits appear limited (Leonard and Wolff, 2013). This may be related to the challenge of the blood-brain barrier, where effective penetration into the brain with agents such as etoposide is limited, though results are very variable (Stewart et al., 1984). Usage of biomarkers to identify resistance or sensitivity genes related to topoisomerases have been under investigation and may assist the efforts in better personalising treatment for GBM, though clinical use remains limited currently (Cerami et al., 2012).

### 5.4.3 Summary of conclusions

- Mean DSB patterns vary across TSS in GSC lines, however all three lines demonstrate an increase in mean DSBs at TTS.

- The neural cell and non-glioma cancer lines investigated demonstrate an overall increase in mean DSBs at TSS but neural cells demonstrate significantly lower peaks at TTS compared to GSCs.

- All cell lines had a lower than predicted DSB frequency at intergenic sites, however, neural cells had a higher proportion of DSBs in intergenic sites compared to GBM and commercial cell lines.

# Chapter 6 Gene transcription, euchromatin enrichment and differential DSBs across GBM

## 6.1 Introduction

Given the preponderance of DSBs within genes and previous evidence that transcriptional activity has been associated with DSBs, DSB density was investigated in transcriptionally active genes (Aymard et al., 2014). This chapter utilised gene expression and euchromatin enrichment to investigate endogenous DSB density in GSCs and differentiated cells. The patterns of DSB density in GSCs and differentiated progeny cells have been broadly similar in the previous chapter's findings. Therefore, in addition to gene expression and chromatin accessibility, differential DSB density was investigated between GSCs and differentiated cells to identify whether there were key areas of disparity.

### 6.1.1 Transcription-related endogenous DSBs

Previous chapters indicated that genic regions have an important role in endogenous DSB location, therefore we explored the influence of transcriptional activity of genes on DSB density. Previous BLISS studies have shown that genes are important sites of interest for DSB locations (Wei et al., 2016). In neural cells, highly transcribed genes have been shown to have higher DSB break frequencies which may indicate that transcriptional activity of a gene has an important bearing on relative frequency of DSBs (Michel et al., 2022, Brambilla et al., 2020). There is also evidence supporting that transcriptionally active regions are preferentially repaired over regions in heterochromatin (Puget et al., 2019). Conversely, highly transcribed genes have also been shown to be repaired via HR rather than NHEJ, potentially resulting in a delay in repair depending on cell cycle dynamics (Aymard et al., 2017).

### 6.1.2 Chromatin profiling

It is well known that chromatin dysregulation and altered chromatin distribution in cancer can have major influences on mutation frequency and gene expression (Polak et al., 2015, Beck et al., 2012). Chromatin remodelling in GBM is a known driver of treatment resistance (Liau et al., 2017, Chen et al., 2022). Regarding the potential likelihood of DSB pattern and chromatin distribution, proximity to

nucleosomes has been linked to a reduced frequency of DSBs, specifically in the context of DSBs occurring following IR (Brambilla et al., 2020). The reason for the association of nucleosome proximity with a reduced incidence of DSBs is uncertain. It is possible that DNA in close proximity to nucleosomes are shielded from IR-associated damage. It could also be that proximity to nucleosomes promotes a faster response to DNA damage or promotes accurate repair. This raises the question of whether the pattern of chromatin condensation and chromatin remodelling acts to facilitate early DSB repair or may protect against the induction of damage, though determining the difference between these possibilities is challenging. Interestingly, chromatin remodelling has also been associated with RS resolution by promoting fork stability and lesion processing (Fournier et al., 2018). R-Loops, which have been alluded to as important players in RS and DSB induction have also been described as sites with the potential to alter chromatin structure (Xu et al., 2021a, Promonet et al., 2020, Skourti-Stathaki et al., 2014).

### 6.1.3 Differential DSB patterns across GSC and differentiated GBM lines

As previously discussed, it has been established that primary cell lines E2, G7 and R10 GSCs have demonstrated radioresistance compared to their differentiated progeny (Carruthers et al., 2018). Given the contrast in cell survival following IR, distinguishing features between GSCs and differentiated cells may give insight into important resistance mechanisms. The elevated RS levels in GSCs have been proposed as a means of priming the aberrant DDR to allow rapid repair from radiotherapy-induced DSBs (Carruthers et al., 2018). Interestingly, GSCs also demonstrate markers of persistent DSBs, even at baseline (Carruthers et al., 2018). Therefore, there may be a subgroup of endogenous DSBs present that do not risk cell death but rather prime the DDR for management of lethal DSBs. Identifying locations of prominent DSBs may allow insight into causative mechanisms with further implications for aberrant phenotypes.

### 6.1.4 Aims

- Investigate the relationship between DSB density and transcriptional activity in E2, G7 and R10 GSCs and differentiated cells.

- Identify links in DSB density across high and low euchromatin-enriched genomic regions in E2, G7 and R10 GSCs and differentiated cells.

- Compare the total number of DSBs detected using BLISS in GSCs and differentiated cells in E2, G7 and R10 cells and identify contrasts in total DSBs.

- Detect sites of contrasting DSB density in E2, G7 and R10 cells in GSCs compared to differentiated cells in 100 kbp regions and in genes

## 6.2 Materials and Methods

### 6.2.1 DSBs in actively transcribed genes

Using RNA-seq data, generated and processed in triplicate in DESeq2 by Dr Emily Clough, raw DNA counts were converted to TPM and plotted against DSB density where DSB density was adjusted gene length (DSBs/kbp) as per previous chapters. To calculate the TPM from the raw RNA-seq reads, these were divided by the length of kilobases per gene giving reads per kilobase. Reads per kilobase were then totalled per sample and divided by 1 million to give a per million scaling factor. Then reads per kilobase was divided by the per million scaling factor to give TPM. This was calculated using the following function in RStudio:

**Equation 1: transcripts per million**

*tpm3 <- function(counts,len) { x <- counts/len return(t(t(x)*1e6/colSums(x)))}.*

RNA-seq data was originally generated using hg19 and therefore gene names were used for linking RNA-seq gene expression and BLISS gene DSB density. Genes that were not identified in RNA-seq data were excluded. As with previous analysis, genes that had no DSBs were excluded. Similarly, genes with no expression from RNA-seq data were also excluded from analysis. Again, genes

with a length of <200 bp were excluded in order to minimise skew from overrepresentation of short genes with 1 or 2 DSBs (8171 genes).

WGS data became available in the latter part of this project and so the effect of gene copy number on DSB density was reviewed retrospectively in the context of gene expression after the majority of analysis had been performed and data processed. Having investigated CNV retrospectively, it gave helpful context to data already analysed, however DSB data was not re-normalised to CNV going forward. For analyses specifically considering copy number, CNVs within the GSC lines E2, G7 and R10, were determined from WGS data generated by Novogene™. Novogene™ files identifying CNV locations were used in conjunction with BLISS data to generate DSB density within CNV and normal copy number base pair start and end sites. Files provided CNV start and end sites across chromosomes and additionally provided genes within these regions with an associated copy number. All CNVs were marked by whole number gain or losses (i.e. CN=1/CN=5) rather than partial gain or loss. Therefore, for each gene, the number of DSBs within that gene was adjusted for both gene length and gene copy number.

### 6.2.2 DSBs in highly euchromatin enriched sites

For investigation of chromatin structure, ATAC-seq libraries were prepared by Karen Strathdee in matched GSC and differentiated cell lines in E2, G7 and R10. Experiments were performed in triplicate. ATAC-seq datasets were processed using Nextflow nf-core ATAC-seq pipeline and consensus peaks were generated from experimental repeats. The consensus peak datasets were used for analysis of euchromatin enrichment. For this, peaks from consensus .broadPeak files were ordered by consensus signal value indicating high to low peak enrichment. The top 1000 peaks and bottom 1000 peaks from each dataset were extracted and the DSB density was corrected for peak length (DSB/kbp). As with genes, peaks of less than 200 bp in length were excluded from analysis to minimise skew. DSB density per peak was compared across the 1000 most enriched peaks and 1000 least enriched peaks to compare regions of high euchromatin enrichment with regions of relatively low euchromatin enrichment.

### 6.2.3 Differential DSB patterns in GSCs and differentiated cells

The total number of DSB reads per experiment was calculated for GSCs and differentiated cells. The BLISS protocol does not allow for the calculation of the absolute number of DSBs to be measured due to the dependence on PCR amplification for library preparation. However, this is internally accounted for by UMIs allowing for the removal of PCR duplicates during the preprocessing of files, to allow comparison between matched groups which in this case were GSCs and differentiated cells.

The differential DSB density between GSCs and differentiated cells was investigated in 100 kbp regions and then separately in genes using the RStudio package “DESeq2” (Love et al., 2014). Significant differential results were determined by any regions with an adjusted p value of <0.05 and a log2 fold change of >1/<-1. DESeq2 FDR correction was performed using the BHP as part of the DESeq2 package analysis. The “Gene Ontology” RStudio package (Ashburner et al., 2000, Thomas et al., 2022) was used for over-representation analysis (ORA) to identify any sets of genes represented in the differentially broken genes.

## 6.3 Results

### 6.3.1 DSB density is increased in actively transcribed genes

As previously described in earlier chapters, data from all three cell lines indicated that DSB density was increased in genes relative to intergenic regions. Replication-transcription collisions have been highlighted as at risk of replication fork collapse and subsequent DSBs (Nickoloff et al., 2021). Transcription has also been cited as requiring physiological DSB induction from topoisomerases in elongation of transcripts (Schoeffler and Berger, 2008). Highly transcribed genes have been shown to be preferentially repaired through HR which may also influence DSB frequency with regards to DNA repair timing (Yasuhara et al., 2018). Therefore, transcriptional activity and DSB density was investigated in genes in E2, G7 and R10 GSCs and differentiated cells.

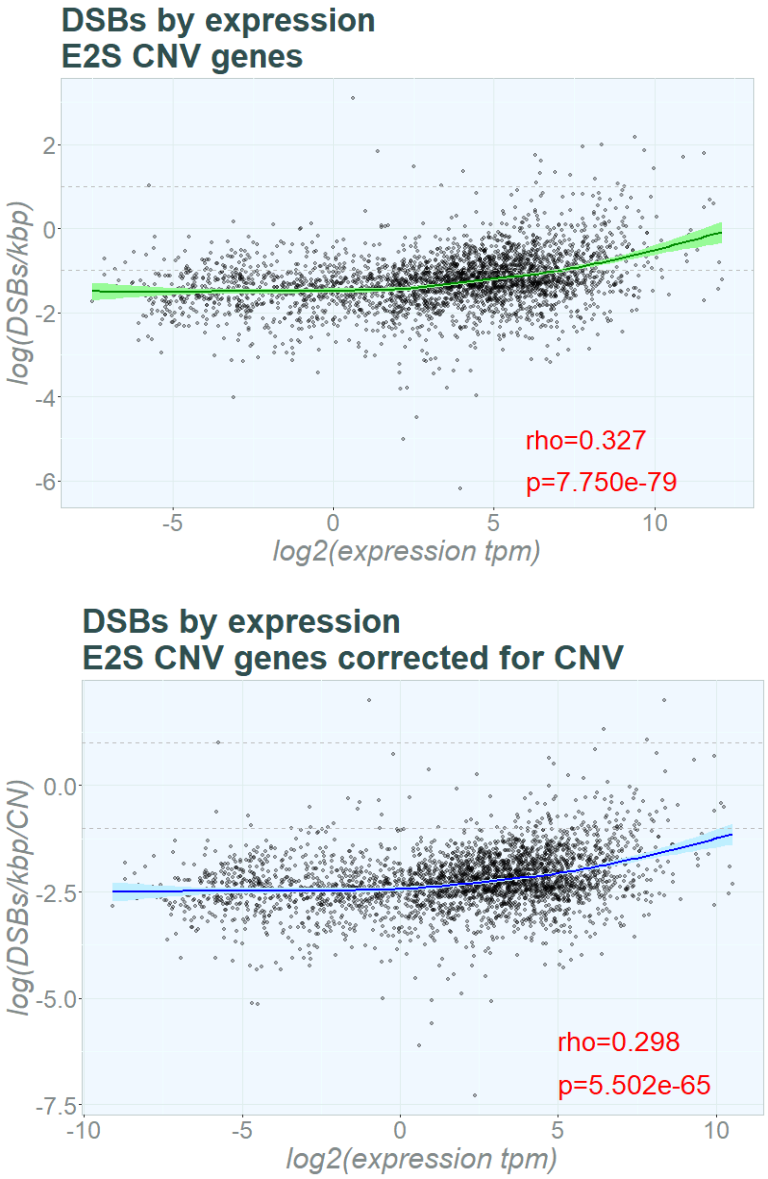
#### 6.3.1.1 Actively transcribed genes in copy number-altered regions show similar distribution in DSB density after copy number adjustment

WGS data became available for GSCs in the latter stages of this project. DSB density was therefore investigated in CNV-affected genes which were selected and adjusted for by copy number. For this, gene DSB density with and without adjustment for copy number was plotted against gene expression counts from RNA-seq analysis (TPM). Given that gene copy number could also result in higher or lower gene expression levels due to gene abundance, expression data was also adjusted for copy number (Figure 6.1).

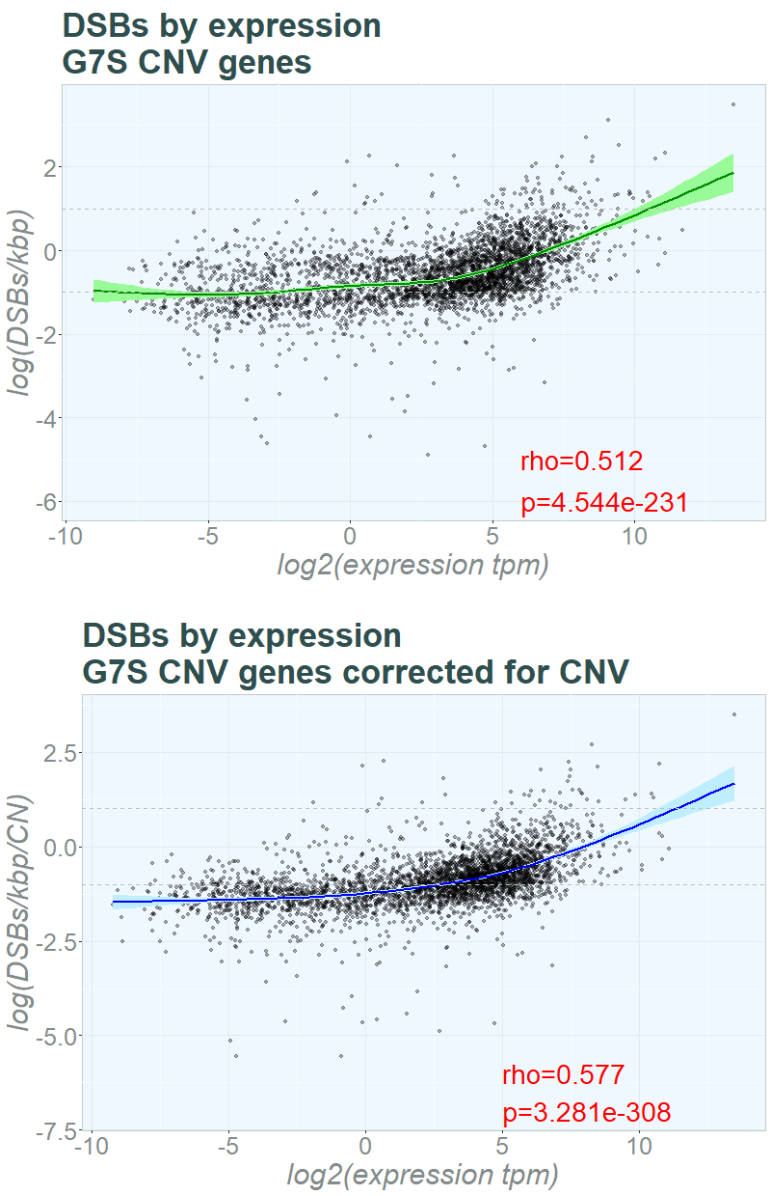
Figure 6.1 shows CNV-affected genes plotted in E2, G7 and R10 GSCs. Both the data normalised for copy number-adjusted (blue) and the uncorrected data (green) demonstrate a conserved positive correlation of DSB density relative to transcript abundance. E2 GSCs demonstrated a rho value of 0.327 in uncorrected CNV genes and a rho of 0.298 in CNV genes adjusted for copy number (Spearman rank correlation). G7 GSCs demonstrated rho values of 0.512 and 0.577 uncorrected and copy number adjusted data respectively, both demonstrating a consistent positive correlation of similar magnitude. Finally, R10 GSCs uncorrected genes had a rho value from 0.374 and after copy number correction a rho value of 0.404. The DSB density for genes unaffected by copy number was also plotted (available in supplementary figures) and showed rho values across GSCs that were consistent with the results of the CNV-affected genes (E2: rho = 0.349, G7: rho = 0.568, R10: rho = 0.392).

Overall, the pattern of DSB density across CNV-corrected genes in relation to gene expression was similar to that of CNV-uncorrected genes, indicating that there was no overall biological significance in terms of the impact of CNV on DSB density in transcriptional expression. Here, the positive correlation showed a late uptick in DSBs in the genes with highest TPM compared to the least transcriptionally active genes.

(a)


DSBs by expression
E2S CNV genes
log(DSBs/kbp)
2
0
-2
-4
-6
-5
0
5
10
log2(expression tpm)
rho=0.327
p=7.750e-79
DSBs by expression
E2S CNV genes corrected for CNV
log(DSBs/kbp/CN)
0.0
-2.5
-5.0
-7.5
-10
-5
0
5
10
log2(expression tpm)
rho=0.298
p=5.502e-65

(b)


DSBs by expression
G7S CNV genes
log(DSBs/kbp)
2
0
-2
-4
-6
-10
-5
0
5
10
log2(expression tpm)
rho=0.512
p=4.544e-231
DSBs by expression
G7S CNV genes corrected for CNV
log(DSBs/kbp/CN)
2.5
0.0
-2.5
-5.0
-7.5
-10
-5
0
5
10
log2(expression tpm)
rho=0.577
p=3.281e-308

(c)

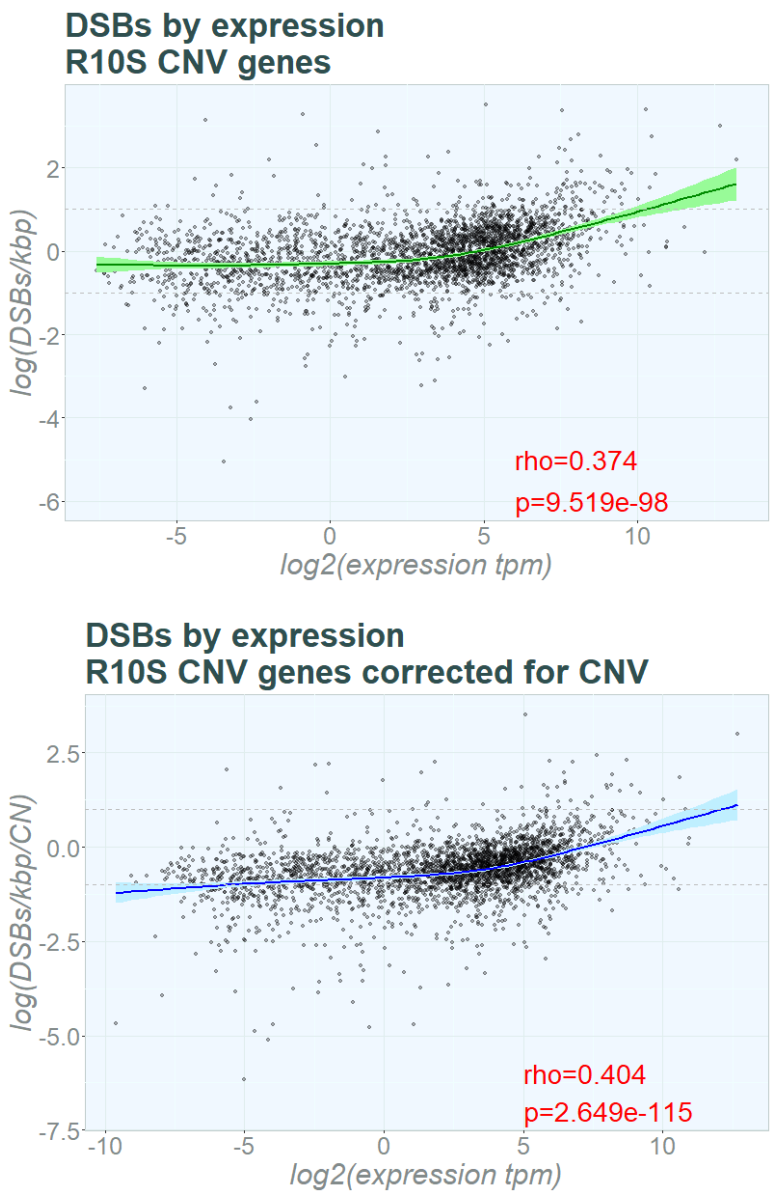


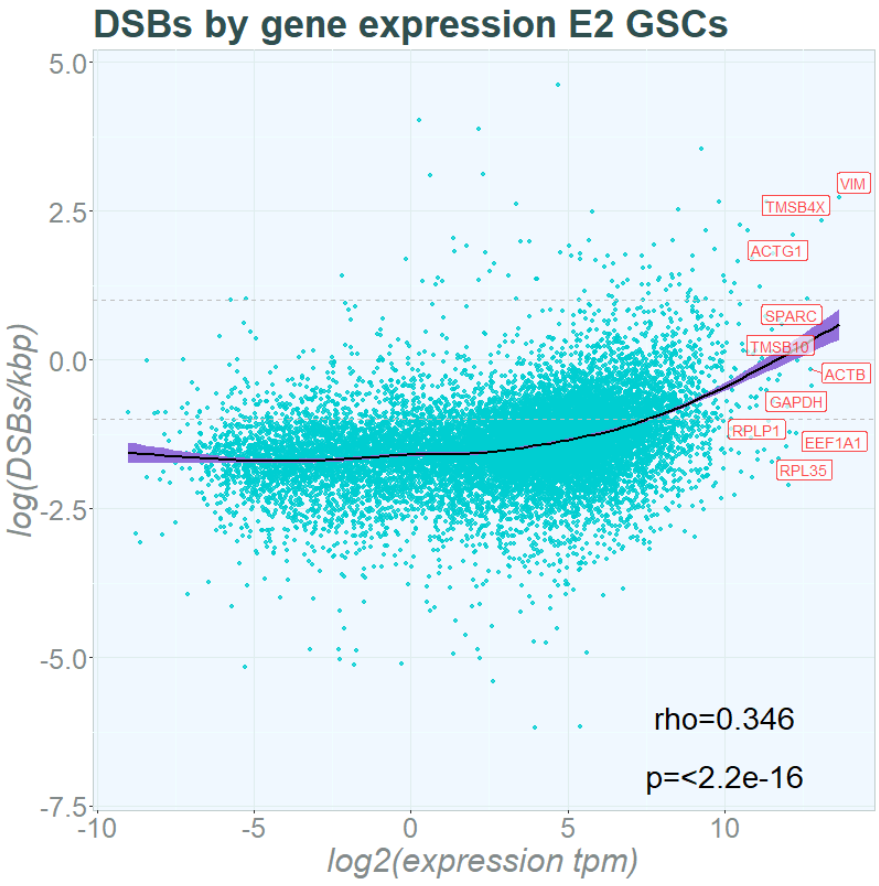

**Figure 6.1. Gene expression and DSB density in genes adjusted for CNV**

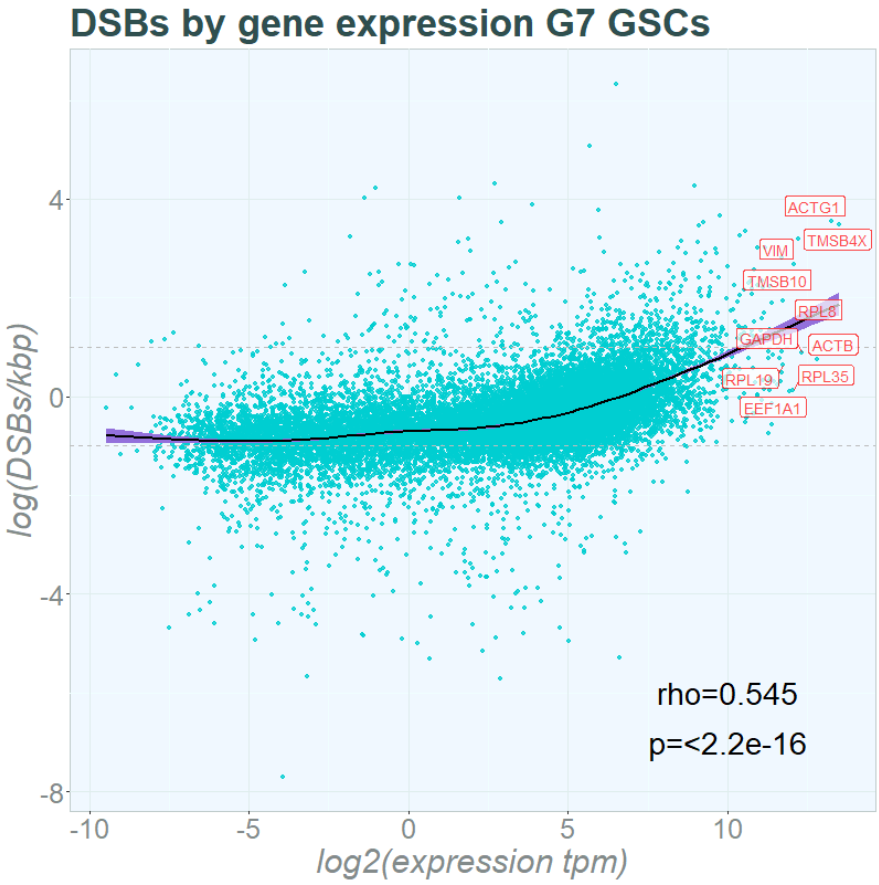
Genes affected by CNV were adjusted for copy number. Mean DSB density from BLISS data and mean gene expression (TPM) from RNA-seq data are displayed before and after copy number adjustment. CNV adjusted data ("CNV normalised") is displayed as the top panel with the blue line and uncorrected data is displayed as the bottom panel with the green line. Green and blue lines represent locally estimated scatterplot smoothing (LOESS) using non-parametric local regression to estimate the shape of the data. Spearman rank correlation estimates with rho and p values displayed in the text. Individual dots mark CNV-affected genes. Data was log transformed for visualisation. (a) E2 GSC line. (b) G7 GSC line. (c) R10 GSC line.

### 6.3.1.2 DSB density is highest in the most highly transcribed genes

Figure 6.2 shows gene expression (TPM) against DSB density (DSB/kbp) for genes containing DSBs and TPM values >0 for E2, G7 and R10 GSCs. In E2, G7 and R10

cell lines DSB density trended towards being higher in the genes with the highest expression (GSCs shown only, differentiated cell plots available in supplementary figures). DSB density in genes was positively correlated with gene expression in all cell lines. E2 lines had the lowest rho values at 0.346 in GSCs and 0.302 in differentiated cells, followed by R10 cells which had a rho value of 0.346 in R10 GSCs and 0.400 in differentiated cells. G7 GSCs demonstrated the strongest positive correlation with rho values of 0.545 and 0.519 for GSCs and differentiated cells respectively. GSCs and differentiated cells demonstrated similar rho values across E2, G7 and R10 cell lines indicating highly similar relationships between DSB density and transcription in both GSCs and differentiated cell populations. The top 10 most highly expressed genes were annotated for reference in red. GSCs and differentiated cells demonstrated overlap in the genes with highest expression with *TSMB4X*, *VIM*, *TSMB10*, *GAPDH* and *ACTB* all identified as highly expressed annotated genes in all lines for GSCs and differentiated cells. Of these, *VIM, GAPDH* and *ACTB* were identified as housekeeping genes (Wang et al., 2023, Cuvertino et al., 2017, Panina et al., 2020). *TSMB10* (Thymosin beta 10) and *TSMB4X* (Thymosin beta X linked) expression have been associated with poorer prognosis in GBM and both appear more widely associated with tumourigenesis across multiple cancer types (Xiong et al., 2022). There was lower DSB density at the least transcriptionally active genes across E2, G7 and R10 lines and minimal gradient change of LOESS curve in the low transcription genes. However, at the most transcriptionally active genes there was an uptick in DSB density indicated by the LOESS curve. This positive trend indicated that the positive rho result may be driven by genes with the highest transcriptional expression. Therefore, gene DSB density was also investigated by genes grouped into expression quintiles.


DSBs by gene expression E2 GSCs
log(DSBs/kbp)
log2(expression tpm)
rho=0.346
p=<2.2e-16
VIM
TMSB4X
ACTG1
SPARC
TMSB10
ACTB
GAPDH
RPLP1
EEF1A1
RPL35


DSBs by gene expression G7 GSCs
log(DSBs/kbp)
log2(expression tpm)
rho=0.545
p=<2.2e-16
ACTG1
TMSB4X
VIM
TMSB10
RPL6
GAPDH
ACTB
RPL19
RPL35
EEF1A1

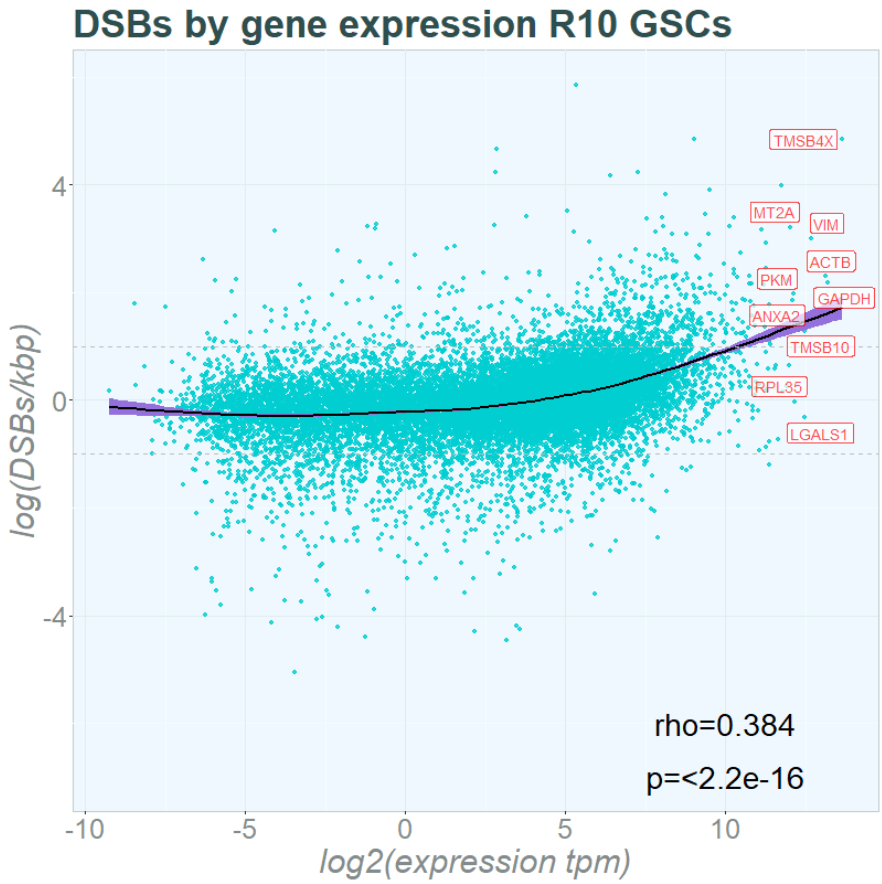



**Figure 6.2. Gene expression and DSB density in E2, G7 and R10 GSC lines.**

Mean gene expression (TPM) plotted against logged mean DSB density (DSB/kbp) per gene. Expression means calculated across RNA-seq repeats, DSB means calculated across BLISS repeats. DSB/kbp and TPM expression are log-transformed for visualisation and represented as a scatterplot. The top ten genes with the highest expression are represented in red text. Horizontal dotted lines represent -1 and +1 log(DSB/kbp). The purple line represents locally estimated scatterplot smoothing (LOESS) using non-parametric local regression to estimate the shape of the data. Statistical analysis was performed by non-parametric Spearman Rank correlation with rho values and p values displayed on the graphs (a) E2 GSC line. (b) G7 GSC line. (c) R10 GSC line.

Gene DSBs/kbp is displayed per percentage quintile in Figure 6.3 from lowest to highest gene expression for E2, G7 and R10 GSCs (differentiated cells data available in supplementary figures). The p-values for comparative significance across quintiles are displayed below each violin plot (ANOVA and post-hoc Tukey test). Genes with the highest expression (80-100%) had a significant fold change in median DSB density from the lowest expression quintiles (0-20%) across all three cell lines.

In E2 GSCs, there was a significant fold increase in DSB density of genes in the top 80-100% of transcriptionally active genes compared to genes in the bottom 60%. This was also the case in E2 differentiated cells (see supplementary

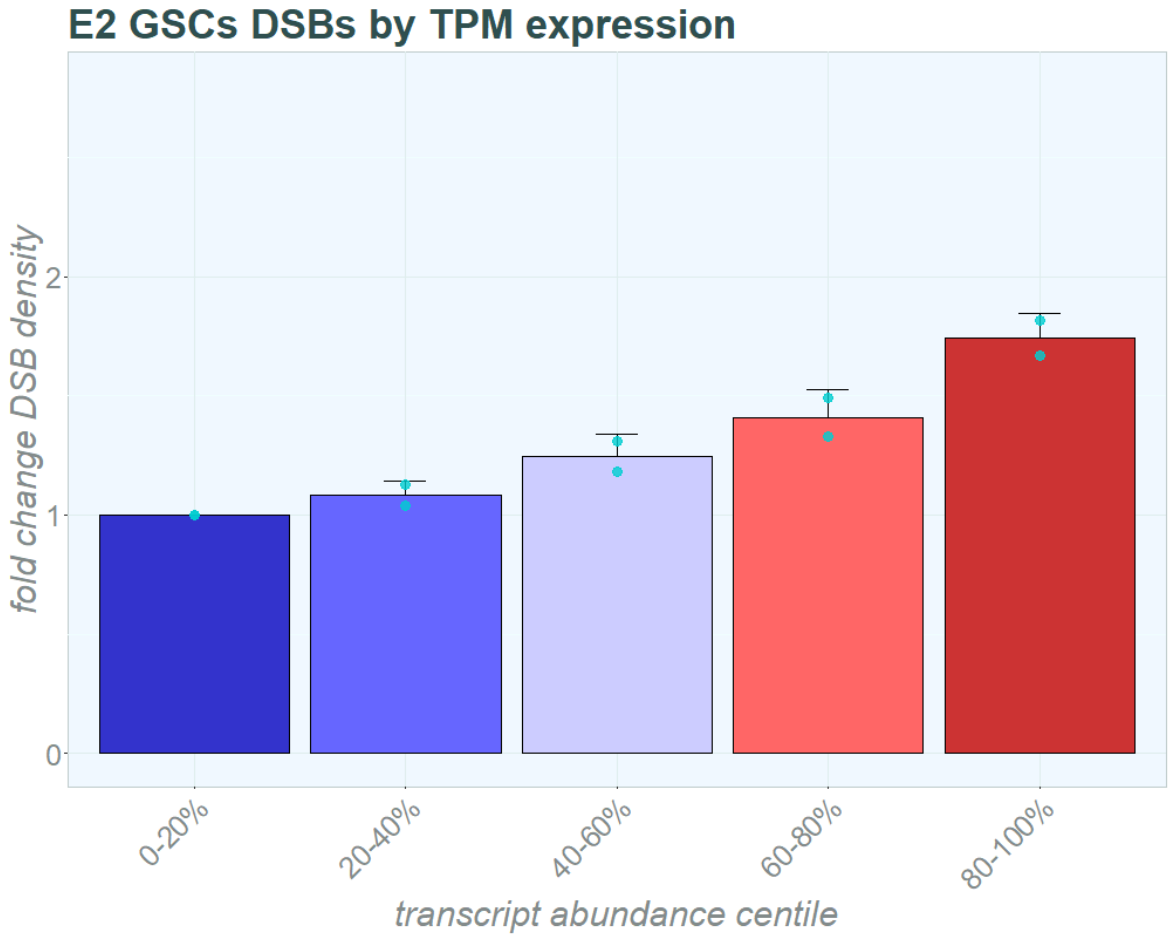
figures). There was no significantly detectable difference between the bottom three quintiles: 0-20%, 20-40% and 40-60%.

G7 GSCs demonstrated the most significant results in differences across the transcriptional activity quintiles. The top quintile (80-100%) a demonstrated highly significant fold increase in DSB density compared to the bottom three quintiles with an adjusted p-value of <0.00001. G7 differentiated cells also demonstrated a significant increase in DSB density in the top quintile (80-100%) compared to all lower quintiles (0-80%) (see supplementary figures). Only fold change between the lowest two quintiles 0-20% and 20-40% were non-significant.

Finally, the R10 GSC top two quintiles demonstrated a significant fold increase in DSB density from genes that were in the bottom 40% of transcriptionally expressed genes. This was also the case in R10 differentiated cells (see supplementary figures). Again, there was no significant fold difference between the bottom two quintiles with regards to DSB density.

Across all three lines, there were no significant differences between the bottom 0-20% transcriptionally active genes and the 20-40% transcriptionally active genes, however genes which were the most transcriptionally active (80-100%) demonstrated a consistently significant increase from the least transcriptionally active (0-20%). Overall, these findings indicated that highly transcriptionally active genes increased in DSB density but genes with the lowest transcriptional activity quintiles did not demonstrate significant differences in DSB density.

**(a)**

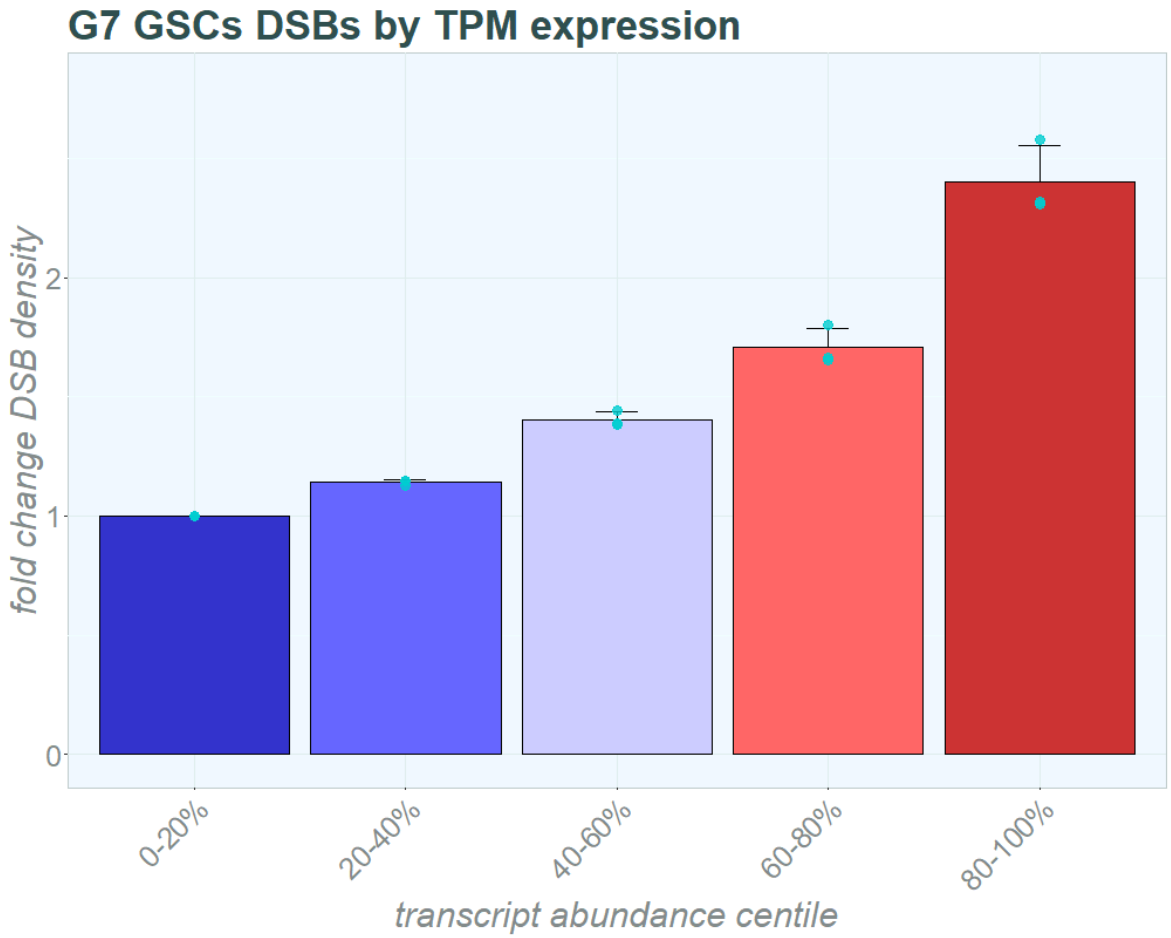



**(b)**

E2 GSC exp. By 5ths p-values

| E2 GSC exp. by 5ths p-values | 0-20% | 20-40% | 40-60% | 60-80% | 80-100% |
|---|---|---|---|---|---|
| 0-20% | 1 | | | | |
| 20-40% | 0.854537 | 1 | | | |
| 40-60% | 0.146192 | 0.40048 | 1 | | |
| 60-80% | 0.023791 | 0.057112 | 0.409785 | 1 | |
| 80-100% | 0.001693 | 0.002927 | 0.01054 | 0.052478 | 1 |

**(c)**

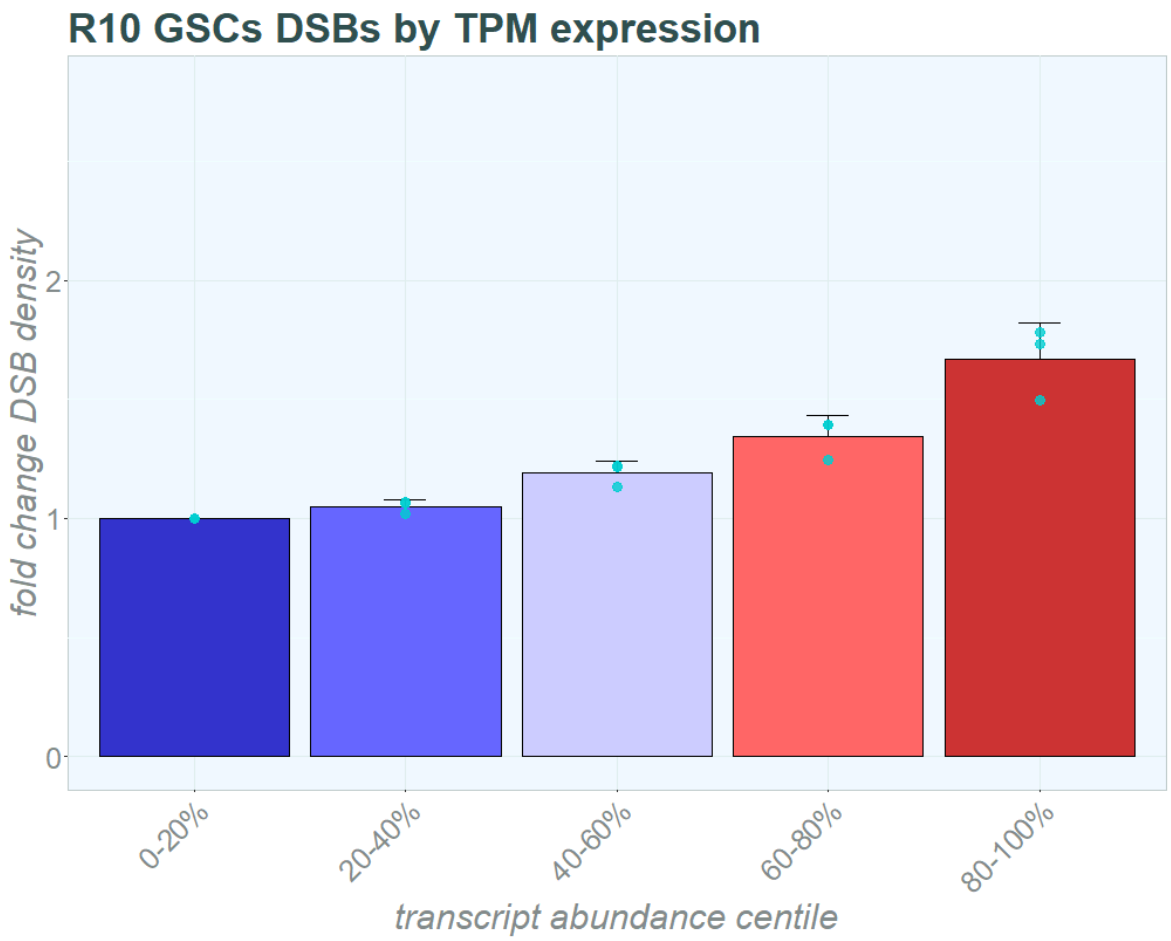



**(d)**

G7 GSC exp. By 5ths p-values

| G7 GSC exp. By 5ths p-values | 0-20% | 20-40% | 40-60% | 60-80% | 80-100% |
|---|---|---|---|---|---|
| 0-20% | 1 | | | | |
| 20-40% | 0.25977 | 1 | | | |
| 40-60% | 0.00070 | 0.01511 | 1 | | |
| 60-80% | 0.00001 | 0.00004 | 0.00622 | 1 | |
| 80-100% | <0.00001 | <0.00001 | <0.00001 | 0.00001 | 1 |

**(e)**




**(f)**

R10 GSC exp. By 5ths p-values

| R10 GSC exp. By 5ths p-values | 0-20% | 20-40% | 40-60% | 60-80% | 80-100% |
|---|---|---|---|---|---|
| 0-20% | 1 | | | | |
| 20-40% | 0.94437 | 1 | | | |
| 40-60% | 0.10336 | 0.29540 | 1 | | |
| 60-80% | 0.00316 | 0.00916 | 0.21892 | 1 | |
| 80-100% | 0.00001 | 0.00003 | 0.00024 | 0.00482 | 1 |

**Figure 6.3 Quintile gene expression and DSB density in GSC lines E2, G7 and R10.**

Mean DSB density per gene by expression quintiles from lowest (0-20% in blue) to highest expression (80-100% in red) displayed as a bar chart. Bar chart displays mean of medians of the fold-change in DSBs/kbp from the lowest expression quintile. Whiskers represent standard deviation. Replicate medians displayed in turquoise. Significance testing performed using ANOVA and post-hoc analysis by Tukey test. Corresponding p-values are displayed in tables below between quintiles. Red highlighted results demonstrate significant results of an adjusted p-value <0.05. (a) Mean of duplicate repeats for E2 GSC lines using RNA-seq and BLISS data. DSBs by transcription quintiles from least to most expressed genes. (b) E2 corresponding p-values generated in post-hoc analysis. (c) Mean of triplicate repeats for G7 GSC

lines using RNA-seq and BLISS data. DSBs by transcription quintiles from least to most expressed genes. (d) G7 corresponding p-values generated in post-hoc analysis. (e) Mean of triplicate repeats for R10 GSC lines using RNA-seq and BLISS data. DSBs by transcription quintiles from least to most expressed genes. (f) R10 corresponding p-values generated in post-hoc analysis.

### 6.3.2 DSB density in euchromatin-enriched regions is variable across GBM cells

Recognising that chromatin modulation holds important roles in transcriptional expression in cancer and has been linked to reduced DNA damage following IR, DSB density was investigated in areas of euchromatin enrichment (Mack et al., 2019, Chen et al., 2022, Brambilla et al., 2020). Actively transcribed genes must be accessible to transcriptional machinery and therefore are intrinsically linked to chromatin modulation, hence the dual approach of assessing DSB density in transcriptionally active genes and areas of accessible euchromatin.

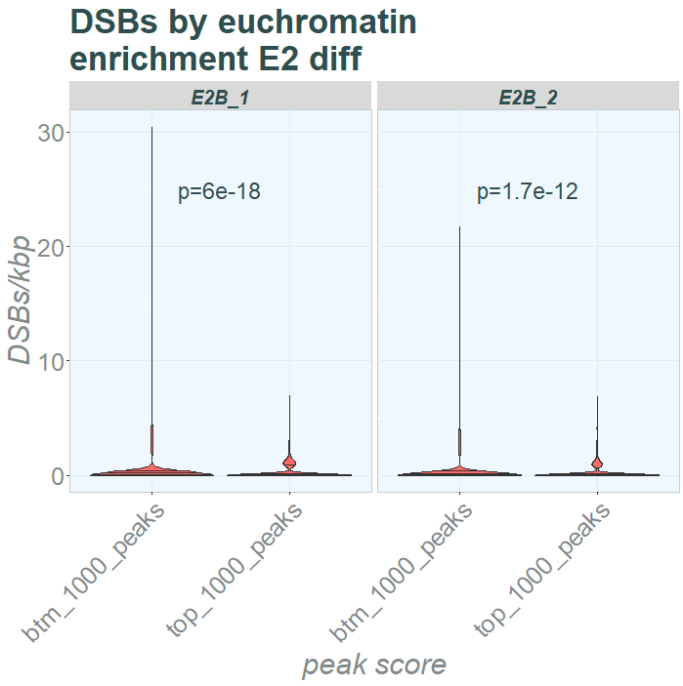
The density of DSBs (DSB/kbp) was calculated at consensus euchromatin peaks in E2, G7 and R10 GSC and differentiated cells. To determine whether highly enriched peaks had contrasting DSB density to low enrichment peaks, the DSB density was compared in the 1,000 peaks with the highest enrichment to the 1,000 peaks with the lowest enrichment (Figure 6.4). Figure 6.4 displays DSB density across GSC and differentiated lines in the top and bottom euchromatin enrichment peaks. Notably, there was a non-Gaussian distribution of DSB density across peaks with a right skew, indicating that there were many peaks with a very low DSB density. Across all lines, there were peaks in both the lowest and highest enrichment groups that had few to no DSBs within the consensus peak sets.

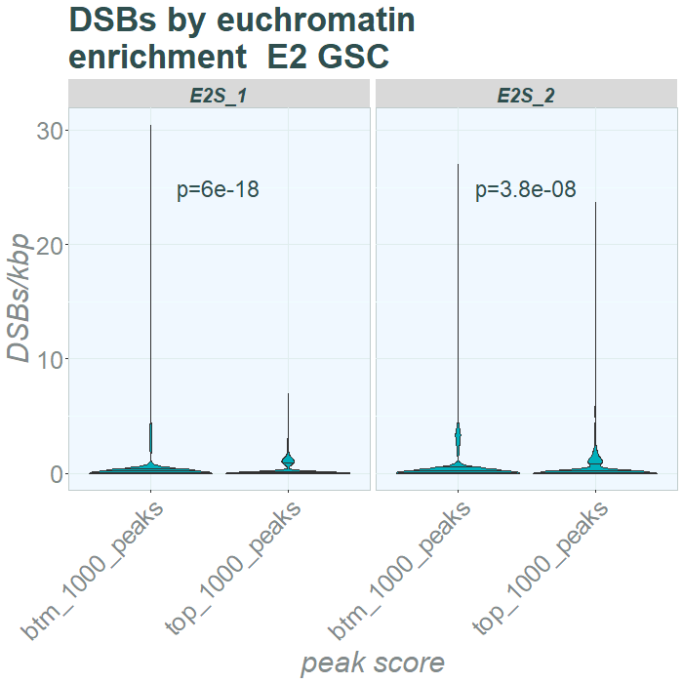
The repeats of E2 differentiated cells and GSCs both had median DSB values of close or equal to 0 DSB/kbp in both the highest and lowest enriched peaks. There was a small subset of regions with a DSB density of above 0 DSB/kbp identifiable as noted by the shape of the violin plot. In E2 differentiated cells, both samples showed an increase in DSB density in the top enriched peaks compared to the least enriched peaks. There was a greater subset of top peaks with >0 DSB/kbp DSB density compared to the bottom peaks in differentiated cells. However, there were several peaks in the lowest enrichment group that had many more DSBs/kbp than any peak in the highest enriched peaks.

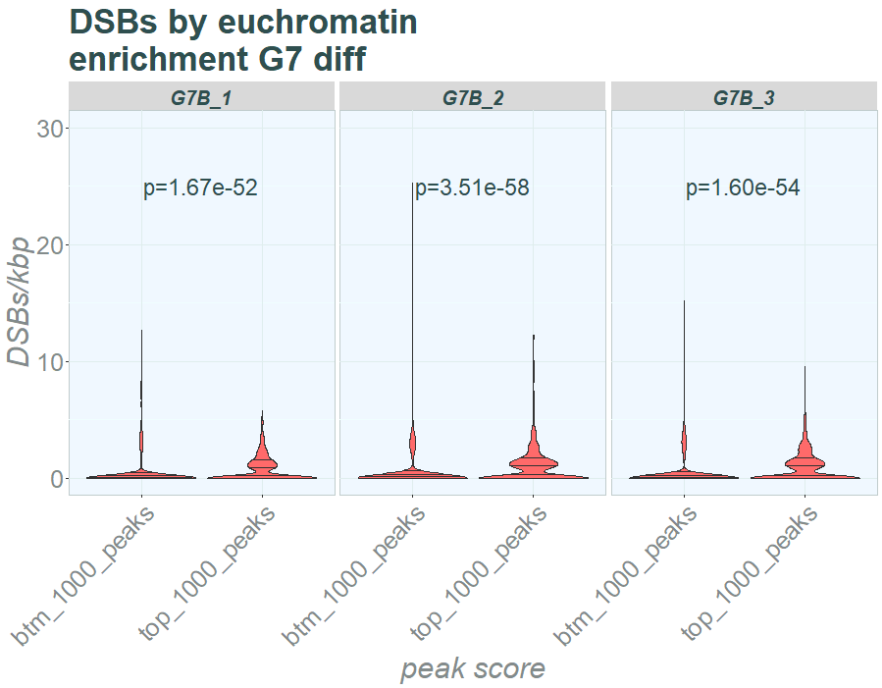
Regarding E2 GSCs, the top enriched peaks had higher DSBs in both samples, though again there was a subset of peaks in the lowest enrichment group which had a DSB density higher than any of the top enrichment group.

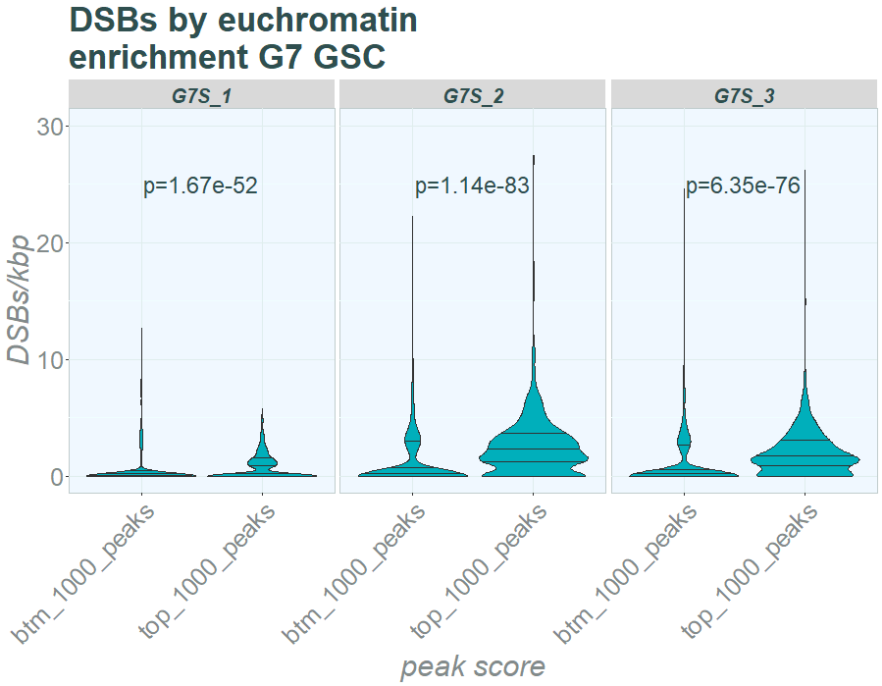
Across G7 GSCs and differentiated cells, there was a significant increase in DSB density in the euchromatin peaks with the highest enrichment compared to the lowest enrichment. Despite this, there were still many peaks that had few to no DSBs, resulting in a clear right skew of the data. G7 GSCs demonstrated the greatest difference between highest and lowest euchromatin peaks, with samples 2 and 3 indicating a median of greater than 1 DSBs/kbp. These two samples were the only two across all cell lines with a median of >1 DSBs/kbp.

The R10 GSCs and differentiated cells also had a higher DSB density in the highest euchromatin enrichment sites, though again there were some peaks in the lowest enrichment group that had a higher DSB density than peaks in the top enrichment group. In R10 differentiated samples, there was a higher DSB density in the top enrichment group compared to the bottom enrichment groups, though differences were small between groups. Similarly, R10 GSCs continued to show a higher DSB density in the peaks with the highest enrichment though the majority of both high and low enrichment sites had 0 DSBs/kbp. Taken together, whilst higher euchromatin enrichment was associated with higher DSB density, these results indicated that this relationship was variable across euchromatin peaks and less predictive than transcriptional activity.


DSBs by euchromatin enrichment E2 diff
E2B_1
E2B_2
p=6e-18
p=1.7e-12
DSBs/kbp
30
20
10
0
btm_1000_peaks
top_1000_peaks
btm_1000_peaks
top_1000_peaks
peak score


DSBs by euchromatin enrichment E2 GSC
E2S_1
E2S_2
p=6e-18
p=3.8e-08
DSBs/kbp
30
20
10
0
btm_1000_peaks
top_1000_peaks
btm_1000_peaks
top_1000_peaks
peak score


DSBs by euchromatin enrichment G7 diff
G7B_1
G7B_2
G7B_3
p=1.67e-52
p=3.51e-58
p=1.60e-54
DSBs/kbp
30
20
10
0
btm_1000_peaks
top_1000_peaks
peak score


DSBs by euchromatin enrichment G7 GSC
G7S_1
G7S_2
G7S_3
p=1.67e-52
p=1.14e-83
p=6.35e-76
DSBs/kbp
30
20
10
0
btm_1000_peaks
top_1000_peaks
peak score

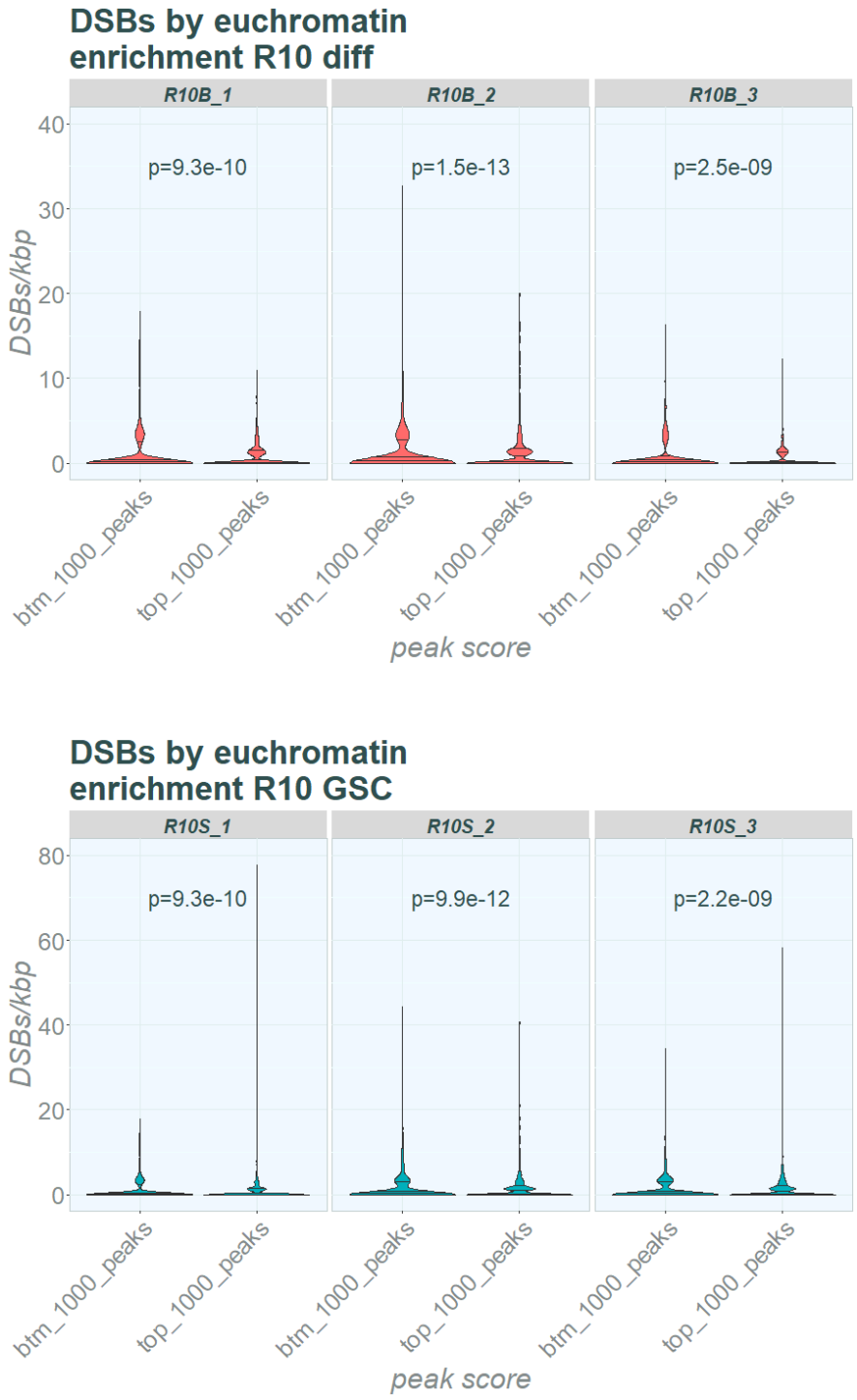



**Figure 6.4. DSB density across greatest and least enriched ATAC-seq peaks**

DSB density across ATAC-seq peaks normalised for peak length. Violin plot of 1000 least ATAC-seq enriched peaks ("btm_1000_peaks") and the 1000 peaks with the greatest ATAC-seq enrichment ("top_1000_peaks"). Significance testing by Wilcoxon signed rank test with BHP correction. Differentiated cell lines in red, GSC lines in turquoise. (a) E2 differentiated repeats 1-2. (E2B_1, E2B_2) (b) E2 GSCs repeats 1-2. (E2S_1, E2S_2). (c) G7 differentiated repeats 1-3. (G7B_1, G7B _2, G7B_3). (d) G7 GSCs repeats 1-3. (G7S_1, G7S _2, G7S_3). (e) R10 differentiated repeats 1-3. (R10B_1, R10B _2, R10B_3). (f) R10 GSCs repeats 1-3. (R10S_1, R10S _2, R10S_3).

### 6.3.3 Comparative analysis of DSB patterns across GSCs and differentiated cells

GSC lines E2, G7 and R10 have been established as radioresistant compared to their differentiated progeny (Carruthers et al., 2018). Comparative analysis of DSB distribution was performed between GSC and differentiated cell populations. Accordingly, DSB frequency across 100 kbp genomic regions and within genes was investigated to determine if there were identifiable distinctions in GSC DSB distribution compared to differentiated cells.

#### 6.3.3.1 DSB frequency is higher across GSCs compared to differentiated cells

To investigate differences in overall DSB frequency in GSCs and differentiated cells, the total DSBs detected using BLISS was aggregated from the three GBM lines. Given the differences in yield across the repeats and cell lines, the DSB yield from GSCs was also expressed as a fold change from differentiated cells. Table *6.1* shows the total number of DSBs detected per sample and contributing frequency of differentiated and GSC DSBs per biological repeat.

The total number of detectable DSB reads was variable between samples. R10 lines had the greatest abundance of DSB reads with >4 million total DSBs reads detected per pair of differentiated-GSC samples. E2 had the lowest detectable DSB reads with repeats 1 and 2 yielding a total number of DSBs of <1.5 million total reads. As described, the yield of detectable DSBs was higher in GSCs than in differentiated cells. GSCs demonstrated a fold increase in DSBs of between 1.069 to 3.352 (R10 repeat 2 and G7 repeat 1 respectively). The fold increase in detectable DSB frequency varied across repeats within cell lines, despite being an overall increase from differentiated cells to GSCs. Altogether, there was a significant fold increase in DSBs in GSCs cells compared to differentiated cells, indicating a higher DSB frequency in GSCs compared to differentiated progeny.

**Table 6.1 DSBs detected by BLISS**

| | DSBs per replicate | Diff cell DSBs | GSC DSBs | DSB fold change diff to GSC |
|---|---|---|---|---|
| ***E2 r1*** | 1,365,713 | 437,834 | 927,879 | 2.119 |
| ***E2 r2*** | 1,123,180 | 353,245 | 769,935 | 2.180 |
| ***G7 r1*** | 3,211,212 | 737,786 | 2,473,426 | 3.352 |
| ***G7 r2*** | 3,069,217 | 1,051,840 | 2,017,377 | 1.918 |
| ***G7 r3*** | 2,620,154 | 909,804 | 1,710,350 | 1.880 |
| ***R10 r1*** | 7,936,496 | 1,977,068 | 5,959,428 | 3.014 |
| ***R10 r2*** | 5,887,246 | 2,845,566 | 3,041,680 | 1.069 |
| ***R10 r3*** | 4,744,397 | 1,672,939 | 3,071,458 | 1.836 |

| | | *Fold change diff vs GSC* |
|---|---|---|
| | ***T-test:*** | p=0.012 |

Table displays the total DSB reads detected per sample for cell lines E2, G7 and R10. Repeats are denoted as r1: repeat 1, r2: repeat 2, r3: repeat 3. Absolute number of detected DSB reads for differentiated cells ("Diff cell DSBs") and GSCs (GSC DSBs) displayed. The difference in DSB frequency from differentiated cells compared to GSCs is represented as a fold change from the total number of differentiated DSBs detected. Statistical testing between differentiated and GSC fold change was performed using a t-test.

#### 6.3.3.2 GSCs do not demonstrate differential DSB locations that account for the biological changes compared to differentiated cells

To determine differential DSB density between GSCs and differentiated cells, 100 kbp regions and gene locations were investigated using "DESeq2" (Love et al., 2014). Figure 6.5 demonstrated a principal component analysis (PCA) of GSC and differentiated cells. Samples clustered by cell line rather than identity as GSCs or differentiated populations. GSCs remained closer to differentiated cells,

rather than GSCs clustering more closely with other GSCs. R10 GSCs and differentiated cells clustered most closely together than other cell lines.

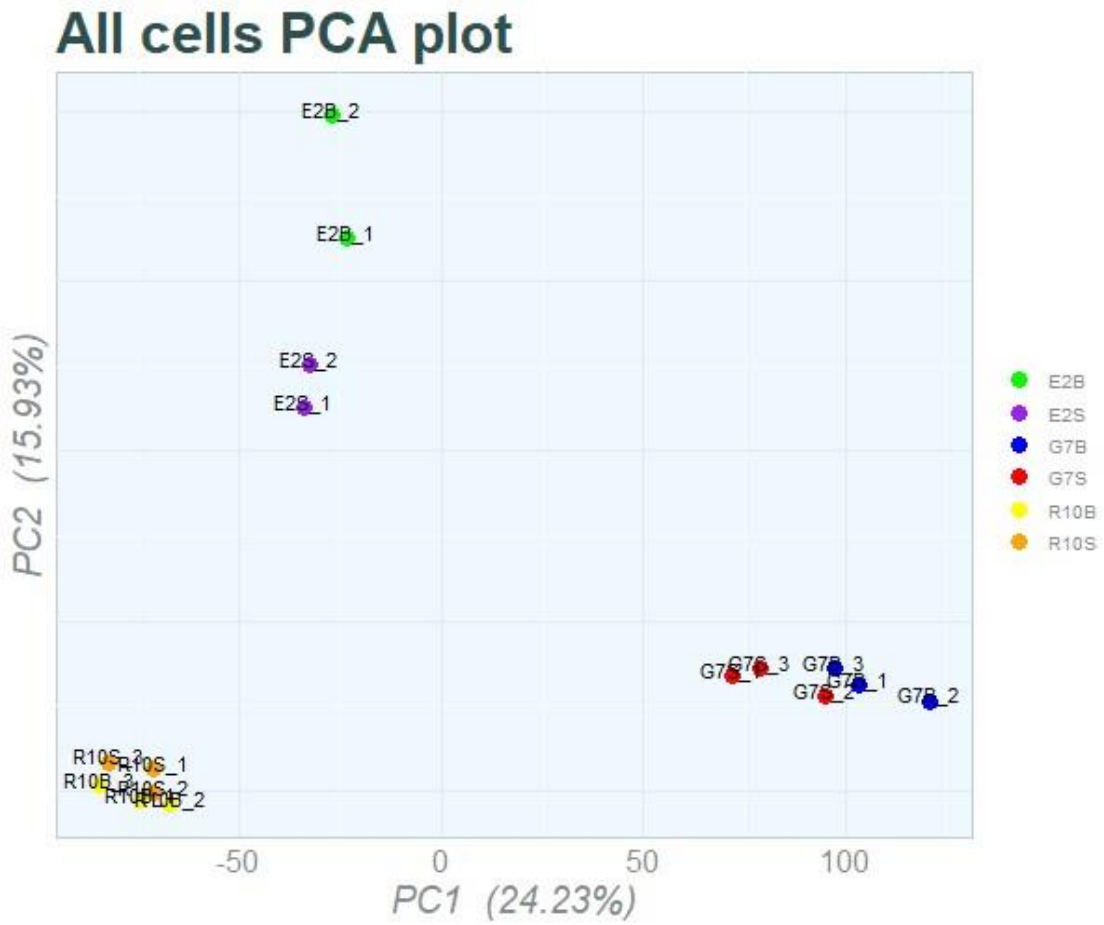



**Figure 6.5**: **PCAs of DSB density per 100 kbp regions in GSCs and differentiated cells**

Principle component analysis (PCA) of all cell lines for 100 kbp regions and individual PCA of E2 cell line. Repeats denoted by number 1-3. (a) PCA plot of all cell lines. GSC lines: E2S (green), G7S (red), R10S (orange). Differentiated lines: E2B (purple), G7B (blue), R10B (yellow). Principle component 1 and 2 shown only.

Figure 6.6 displays the 100 kbp regions and genes with significant differential DSB distribution between E2 GSCs and differentiated cells. There were three 100 kbp regions that contained significantly differential DSB distribution between GSC and differentiated sites, two of which had significantly higher relative DSB distribution in GSCs than differentiated and one with significantly higher relative DSB distribution in differentiated compared to GSCs. Regarding genes, there were four genes with significant contrasting relative DSB distribution between GSCs and differentiated cells: *TNPO1, RPL14, EIF3A* and *ENPP3*. Given the small number of gene results, Gene Ontology over-representation analysis (ORA) could

not be performed in these genes to determine any unifying themes (Ashburner et al., 2000, Thomas et al., 2022). These genes were all identified as protein coding. The gene *EIF3A* was present within the chromosome 10 region that was also identified as having a significantly higher relative DSB distribution in GSCs. Similarly, *ENPP3* was also present within chromosome 6 region 131,600,000-131,700,000 where relative DSB distribution was significantly lower in GSCs.

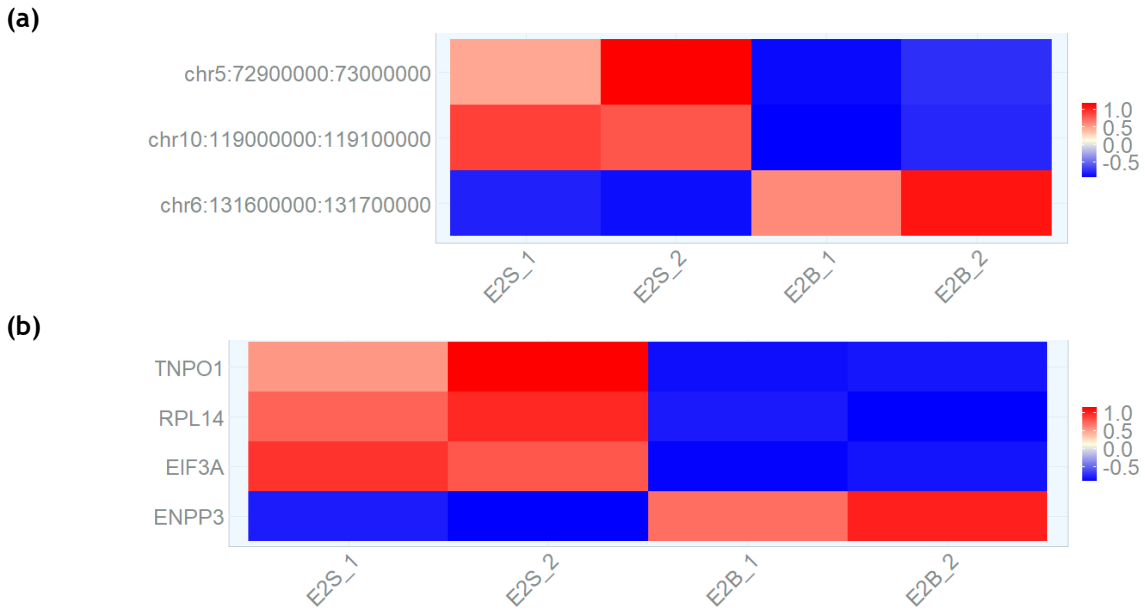



**Figure 6.6. E2 differential DSBs in 100 kbp regions and genes in GSCs and differentiated cells**

Differential DSB distribution identified using "DESeq2" analysis. Significant results were considered as a log2 fold change of >1 or <-1 and an adjusted p-value of <0.05. Differential analysis performed between E2 GSCs repeats 1 and 2 (E2S_1, E2S_2) and E2 differentiated cells repeats 1 and 2 (E2B_1, E2B_2). Higher DSB distribution indicated in red and lower DSB distribution indicated in blue as representative of z-scores. (a) Heatmap of significant differential DSB distribution between E2 GSCs and differentiated cells in 100 kbp regions. (b) Heatmap of significant differential DSB distribution between E2 GSCs and differentiated cells in genes.

Figure 6.7 displays the differential DSB distribution 100 kbp regions and genes that were significantly divergent in distribution across G7 GSCs and differentiated cells. Regarding significant regions with relative contrasting DSB frequency, there were 7 regions identified, all of which had significantly lower relative DSB distribution in GSCs compared to differentiated cells. Regarding genes, there were 36 genes identified with significant divergent DSB distribution. Of these, 34 genes were significantly lower in relative DSB distribution in GSCs compared to differentiated cells, with 2 genes being significantly higher in relative DSB distribution in GSCs (*KANSL2* and *RPS6*). Gene

Ontology ORA was performed, however did not show any over-representation of gene sets within these genes with significantly divergent relative DSB distribution. There were 5 genes identified that did not have corresponding gene symbols (*ENSG00000240401*, *ENSG00000251523*, *ENSG00000254687*, *ENSG00000266997*, and *ENSG00000224738*). All of these except for *ENSG00000266997* were associated with anti-sense or lncRNA class genes. The gene *ENPP3*, a protein-coding gene associated with nucleic acid binding and scavenger receptor activity, had significantly lower relative DSB distribution in GSCs compared to differentiated cells. This had also been the case in E2, where GSCs had a significantly lower relative DSB distribution than differentiated cells in this gene.

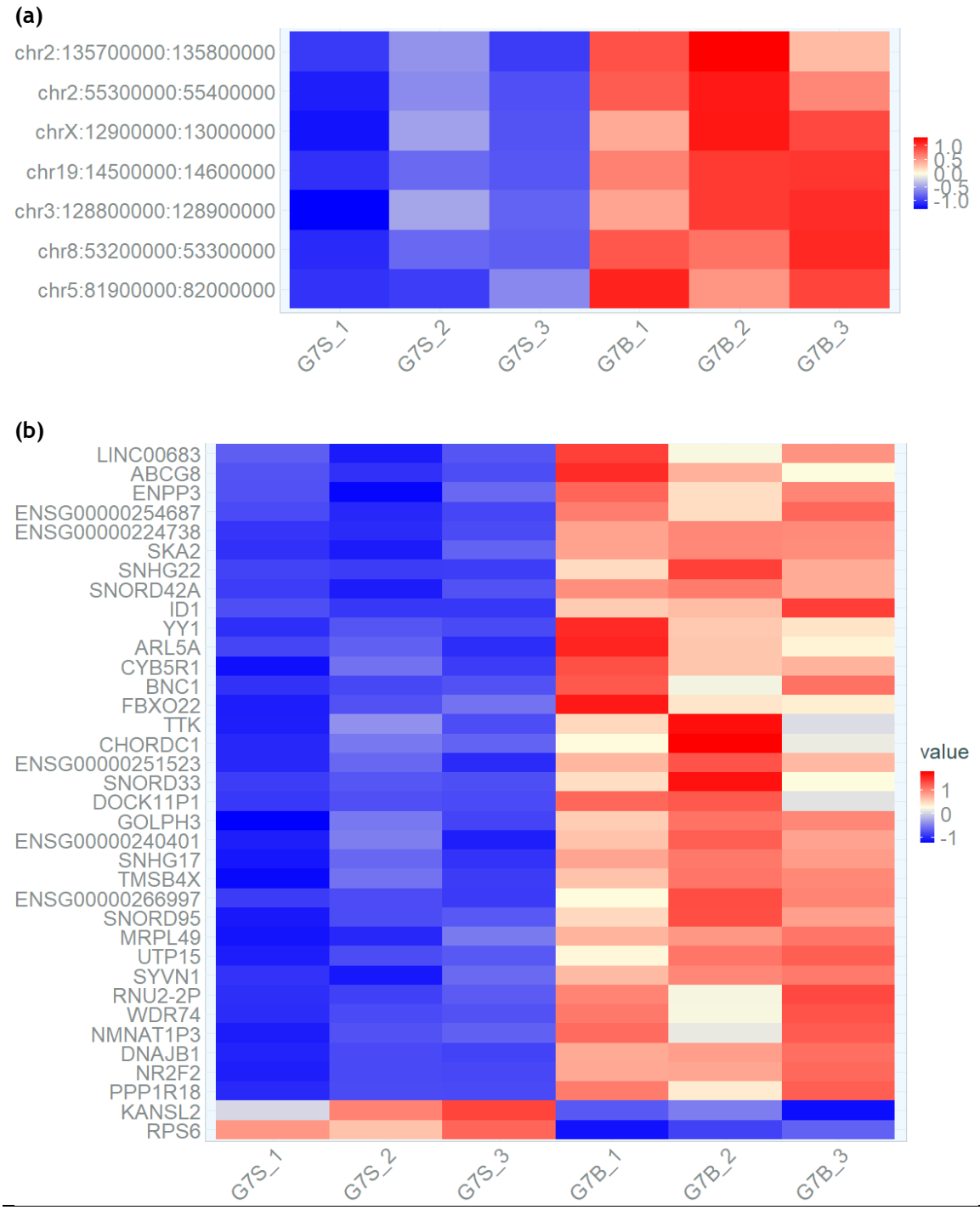



**Figure 6.7. G7 differential DSBs in 100 kbp regions and genes in GSCs and differentiated cells**

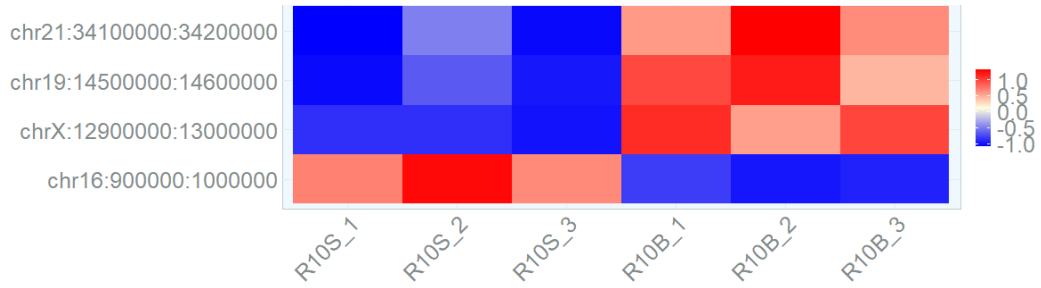
Differential DSB distribution identified using “DESeq2” analysis. Significant results were considered as a log2 fold change of >1 or <-1 and an adjusted p-value of <0.05. Differential analysis performed between G7 GSCs repeats 1-3 (G7S_1, G7S_2, G7S_3) and G7 differentiated cells repeats 1-3 (G7B_1, G7B_2, G7B_3). Higher DSB distribution indicated in red and lower DSB distribution indicated in blue as representative of z-scores. (a) Heatmap of significant differential DSB distribution between G7 GSCs and differentiated cells in 100 kbp regions. (b) Heatmap of significant differential DSB distribution between G7 GSCs and differentiated cells in genes.

Finally, Figure 6.8 shows the 100 kbp regions and genes with significant sites of divergent relative DSB distribution in R10 GSCs compared to differentiated cells. Four 100 kbp regions were identified as having a significantly contrasting relative DSB distribution between R10 GSCs and differentiated cells. Of these, 3 had a significantly lower relative DSB distribution in GSCs compared to differentiated cells with one 100 kbp region having a significantly higher relative DSB distribution in GSCs. Two of the regions of contrasting relative DSB distribution in R10 were also present in G7 sites; chrX:12,900,000-13,000,000 and chr19:14,500,000-14,600,000. In both cases the GSC DSB distribution was significantly lower in GSCs compared to differentiated cells. The region in chromosome X contained the *TMSB4X* gene, encoding an actin sequestering protein, which also appeared in both G7 and R10 as divergent in relative DSB distribution. Similarly, the gene *DNAJB1*, encoding a protein in the heat shock protein 40kD (Hsp40) family, was within the chromosome 19 region and present in both differential relative DSB distribution genes for G7 and R10. Regarding genes in R10 cells, there were 65 genes with significantly divergent relative DSB distribution between GSCs and differentiated cells. Twenty of these did not have corresponding gene symbols: *ENSG00000250643*, *ENSG00000234938*, *ENSG00000225879*, *ENSG00000220918*, *ENSG00000224738*, *ENSG00000217060*, *ENSG00000234337*, *ENSG00000280128 (uncategorized)*, *ENSG00000285645* (protein coding), *ENSG00000291258*, *ENSG00000259900* (protein coding), *ENSG00000282885* (uncategorised), *ENSG00000272973*, *ENSG00000259299*, *ENSG00000262898*, *ENSG00000278144* (miscRNA), *ENSG00000277873*, *ENSG00000274525*, *ENSG00000289835* and *ENSG00000276272*. Of these, 16 were within the lncRNA or pseudogene class. There were also additional named lncRNA genes present. These were present in both genes with significantly higher and lower DSB distribution in GSCs compared to differentiated cells. A Gene Ontology ORA was also performed in the genes with significantly higher and lower relative DSB distribution however there were no over-represented gene sets identified. The gene *RPL14*, encoding a large ribosomal subunit protein, had a significantly increased relative DSB distribution in R10 GSCs as well as E2 GSCs. Additionally, there were 12 genes that were shared as having significantly divergent relative DSB distribution between GSCs and differentiated cells in R10 and G7 lines. In addition to *DNAJB1* and *TMSB4X* the following genes also displayed lower relative DSB distribution across R10 and G7 GSCs compared

to differentiated cells: *CHORDC1, ID1, RPS6, PPP1R18, CYB5R1, ARL5A, BNC1, NMNAT1P3, DOCK11P1* and *ENSG00000224738*. These genes all had a significantly lower DSB distribution in GSCs compared to differentiated cells in both G7 and R10.

Overall, when considering differential analysis across DSB density in GSCs and differentiated cells, results did not show a definitive correlation between differences in DSB distribution across populations that related to gene function in any of the GBM lines. Notably there were a number of lncRNA and non-protein coding genes represented within the G7 and R10 genes identified through DESeq analysis.

(a)


chr21:34100000:34200000
chr19:14500000:14600000
chrX:12900000:13000000
chr16:900000:1000000
1.0
0.5
0.0
-0.5
-1.0
R10S_1
R10S_2
R10S_3
R10B_1
R10B_2
R10B_3

(b)

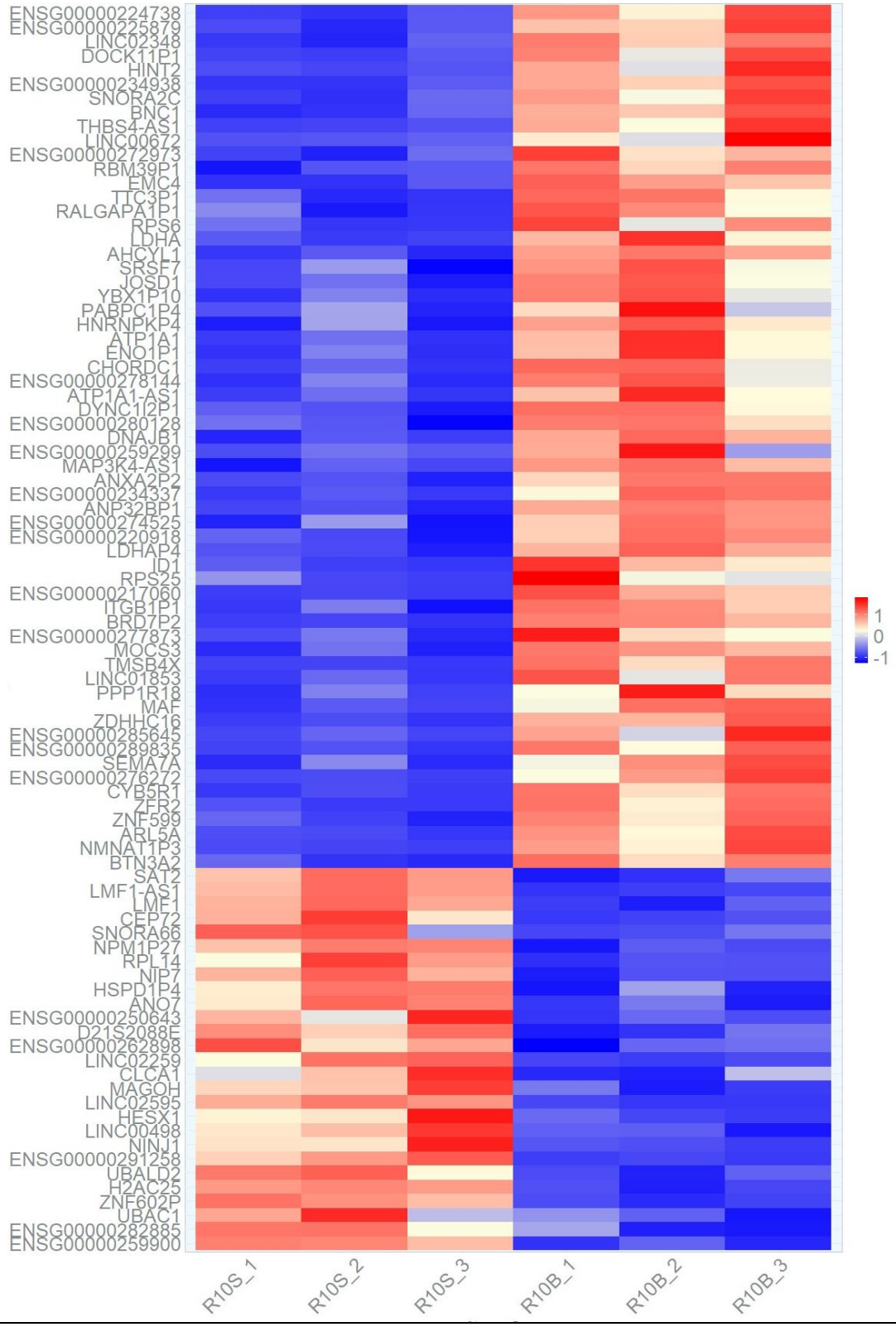

ENSG00000224738
ENSG00000225879
LINC02348
DOCK11P1
HINT2
ENSG00000234938
SNORA2C
BNC1
THBS4-AS1
LINC00672
ENSG00000272973
RBM39P1
EMC4
TTC3P1
RALGAPA1P1
RPS6
LDHA
AHCYL1
SRSF7
JOSD1
YBX1P10
PABPC1P4
HNRNPKP4
ATP1A1
ENO1P1
CHORDC1
ENSG00000278144
ATP1A1-AS1
DYNC1I2P1
ENSG00000280128
DNAJB1
ENSG00000259299
MAP3K4-AS1
ANXA2P2
ENSG00000234337
ANP32BP1
ENSG00000274525
ENSG00000220918
LDHAP4
ID1
RPS25
ENSG00000217060
ITGB1P1
BRD7P2
ENSG00000277873
MOCS3
TMSB4X
LINC01853
PPP1R18
MAF
ZDHHC16
ENSG00000285645
ENSG00000289835
SEMA7A
ENSG00000276272
CYB5R1
ZFR2
ZNF599
ARL5A
NMNAT1P3
BTN3A2
SAT2
LMF1-AS1
LMF1
CEP72
SNORA66
NPM1P27
RPL14
NIP7
HSPD1P4
ANO7
ENSG00000250643
D21S2088E
ENSG00000262898
LINC02259
CLCA1
MAGOH
LINC02595
HESX1
LINC00498
NINJ1
ENSG00000291258
UBALD2
H2AC25
ZNF602P
UBAC1
ENSG00000282885
ENSG00000259900
1
0
-1
R10S_1
R10S_2
R10S_3
R10B_1
R10B_2
R10B_3

**Figure 6.8. R10 differential DSBs in 100 kbp regions and genes in GSCs and differentiated cells**

Differential DSB distribution identified using “DESeq2” analysis. Significant results were considered as a log2 fold change of >1 or <-1 and an adjusted p-value of <0.05. Differential analysis performed between R10 GSCs repeats 1-3 (R10S_1, R10S_2, R10S_3) and R10 differentiated cells repeats 1-3 (R10B_1, R10B_2, R10B_3). Higher DSB distribution indicated in red and lower DSB distribution indicated in blue as representative of z-scores. (a) Heatmap of significant differential DSB distribution between R10 GSCs and differentiated cells in 100 kbp regions. (b) Heatmap of significant differential DSB distribution between R10 GSCs and differentiated cells in genes.

## 6.4 Discussion and conclusions

This chapter has investigated the relationship of DSBs in genes with gene expression as well as identifying areas of highly enriched euchromatin and corresponding DSB distribution. Finally, a comparative analysis of DSB distribution was performed across the three GBM cell lines comparing GSCs and differentiated cells in E2, G7 and R10.

### 6.4.1 DSBs at transcriptionally active genes

Actively transcribed genes were investigated using previously generated RNA-seq data on the three GSC lines E2, G7 and R10. This chapter identified a clear positive correlation between DSBs and transcription with the highest transcribed genes trending towards a higher DSB density. This was also true of genes affected by CNV which demonstrated a similar overall pattern in DSB density at the most highly transcribed genes even after adjustment for copy number. This was a helpful confirmation as genes affected by CNV could artificially over or underestimate the actual DSB density because of DNA abundance rather than DSBs. Increases in DSB density in highly transcribed genes have previously been described by Brambilla et al. in mesothelioma cell lines (Brambilla et al., 2020). In this paper they demonstrated that specifically the TSS of the top 10% of genes with highest expression correlated with an increase in DSBs compared to the least expressed genes of malignant mesothelioma cells. Transcription-associated DSBs have also been described in neural cells using a related DSB mapping technique: DSBcapture (Michel et al., 2022, Lensing et al., 2016). Interestingly their findings also demonstrated that transcription-related DSBs could trigger a DDR from p53. However, they were not able to show a DDR following induced RS.

Our data has demonstrated that when transcribed genes were taken as a group, there was an overall increase in DSBs in the most highly transcribed genes. Interestingly, this trend was not linear and DSB density appeared to be much less affected in the middle to lower transcribed genes. It was primarily the top 20-40% of expressed genes that appeared to exhibit the greatest increase in DSB density relative to genes with low expression. This was consistent across both GSCs and differentiated cells through the three cell lines. Whilst the observation of highly transcribed genes being prone to DSBs compared to genes with low transcriptional activity has been observed in other cell types this non-linear association of DSBs at lower levels of transcription has not been widely documented to our knowledge.(Ballarino et al., 2022, Michel et al., 2022, Brambilla et al., 2020). This may be related to the even division of transcriptional activity into quintiles rather than using a "cut off" TPM. It is possible that the fold change across highly transcribed genes displays a much greater biologically significant difference than the fold change across genes that are much less frequently transcribed, giving the steeper uptick at highly transcribed genes seen in Figure 6.2. This uptick only at the highly transcribed genes may also represent a relationship between highly transcribed genes and highly euchromatin-enriched sites where the intersection of both together has an additive impact on DSB density. Conversely, locations where there is low transcriptional activity may less commonly intersect with a euchromatin-enriched region and therefore DSB density may be largely similar across these sites.

### 6.4.2 Chromatin distribution and DSBs

The epigenetic DNA landscape is increasingly understood to have important implications for gene expression, function and survival (Martí et al., 2021, Aleksandrov et al., 2020, Mack et al., 2019). To interrogate DSB density at euchromatin-enriched sites, DSBs at high and low enrichment sites were investigated for GSCs and differentiated progeny. GSCs and differentiated G7 cells demonstrated a clear increase of DSBs occurring at the most enriched euchromatin peaks. However, this was much less pronounced across R10 and E2 lines. Whilst E2 GSCs and differentiated cells demonstrated an increase in DSB density at highly enriched sites, these results were biologically challenging to interpret, given the number of peaks with 0 DSBs/kbp. Similarly, whilst R10 had

a higher DSB density at highly enriched peaks, the differences were small and less significant than evidenced in G7 cells. However, it should be observed that the data followed a non-Gaussian distribution with a heavy right skew and the presence of peaks with few to no DSBs across all GSC and differentiated cells, particularly in E2 and R10 cell lines. The high number of peaks with no DSBs raised an interesting quandary. Regarding transcription, only transcribed genes were investigated for DSB density, however, because euchromatin peaks had been mapped specifically to these cell lines, these peaks were not excluded. These regions were in mappable areas and had been called as consensus peaks across the three ATAC-seq samples. This group of peaks with a DSB density of zero may represent an interesting set of locations. For example, these locations may be easily accessible but rarely transcribed due to cell activity, leaving these locations at low risk of physiologically induced DSBs from topoisomerases. Additionally, such regions which are easily accessible but not in frequent use may also be more rapidly repaired due to ease of access. It is interesting, however that G7 cells had a much more consistent relationship with DSB density in euchromatin than either E2 or R10. The reasons for this are not clear, however it may be related to the differences in DSBs across TSS seen in previous chapters. As previously described, G7 cells demonstrated an overall increase in DSBs around TSS. As has been established, euchromatin distribution across TSS also demonstrates enrichment. This, together with the increase of DSB density in the highly enriched euchromatin peaks in G7 may go some way to reflect the difference in G7 DSB density for euchromatin peaks compared to E2 and R10 cell lines. Overall, the variability of DSBs across these GBM lines indicated that euchromatin enrichment does not universally indicate high DSB density. Whilst euchromatin and transcriptional activity would be expected to go hand in hand, it is important to bear in mind that euchromatin is dynamic and also not specific to genes. Therefore, euchromatin enrichment will be dependent on cell cycle timing as well as the directed transcriptional activity and replication. Additionally, there is conflicting evidence on DSB repair in areas of heterochromatin and euchromatin. Whilst DSB repair is prioritised in highly transcribed regions, these sites have also been shown to be preferentially repaired via HR (Brambilla et al., 2020, Oster and Aqeilan, 2020). Regions of non-coding heterochromatin have also been cited as locations where repair is neglected and therefore can result in accumulation of breaks (Qiu, 2015).

Proximity to nucleosomes has also been established as a potential influencer on DSB location. DNA in euchromatin regions will likely have lower nucleosome occupancy. As previously discussed, nucleosome occupancy has been associated with a lower DSB density following IR and so it is possible that this association might also extend to endogenous DSBs (Brambilla et al., 2020).

### 6.4.3 Differential DSBs across GSCs and differentiated cells

So far, DSB density and DSB patterns across GSCs and differentiated cells have appeared highly similar. Given the contrast across GSC and differentiated cells with regards to treatment resistance, it is perhaps surprising to see little in the way of clear changes across these results regarding DSB distribution so far. Therefore, differential analysis across GSCs and differentiated cells was an important step in identifying if there were other sites of divergence in DSB distribution between GSC and differentiated cell lines. At a fundamental level, the clearest difference to note was the disparity in the total number of DSB reads detected via BLISS between GSCs and differentiated cells. Whilst BLISS-detected DSBs cannot be used as an absolute measure of DSB frequency, the use of UMIs allows for a comparison of DSB frequency within paired experiments comparing GSCs and differentiated cells. There was an increase in BLISS-detected DSBs in GSCs compared to differentiated cells. The cause of this higher DSB frequency in GSCs has a few possible reasons. In previous research, GSCs have been identified as having a subtly higher level of immunofluorescence-detectable DSBs compared to differentiated cells (Carruthers et al., 2018). This may be reflective of the established elevated levels of baseline RS that have also been seen in GSCs and so may reflect RS-induced DSBs. Furthermore, it is possible that DSB frequency could be affected by cell cycle dynamics. Cell cycle phase was not controlled for in these experiments, however DNA content was. Given that different DDR pathways are active at different phases of the cell cycle there may be variation in time to repair for GSCs compared to differentiated cells depending on cell cycle status (Averbeck et al., 2014, Tachon et al., 2018). As well as this, GSCs have been identified as preferentially utilising the HR DDR pathway which may also give rise to the disparity in DSB frequency between GSC and differentiated populations (Lim et al., 2014).

With regards to comparative DSB distribution across GSCs and differentiated cells, there appeared to be a relatively small number of changes in DSB distribution across these groups. In particular, the E2 cell line was very limited in alterations in DSB distribution between GSCs and differentiated cells, however this may be because E2 was n of 2 repeats. Overall, there were few significant changes in DSB distribution between 100 kbp regions and several of these reflected the DSB distribution seen in genes within these regions. This may be indicative that 100 kbp was not an optimal length of region or that the meaningful changes in DSB distribution occurred within genes. This would be in keeping with the fact that much of the variation in DSB distribution has appeared to occur either within or around genic locations. There were no gene ORA sub-groups identified to suggest particular sets of genes that were more or less susceptible to DSBs in GSCs compared to differentiated cells. This may indicate that function of genes is less important in predicting DSB distribution than the overall activity of the gene itself. Finally, many lncRNA genes were present displaying changes in DSB distribution between GSCs and differentiated cells. Whilst this was interesting to note, it is important to bear in mind that a meta-analysis on lncRNA described close to 60,000 lncRNA compared to around 30,000 protein coding genes (Iyer et al., 2015). This may therefore be partially representative of the abundance of lncRNA-encoding genes in the genome. LncRNA still remain an area of interest given their various emerging roles as cancer drivers (Huarte, 2015) and given that their role and activity within the cells remain somewhat opaque, there may yet be links to DSB distribution to lncRNA genes in GSCs. Overall, these data do not demonstrate stark changes of endogenous DSB distribution between GSCs and differentiated lines.

### 6.4.4 Summary of conclusions

- Across E2, G7 and R10 GSCs and differentiated cells, the genes with the highest expression have a higher DSB density than those with the lowest gene expression levels.

- Regions with the highest euchromatin enrichment had a higher DSB density compared to regions with the lowest euchromatin enrichment which was most apparent within G7 GSCs. However, even in the euchromatin peaks with the highest enrichment, there were sites with

few or no DSBs, indicating that euchromatin enrichment is not a sole driver of DSB density.

- GSCs demonstrate a higher overall number of detectable DSB reads compared to differentiated cells across E2, G7 and R10 lines.

- DSB distribution across GSCs and differentiated cells do not demonstrate changes that correlate with the diverging biological radioresistance seen between populations. Regions of divergent DSB distribution across GSCs and differentiated sites appear to primarily reflect DSBs occurring within genes rather than across regions.

# Chapter 7 Mapping DSBs before and after irradiation

## 7.1 Introduction

The previous chapters have outlined the DSB patterns across three GSC cell lines with reference to their differentiated progeny, neural cells and commercial cancer lines. Regarding GSCs and differentiated cells, there have been few differences in terms of contrasting DSB locations and hotspots. Therefore, DSB density distribution was explored following IR in this chapter. Finally, the more recent DSB mapping technique of INDUCE-seq (Dobbs et al., 2022) using one of the GBM lines was used in conjunction with BLISS to assess overall DSB frequency following IR in GSCs and differentiated cells.

### 7.1.1 DNA damage response to IR in GBM

As a means of describing the DDR to DSBs in our cell lines, immunofluorescence staining of IRIFs were used (Belyaev, 2010). Two DDR IRIF markers following IR; namely γH2AX and 53BP1, were used in cell lines R10 and E2. Both 53BP1 and γH2AX are commonly used as surrogates for measuring DSB DNA damage (Yang et al., 2015, Panier and Boulton, 2014, Valdiglesias et al., 2013, Rothkamm and Löbrich, 2003). With both 53BP1 and γH2AX IRIFs, it is established that their use as DSB surrogate markers do not equate to a single protein of interest but can indicate many colocalising proteins of the same type to an area of DSB damage. Indeed, γH2AX foci are known to span across megabases as part of the DDR (Noubissi et al., 2021b). Additionally, 53BP1 foci have also been known to organise into 53BP1 nuclear bodies as a means of handling DNA damage yet to be processed post-replication (Fernandez-Vidal et al., 2017). Whilst these foci can be used as surrogates for DNA DSB damage, it is important to bear in mind that these will span large chromatin domains rather than indicating single repair sites.

### 7.1.2 DSBs following IR

Whilst directly IR-induced DSBs per-se may not be realistically distinguishable from endogenous DSBs due to their relatively small number, IR-induced global changes may have important impacts on the DSB distribution and density.

Throughout this thesis, in previous chapters it has been demonstrated that there were several consistent regions of endogenous DSBs across repeats and, in some cases, across cell lines. Exposure to IR will undoubtedly change expression pattern and will stimulate the DDR which may result in differences in the previously observed regional patterns of DSB density and distribution. Given that GSC and differentiated cells have demonstrated differential survival following IR, determining whether differences in DSB patterns occur post-IR is pertinent to investigate.

### 7.1.3 Aims and outline

- Outline 53BP1 and γH2AX IRIF pattern in GSC and differentiated cell lines in R10 and E2 GSCs and differentiated cells following IR.

- Characterise BLISS-detected DSB density and changes in DSB distribution following IR in R10 GSCs and differentiated cells following IR and in E2 GSCs following IR.

- Quantify DSB density following IR using INDUCE-seq and compare with BLISS DSB density in R10 GSCs and differentiated cells following IR.

## 7.2 Materials and Methods

### 7.2.1 Ionising radiation induced foci analysis

The detailed procedure of immunofluorescence IRIF staining can be found in chapter 2. In brief, R10 GSCs and differentiated progeny were treated with 0 Gy or 10 Gy irradiation and fixed at 24 hours post-treatment. Cells were stained for DAPI, 53BP1 and γH2AX. IRIF numbers and integrated density per cell was measured using manual confocal microscopy.

**Equation 2 Calculated in ImageJ:**
*Integrated Density = sum of all pixel values in nucleus * area of one pixel*

The automated Opera Phenix™ high content screening system was used to measure foci and mean cell intensity in E2 GSCs and differentiated cells which were treated with 0 Gy or 10 Gy of irradiation and fixed at 6 hours post-

treatment. Mean fluorescence intensity of cell nuclei was calculated as part of the automated software analysis.

**Equation 3 Calculated in Opera:**
*Mean Fluorescence Intensity = Sum of the values of all nuclear pixels / Number of pixels in nucleus*

Cells were also stained for DAPI, 53BP1 and γH2AX. Both R10 and E2 experiments were performed in triplicate. Nuclear IRIF were counted per cell and measured across 3 repeats. The foci per nuclei and nuclear intensity mean of medians were statistically tested via ANOVA and post-hoc Tukey analysis. A p-value of $<0.05$ was considered significant.

### 7.2.2 BLISS-detected DSBs post-IR

For BLISS DSB mapping, cells were fixed and samples extracted as per chapter 2. Samples were processed in the BLISS pipeline described in chapter 2 into bedfiles for DSB locations. Differentially broken regions were investigated in 0 Gy and 10 Gy IR-treated cells using the "DESeq2" analysis pipeline as described previously in chapter 6 and with previous samples. DESeq2 FDR correction was performed using the BHP as part of the DESeq2 package analysis. The "Gene Ontology" RStudio package (Ashburner et al., 2000, Thomas et al., 2022) was used for ORA to identify any unifying themes or gene sets in the differentially broken genes. Differential analysis of DSBs was performed using genomic bins of a set region of 100 kbp which spanned the whole genome. Additionally differential analysis was performed on DSBs in genes. Differential analysis was performed to compare DSBs in samples treated with 0 Gy or 10 Gy. For R10 log2 fold change of $>1/<-1$ and an adjusted p-value of $<0.05$ was used. Given the few differences seen between 0 Gy and IR treated cells in E2, the significance threshold for E2 was also reduced to a log2 fold change of 0.5 to identify any other differences in pattern.

### 7.2.3 INDUCE-seq DSB mapping

For INDUCE-seq DSB mapping, cells were fixed and plated as described in chapter 2. Cells were sent for processing and alignment to the Reed lab who first established and published INDUCE-seq as a DSB mapping method (Dobbs et al., 2022). Two samples per condition were performed due to the limitation on plating requirements for processing. Cells were required to be set out in columns of 8 on a 96 well plate format. Therefore 4 conditions were investigated as a comparator to BLISS; R10 GSCs 0 Gy IR, R10 GSCs 10 Gy 24 hours, R10 differentiated cells 0 Gy IR and R10 differentiated cells 10 Gy 24 hours.

### 7.2.4 Total DSBs

The total number of DSBs per sample was calculated for BLISS and INDUCE-seq. For each, the fold change from 0 Gy to 10 Gy in total DSBs detected was calculated with the 0 Gy sample being normalised to 1.

**Equation 4**
*Fold Change = 10 Gy sample/0 Gy sample*

This was to account for the difference in total cells investigated per technique and to account for the differences in DSB yield and plating adherence within conditions.

## 7.3 Results

### 7.3.1 DNA damage response to IR in GBM

The use of IRIF has been demonstrated as an effective tool for measuring DSBs post-IR where it has been shown that a linear relationship to IR dose of IR induced DSBs can be drawn from DSB surrogate foci such as γH2AX and 53BP1 (Rothkamm and Löbrich, 2003). It has been estimated that exposure to 1 Gy produces roughly 35-40 DSBs per cell. Therefore, to characterise the general DSB response pattern in the analysed GBM lines IF IRIF were employed as DSB surrogate markers. Figure 7.1 displays representative images of dual-stained R10 nuclei with DAPI nuclear stain 24 hours following either 0 Gy or 10 Gy IR. Imaging of R10 cells demonstrated the presence of both 53BP1 foci and γH2AX foci at

baseline and following IR. Notably some cells also demonstrated pan-nuclear staining, predominantly following IR in both GSCs and differentiated cells.

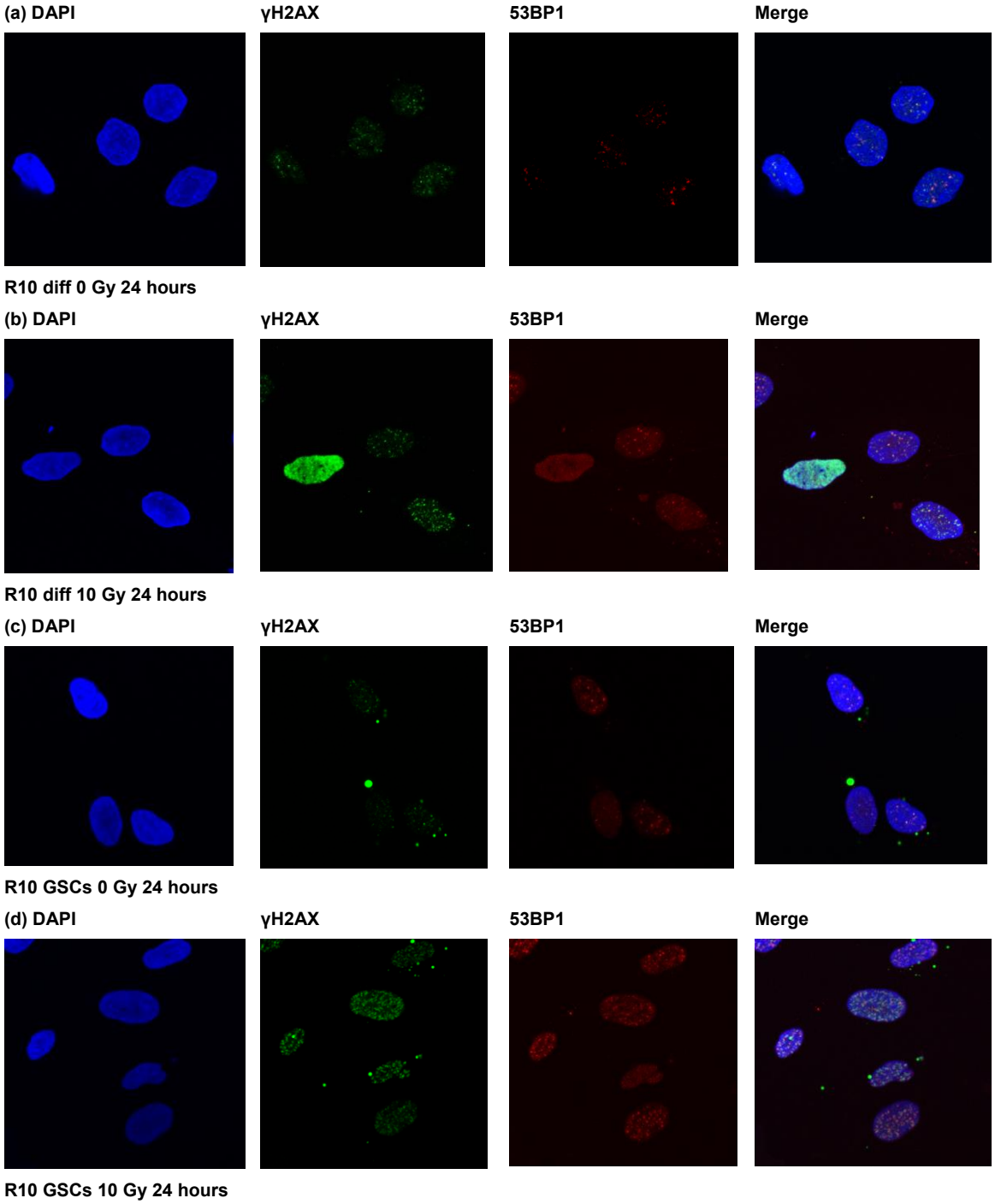



**Figure 7.1. Immunofluorescence imaging of ionising radiation-induced foci at 0 Gy and 10 Gy 24 hours in R10 GSCs and differentiated cells**

R10 cells fixed at 24 hours following 0 Gy or 10 Gy IR. Cell nuclei stained with DSB DDR markers 53BP1 and γH2AX and DAPI nuclear stain. From left to right columns: DAPI nuclear staining, γH2AX foci (green), 53BP1 foci (red), merged blue/red/green image. From top to bottom rows: (a) R10 differentiated cells 0 Gy IR, (b) R10 differentiated cells 10 Gy IR, (c) R10 GSCs 0 Gy IR, (d) R10 GSCs 10 Gy IR.

Figure 7.2 displays 53BP1 and γH2AX foci numbers per cell and integrated density values in R10 GSCs and differentiated cells following 10 Gy IR at 24 hours. The median 53BP1 and γH2AX foci per cell per condition repeat are displayed below. In both differentiated cells and GSCs, there was a trend towards increase in IRIF following 10 Gy IR at 24 hours. Both 53BP1 and γH2AX foci demonstrated a significant increase of foci in differentiated cells following IR. GSCs showed a significant increase in 53BP1 foci but a non-significant increase in γH2AX foci post-IR 10 Gy at 24 hours. In supplement to IRIF formation, overall 53BP1 and γH2AX abundance can be evaluated by determining the integrated density of the nuclear staining for both 53BP1 and γH2AX. Whilst IRIF have been validated as surrogate markers for DSBs in cells, integrated density can also be used as a broader marker of protein expression within the nucleus (Equation 2). Integrated density has also been demonstrated to increase linearly with increasing IR in γH2AX staining (Cai et al., 2009). Additionally, methods quantifying overall protein abundance have the advantage of including cells with pan-nuclear staining where foci may not be easily quantifiable. Both GSCs and differentiated cells showed a trend towards an increased γH2AX increased integrated density following IR, though this was non-significant in both groups. Measurement of 53BP1 integrated density in differentiated cells and GSCs showed minimal differences in nuclear integrated density following 10 Gy IR at 24 hours, though there was a variable spread across median repeats.

Overall, these findings indicated a trending increase in IRIF in differentiated cells and GSCs even at the later timepoint of 24 hours following IR exposure for both 53BP1 and γH2AX, though γH2AX was non-significant in GSCs. This may be reflective of the greater efficiency in DDR in GSCs compared to differentiated cells.

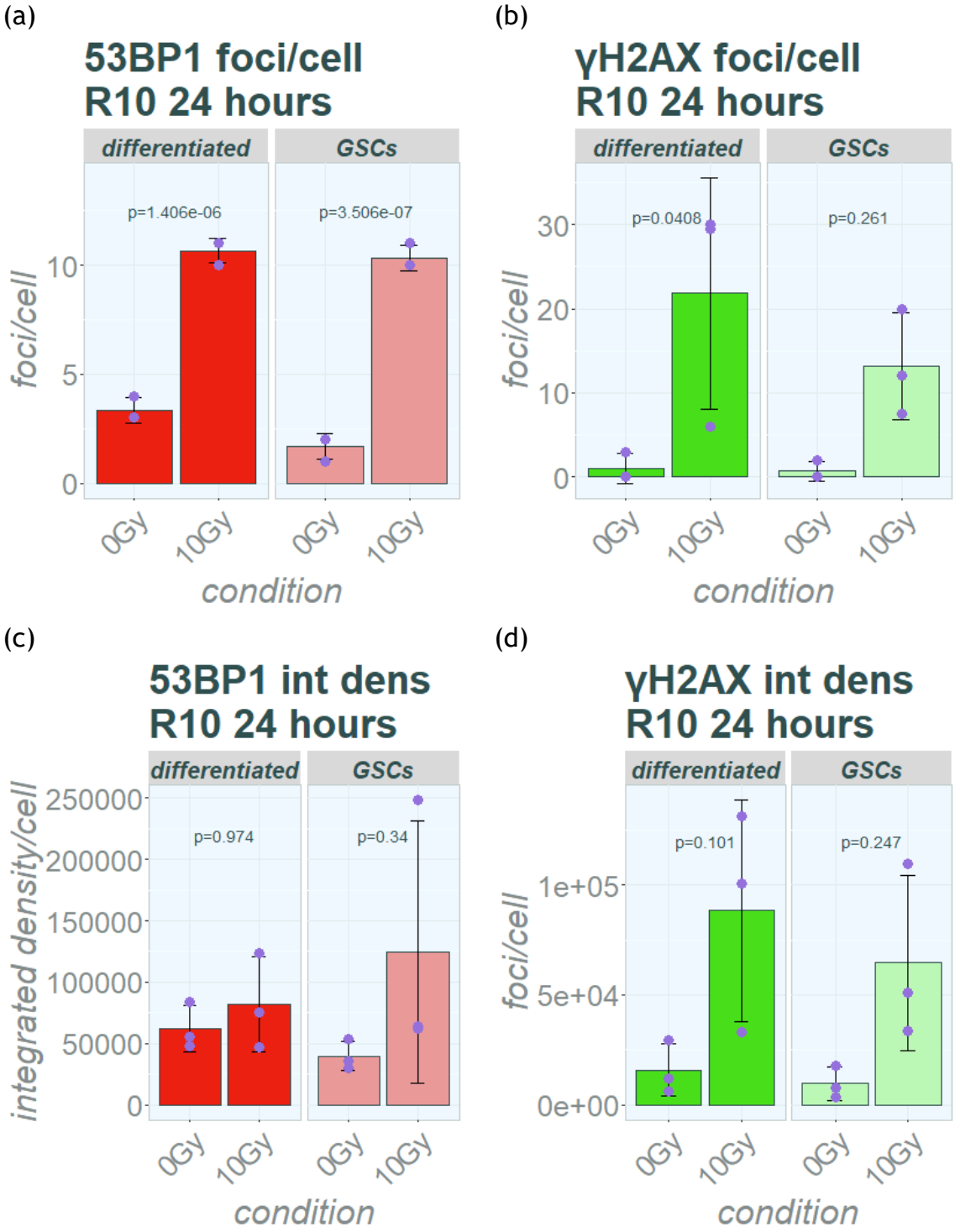



**Figure 7.2. 53BP1 and γH2AX in R10 following 10 Gy 24 hours post IR in GSCs and differentiated cells.**

R10 GSCs and differentiated cells fixed at 24 hours following 0 Gy or 10 Gy IR (n=3 experiments). Bar charts of means of median IRIF per cell nuclei and integrated density per cell nuclei (Equation 1). Medians of the individual experimental repeats are shown and represented as purple dots. Whiskers represent standard deviation across medians. R10 differentiated cells on left of each image ("differentiated": dark red or dark green). R10 GSCs on right of each image ("GSCs": pink or pale green). Statistical testing was performed by an ANOVA with post-hoc Tukey test. Adjusted p-values reported. (a) 53BP1 foci per cell in differentiated cells and GSCs. (b) γH2AX foci per cell in differentiated cells and GSCs. (c) 53BP1 integrated density per cell in differentiated cells and GSCs. (d) γH2AX integrated density per cell in differentiated cells and GSCs.

As with the R10, E2 IRIF were also used to describe the DDR to DSBs, this time at the earlier timepoint of 6 hours. E2 differentiated cells and GSCs were subjected to 0 Gy or 10 Gy and fixed 6 hours post exposure. Both GSCs and differentiated cells were stained following 10 Gy 6 hours post IR for DDR DSB markers 53BP1 and γH2AX (Figure 7.3). The Opera imager microscope was used for quantification of foci per cell nuclei and mean fluorescence intensity of cell nuclei for 53BP1 and γH2AX activity in E2 cells (Equation 3). The Opera imager provided high throughput image capture across multiple wells and an integrated analysis system as part of image processing in addition to providing alternative microscope availability to the Zeiss confocal microscope. Therefore, set up and fixation methods were different between R10 and E2 experiments which are detailed in chapter 2. Automated imaging software allowed for mean cell nuclear intensity measurements but not integrated density of cell nuclei meaning R10 and E2 experiments were not compared directly, however mean cell nuclear intensity has also been used to measure γH2AX activity (Noubissi et al., 2021b).

Both E2 GSCs and differentiated progeny demonstrated a significant increase in γH2AX foci at 6 hours following 10 Gy IR (Figure 7.3). Differentiated cells showed a significant increase in 53BP1 foci post-IR at 6 hours. E2 GSCs did not demonstrate a significant increase in 53BP1 foci at 6 hours post-IR 10 Gy. There was no significant increase in 53BP1 mean intensity post-IR in either GSCs or differentiated cells. This may reflect the difference in mean cell nuclear intensity and foci per cell nucleus where 53BP1 staining had a greater propensity of background staining which may have influenced the lack of difference between 0 Gy and IR-treated cells as mean cell intensity does not account for background staining. On the other hand, γH2AX mean cell intensity demonstrated a significant increase in both GSCs and differentiated cells at 6 hours following exposure to 10 Gy. Taken together, E2 showed an increase in γH2AX foci at 6 hours, though this was again a smaller increase in GSCs compared to differentiated cells which may indicate a more effective resolution of DSBs in GSCs compared to differentiated cells.

R10 differentiated cells had a higher number of 53BP1 foci at 0Gy compared to R10 GSCs. This was not seen in E2 differentiated cells compared to E2 GSCs where baseline 53BP1 foci were the same. There were no significant differences between baseline γH2AX foci across R10 differentiated cells and GSCs or across E2 differentiated cells and GSCs. These baseline differences in 53BP1 may reflect differences across GBM cell lines and may also indicate differences between DDR pathways and repair, where 53BP1 is more commonly associated with NHEJ repair over HR.

Overall, these findings showed an efficient engagement of DDR in the irradiated cells and demonstrated similar responses of foci induction following IR using both 53BP1 and γH2AX foci. However, integrated density and mean intensity across cell nuclei appeared less clear metrics in identifying IR induction with 53BP1 than in γH2AX.

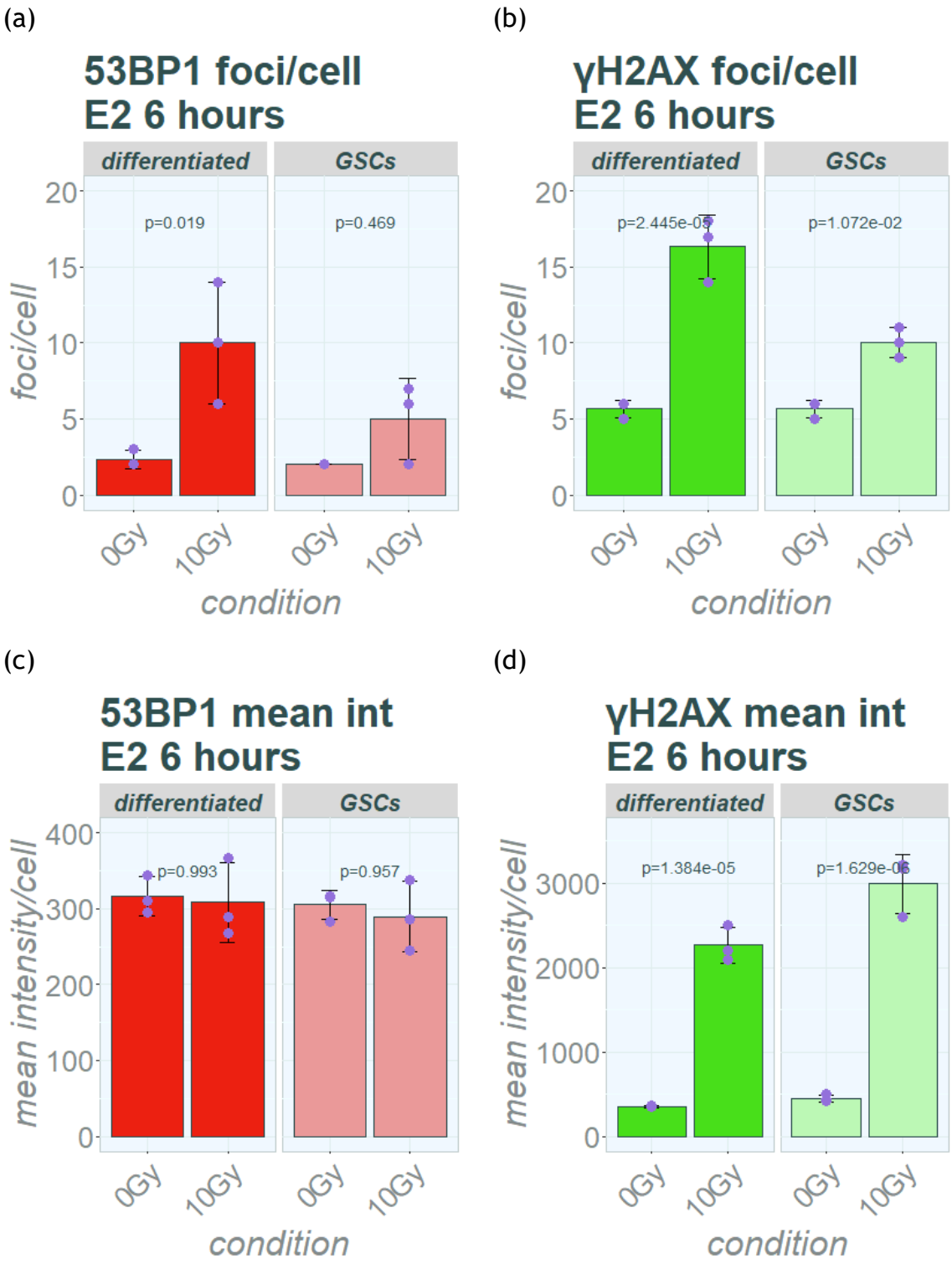



**Figure 7.3 53BP1 and γH2AX foci and mean fluorescence intensity in E2 following 10 Gy 6 hours post IR in GSCs and differentiated cells.**

E2 GSCs and differentiated cells fixed at 24 hours following 0 Gy or 10 Gy IR (n=3 experiments). Bar charts of the mean of medians for IRIF per cell nuclei and mean fluorescence intensity per cell nuclei (Equation 2). Medians of repeats displayed and represented as purple dots. Whiskers represent standard deviation across medians. E2 differentiated cells on left of each image ("differentiated": dark red or dark green). E2 GSCs on right of each image ("GSCs": pink or pale green). Statistical testing was performed by an ANOVA with post-hoc Tukey test. Adjusted p-values reported. (a) 53BP1 foci per cell in differentiated cells and GSCs. (b) γH2AX foci per cell in differentiated cells and GSCs. (c) 53BP1 mean fluorescence intensity per cell in differentiated cells and GSCs. (d) γH2AX mean fluorescence intensity per cell in differentiated cells and GSCs.

## 7.3.2 BLISS-detected DSBs following IR

As previously described, GBM presents a significant challenge in treatment due to radioresistance of GSCs, which are capable of repopulating tumours (Gimple et al., 2019). Hence, the investigation of DSB density and distribution following IR was of interest in determining how IR treatment influences DSB distribution in GSCs. To investigate this, BLISS-detected DSBs across GSCs following IR in R10 GSCs and differentiated cells were used in the first instance. Samples were irradiated with 10 Gy and collected at 24 hours post-IR.

### 7.3.2.1 R10 IR-treated cells demonstrate different yields of DSBs in GSCs and differentiated cells at 24 hours

As previously reported in chapter 6, GSCs tended toward a higher number of unique mappable DSB reads in comparison to their differentiated progeny. Here, DSB frequency in cells treated with 10 Gy was matched against 0 Gy-IR cells (Table *7.1*). Table 7.1 shows the read frequency and normalised break frequency across three repeats for R10 GSC and differentiated progeny cells. Fold change from 0 Gy to 10 Gy was calculated for GSCs and differentiated cells across repeats.

Overall, there were no significant differences between 0 Gy and IR treated groups. Differentiated lines demonstrated a consistent trend of an increase in DSBs following IR at 24 hours of between 1.2 to 1.8-fold increase in DSBs detected (Table *7.1*). On the other hand, two of the three GSC repeats showed a drop in DSBs following 10 Gy of 0.2 and 0.7 compared to 0 Gy-treated cells. Notably, the number of DSB reads from the GSC repeats had a minimum read yield of 1.4 million per condition repeat which was at least 2.7-fold higher than differentiated progeny repeats. The difference in fold change between normalised GSC-IR and differentiated-IR results approached significance with a $p=0.081$, however remained outwith the significance threshold.

**Table 7.1: DSBs detected in R10 by BLISS following 10 Gy IR 24 hours**

| | Total DSBs (0 Gy + 10 Gy) | 0 Gy IR | 10 Gy 24 hours | DSB fold change from 0 Gy to 10 Gy |
|---|---|---|---|---|
| ***R10 DIFF r1*** | 1,054,134 | 472,758 | 581,376 | 1.23 |
| ***R10 DIFF r2*** | 948,436 | 324,468 | 623,968 | 1.923 |
| ***R10 DIFF r3*** | 955,091 | 437,139 | 517,952 | 1.185 |
| ***R10 GSC r1*** | 10,109,185 | 5,959,428 | 4,149,757 | 0.700 |
| ***R10 GSC r2*** | 9,128,715 | 4,442,180 | 4,686,535 | 1.05 |
| ***R10 GSC r3*** | 7,557,108 | 6,152,960 | 1,404,148 | 0.228 |

| | *Fold change Diff vs Diff IR* | *Fold change GSC vs GSC IR* | *Fold change Diff IR vs GSC IR* |
|---|---|---|---|
| ***T-test*** | p=0.203 | p=0.291 | p=0.081 |

Total number of DSB reads detected per sample following 0 Gy IR or 10 Gy IR. Cells were fixed and collected at 24 hours. The table shows total number of reads per sample and reads per condition. The fold change from 0 Gy to IR was calculated by normalising 0 Gy treated cells to 1. Significance testing was performed using a t-test across fold-change groups.

The total number of reads across each repeat were collated together and displayed with annotated regions of interest in Figure 7.4. GSC read counts (annotated as stem) were higher than differentiated progeny lines, as noted in Table 7.1.

Overall, GSCs displayed a drop in BLISS-detected DSBs following IR whereas the differentiated cell lines displayed an increase in BLISS-detected DSB reads. Proportions of DSBs across annotated regions in differentiated cells were very similar across 0 Gy and IR-treated groups. GSCs showed a slightly higher proportion of DSBs in IR compared to 0 Gy-treated in exons and at intron/exon boundaries. Additionally, IR-treated GSCs showed proportionately slightly less DSBs at intergenic sites. These findings signalled a potential difference in the DDR to IR between GSCs and differentiated cells, influencing the DSB landscape in these populations.

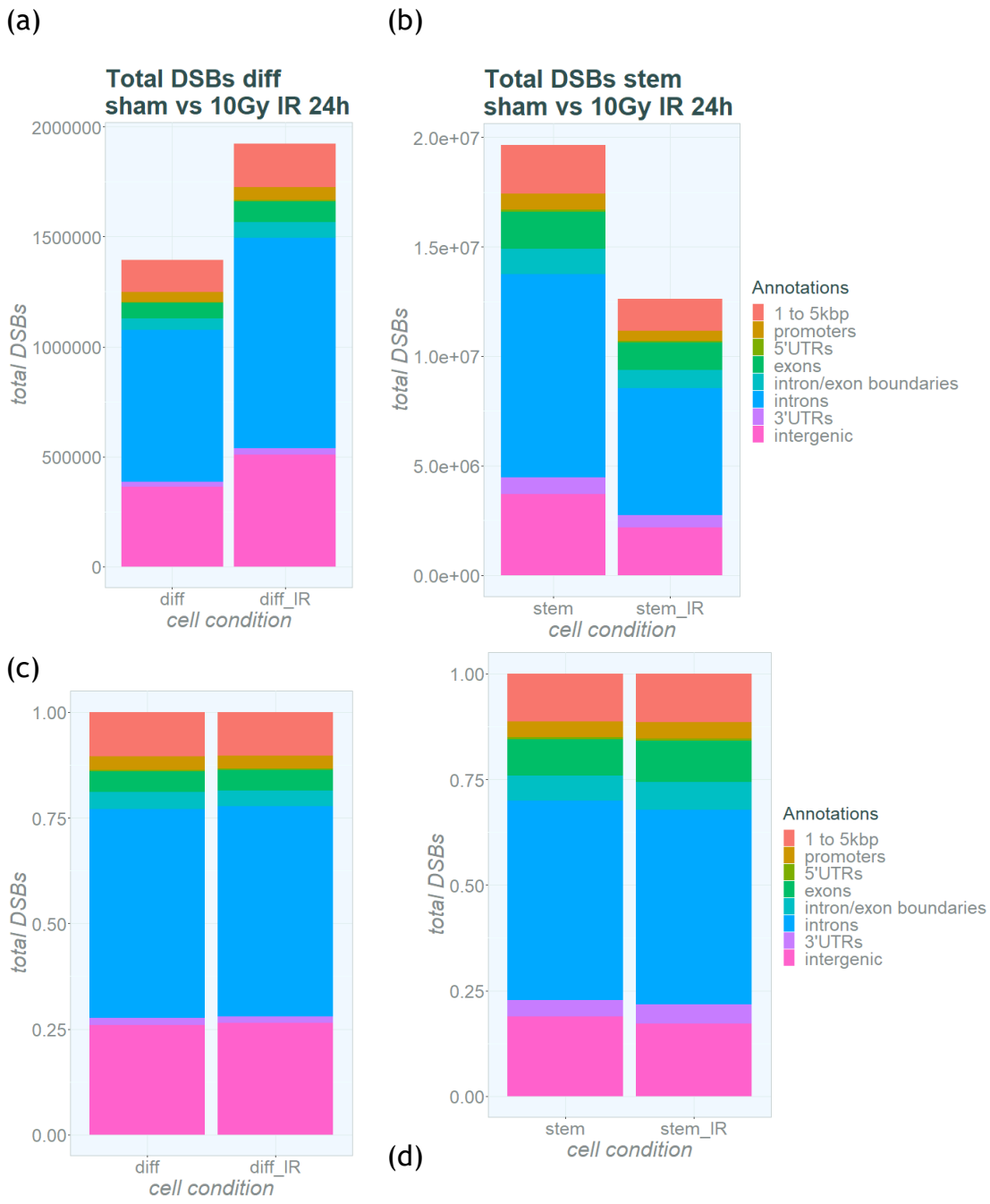


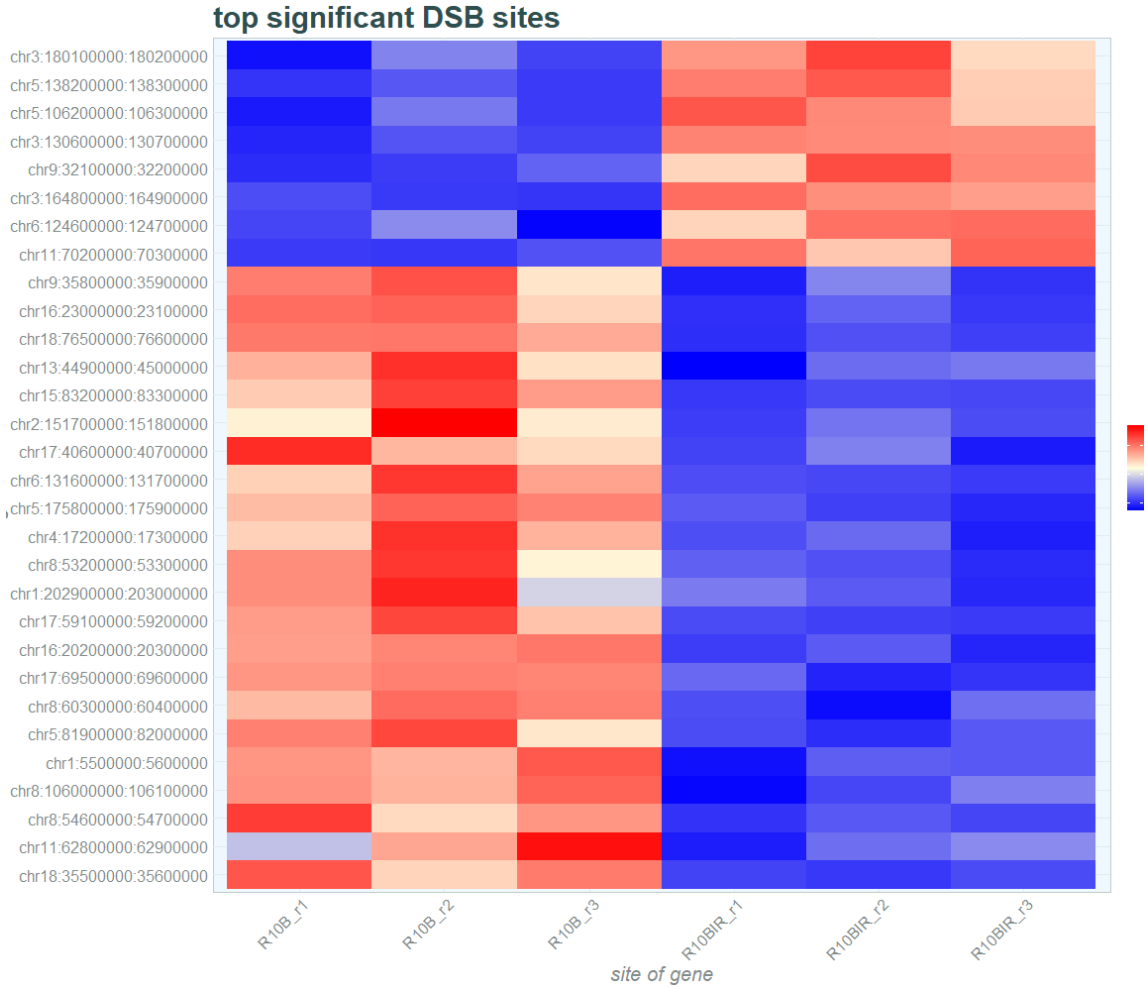

**Figure 7.4: Collated DSBs detected in R10 by BLISS following 10 Gy IR 24 hours**

The total number of DSBs across all samples (n=3). Annotated regions represent total number of DSBs occurring within interest regions. Top to bottom annotated regions: regions between 1-5 kbp from TSS, promotor <1 kbp from TSS, 5' UTRs, exons, intron/exon boundaries, introns, 3' UTRs and intergenic regions. Differentiated 0 Gy-treated cells are denoted as "diff", differentiated IR-treated cells are denoted as "diff_IR". GSC 0 Gy-treated cells are denoted as "stem", GSC IR-treated cells are denoted as "stem_IR". (a) R10 differentiated 0 Gy vs 10 Gy 24 hours cells total DSBs across 3 repeats. (b) R10 differentiated 0 Gy vs 10 Gy 24 hours cells proportions. (c) R10 GSCs 0 Gy vs 10 Gy 24 hours cells total DSBs across 3 repeats. (d) R10 GSCs 0 Gy vs 10 Gy 24 hours cells proportions.

Differentially broken regions were investigated in 0 Gy and 10 Gy IR-treated cells using DESeq2 analysis as with previous samples in chapter 6. Differentially

broken 100 kbp regions demonstrated 30 differentially broken regions in R10 differentiated 0 Gy vs 10 Gy and 6 differentially broken regions in R10 GSCs 0 Gy vs 10 Gy (Figure 7.5). Significant results were described as a log2 fold change of >1/<-1 and an adjusted p-value of <0.05. Between these differentially broken 100 kbp sites, there were no shared regions between the 6 GSC sites and the 30 differentiated cell sites. Differentially broken genes were also investigated, however there were no significant differentially broken genes in either GSCs or differentiated progeny that demonstrated a log2 fold change of >1/<-1 and an adjusted p-value of <0.05. Differences across 0 Gy and IR-treated cells were primarily confined to the total number of BLISS-detected DSBs, rather than changes in DSB distribution or regions and genes with divergent DSB locations. Overall, distribution of DSBs in GSCs and differentiated cells did not appear to demonstrate clear changes in DSB location following IR exposure, despite differences in overall DSB yield.

(a)




(b)

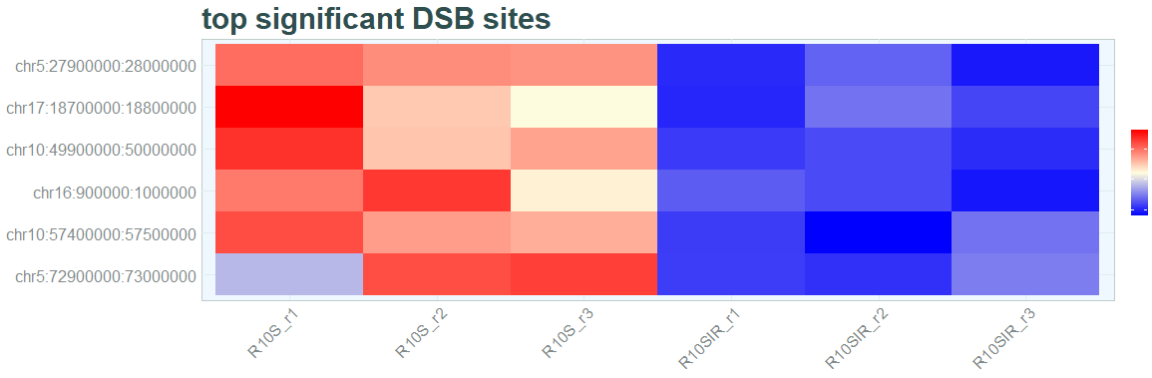



**Figure 7.5: Differential DSBs R10 0 Gy vs 10 Gy IR 24 hours**

100 kbp regions with differential DSB distribution between 0 Gy and irradiated conditions after 24 hours. Differentially broken region constraints: Log2 fold change of >1/<-1 and adjusted p value <0.05. (a) R10 differentiated cells 100 kbp regions with differential DSB frequency, 0 Gy vs 10 Gy IR. (b) R10 GSC 100 kbp regions with differential DSB frequency, 0 Gy vs 10 Gy IR.

#### 7.3.2.2 E2 GSC DSB yield decreases at 6 hours after 10 Gy

Having identified that the number of R10 GSC BLISS-detected DSBs after 10 Gy at 24 hours did not show an increase in BLISS-detected DSBs post-IR but rather a trending decrease, this highlighted the question as to whether this was also consistent with other cell lines. This result was not in keeping with the IRIF data and therefore warranted further investigation. To examine this, E2 GSCs were subject to 0 Gy or 10 Gy IR and fixed at the earlier timepoint of 6 hours. An earlier timepoint was used, since fewer IR-induced DSBs would be repaired 6 hours post-IR compared to 24 hours post-IR (Foray et al., 2005).

Table 7.2 displays the total number of BLISS-detected DSBs per sample in 0 Gy and IR-treated E2 GSCs. As noted in Table 7.2 there was an overall reduction in BLISS-detected DSBs post-10 Gy at 6 hours which was significant ($p=0.026$). This also appeared to show a more consistent overall reduction across repeats from 0 Gy to IR-treated GSCs with a fold change of 0 Gy to IR of between 0.4-0.68. The total number of reads detected per sample was also higher than in R10 GSC samples with the lowest number of E2 GSC DSB reads being 3.4 million in repeat 3 of E2 GSC IR-treated cells.

**Table 7.2 DSBs detected in E2 by BLISS following 10 Gy IR 6 hours**

| | Total DSBs (0 Gy + 10 Gy) | 0 Gy IR | 10 Gy IR 6 hours | DSB fold change 0 Gy to IR |
|---|---|---|---|---|
| ***E2 GSC r1*** | 18,019,540 | 12,772,035 | 5,247,505 | 0.411 |
| ***E2 GSC r2*** | 9,275,510 | 5,532,043 | 3,743,467 | 0.677 |
| ***E2 GSC r3*** | 10,361,338 | 6,915,280 | 3,446,058 | 0.499 |

| | *Fold change GSC vs GSC IR* |
|---|---|
| ***T-test:*** | p=0.026 |

Total number of DSB reads detected per sample following 0 Gy IR or 10 Gy IR in E2 GSCs. Cells were fixed and collected at 6 hours. The table shows total number of reads per sample and reads per condition. The fold change from 0 Gy to IR was calculated by normalising 0 Gy treated cells to 1. Significance testing was performed using a t-test across fold-change groups.

This is also represented graphically in Figure 7.6 which displays the total number of DSBs collated across 3 repeats with the respective annotations of interest. The total number of DSB reads following IR at 6 hours was lower than the 0 Gy-treated cells as seen with the overall fold change in Table 7.2. With regards to annotated regions of interest, there were no clearly discernible differences in annotated regions in 0 Gy compared to IR-treated E2 GSCs. The distribution of DSBs across annotated regions appeared broadly unchanged. The BLISS-detected DSBs demonstrated a significant reduction in frequency following IR. This was in contrast to the IRIF in which γH2AX foci demonstrated a significant increase following IR. Whilst IRIF showed an increase of DSBs induced following IR, the

global decrease in BLISS-detected DSBs may reflect that the endogenous DSB population is not represented through IRIF.

(a) (b)

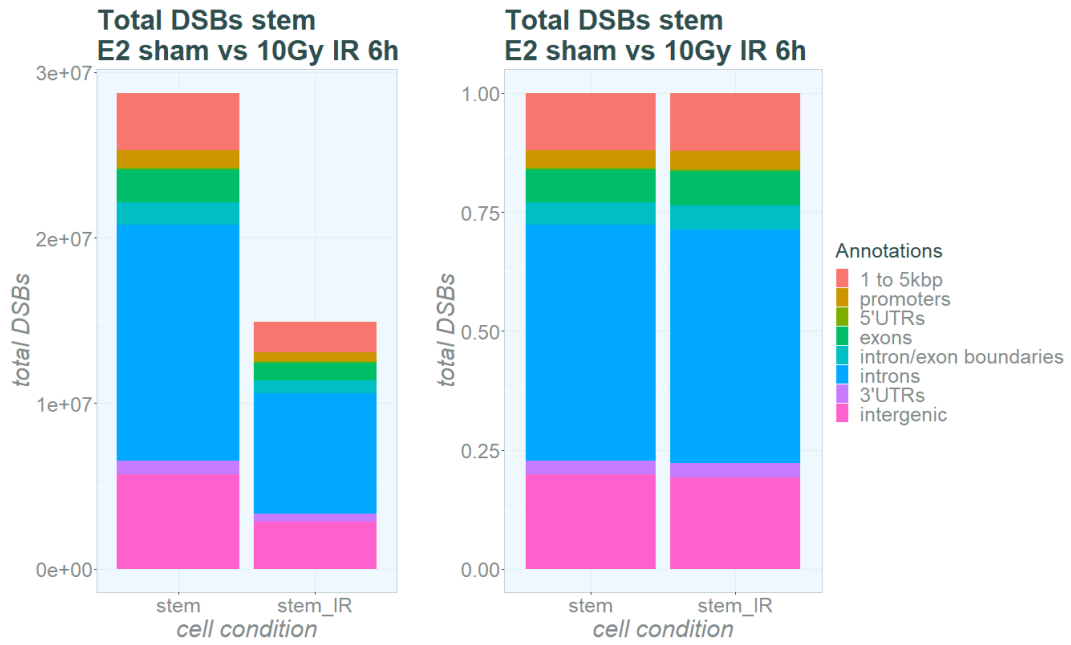



**Figure 7.6 Collated DSBs detected in E2 by BLISS following 10 Gy IR 6 hours**

The total number of DSBs across all samples (n=3). Annotated regions represent total number of DSBs occurring within interest regions. Top to bottom annotated regions: regions between 1-5 kbp from TSS, promotors <1 kbp from TSS, 5' UTRs, exons, intron/exon boundaries, introns, 3' UTRs and intergenic regions. (a) E2 GSCs 0 Gy vs 10 Gy 24 hours total DSBs across 3 repeats. (b) E2 GSCs 0 Gy vs 10 Gy 24 hours proportions.

Differentially broken regions between 0 Gy and IR E2 GSCs were also investigated. No 100 kbp regions or genes showed any significant differences with a log2 fold change of >1/<-1 and an adjusted p-value of <0.05. Therefore, to identify whether there was a smaller signal change that was masked by inter-sample variation, a log2 fold change of >0.5/<-0.5 with an adjusted p values of <0.05 was also investigated. There were a small number of regions at both 100 kbp sites and genic sites that showed some differences at these values (see supplementary figures). There were 20 differentially broken 100 kbp regions and 15 genes demonstrating a log2 fold change of >0.5/<-0.5 and an adjusted p-value of <0.05. E2 GSC differentially broken genes showed that there were 15 genes with a relatively higher DSB frequency in 0 Gy compared to IR-treated cells. However, a Gene Ontology ORA did not indicate any over-representation of gene

sets in the 15 genes with a log2 fold of >0.5/<-0.5. At E2 GSC differentially broken 100 kbp regions there were 9 regions that were relatively higher in DSBs in 0 Gy compared to IR-treated cells and 11 regions that were relatively lower in DSBs in 0 Gy compared to IR-treated cells. These findings were in keeping with the trend seen in R10 GSCs, where there was a global decrease in DSBs but not a change in differential distribution of DSBs following IR.

The above investigation of E2 GSCs following IR exposure showed a significant decrease in DSB yield which correlated with R10 GSC results. However, despite the significant overall decrease in DSBs, there was not an indication of differences in DSB location following IR. This suggested that, whilst IR had a significant impact in DSB frequency at the earlier DDR timepoint, the overall distribution of these DSBs did not change.

### 7.3.3 INDUCE-seq-detected DSBs post-IR demonstrate a yield pattern similar to BLISS-detected DSBs at 24 hours

As a means of directly addressing absolute DSB frequency to validate BLISS results alternative mapping methods for DSBs were also sought. Therefore the mapping of DSBs post-IR exposure was undertaken using INDUCE-seq in collaboration with the Reed Lab (Cardiff, UK) (Dobbs et al., 2022). Of particular importance, INDUCE-seq utilises Illumina flow cells for direct DSB end mapping, which the authors describe as equating to one DSB per read and removes the need for PCR-based library preparation that could introduce bias. Details of the INDUCE-seq method are available in the original paper by Dobbs et al and a brief summary of the preparation steps is also available in chapter 2. This was utilised as a comparator of the changes seen in DSB yield in BLISS-detected DSBs following IR in R10 GBM cells. For this reason, INDUCE-seq appeared to be a complementary technique to determine the absolute frequency of DSBs per sample and conditions.

To investigate DSB frequency following IR in both GSC and differentiated cells in parallel with BLISS, R10 cells were investigated at 24 hours post 10 Gy. Notably, the overall yield of INDUCE-seq-detected DSB reads was much lower in GSCs than in differentiated cells (Table 7.3). R10 GSCs had less than 0.5 million DSB reads in both 0 Gy samples and less than 0.4 million DSB reads in 10 Gy samples.

Conversely, R10 differentiated cells had 0.58 million and 0.9 million DSB reads in 0 Gy samples and 1.0-2.7 million reads in 10 Gy samples. Notably, INDUCE-seq required multiple washing steps and plating for adequate cell adherence of GSCs required optimisation to obtain sufficient DNA yield. It is therefore feasible that the differences in DSB read yield between GSCs and differentiated cells may relate to lower plate adherence of GSCs. This is discussed in more detail in chapter 2 methods. The yield of INDUCE-seq-detected DSBs in differentiated cells was higher following 10 Gy IR 24 hours in both repeats (Table 7.3). This was in keeping with the trending increase seen in BLISS-detected DSBs in differentiated cells in R10 at 24 hours following 10 Gy. In the GSC repeats, IR-treated cells did not demonstrate an increase in INDUCE-seq-detected DSBs, rather, both GSC IR-treated repeats had a small reduction in INDUCE-seq-detected DSBs. This was consistent with the BLISS-detected DSBs following IR where GSCs trended towards a reduction in DSB yield following 10 Gy at 24 hours in R10 and where E2 GSCs demonstrated a significant reduction in DSBs at 6 hours. However, there were no significant differences between 0 Gy and IR-treated cells in INDUCE-seq-detected DSBs (Table 7.3). For each cell line and repeat, the IR-treated cells were normalised to the total number of 0 Gy INDUCE-seq-detected DSB reads. The trend in INDUCE-seq-detected DSBs was consistent with the trend seen in R10 IR-treated cells where there was an increase in DSB yield following IR in differentiated cells and conversely a trending decrease in DSB yield in GSCs. Overall, INDUCE-seq-detected DSBs showed a trend towards higher DSB yield in differentiated cells at 24 hours and conversely in GSCs a reduction in DSB yield.

**Table 7.3 DSBs detected in R10 by INDUCE-seq following 10 Gy IR 24 hours**

| | 0 Gy IR | 10 Gy IR 24 hours | DSB fold change 0 Gy to IR |
|---|---|---|---|
| ***R10 DIFF r1*** | 586,423 | 2,717,324 | 4.634 |
| ***R10 DIFF r2*** | 909,399 | 1,072,794 | 1.180 |
| ***R10 GSC r1*** | 343,941 | 324,570 | 0.944 |
| ***R10 GSC r2*** | 477,682 | 363,141 | 0.760 |

| | *Fold change Diff vs Diff IR* | *Fold change GSC vs GSC IR* | *Fold change Diff IR vs GSC IR* |
|---|---|---|---|
| ***T-test*** | p=0.333 | p=0.667 | p=0.333 |

Total number of DSB reads detected per sample following 0 Gy IR or 10 Gy IR mapped by INDUCE-seq. Cells were fixed at 24 hours and sent for processing and mapping. The table shows total number of reads per sample and reads per condition. The fold change from 0 Gy to IR was calculated by normalising 0 Gy-treated cells to 1. R10 differentiated cells denoted as “R10 diff” and R10 GSCs denoted as “R10 GSCs”. T-test across fold-change groups was used for statistical testing.

### 7.3.4 BLISS and INDUCE-seq read counts show consistent overall patterns in DSB changes in GSCs and differentiated cells

Finally, the coverage and read count for R10 DSBs mapped by BLISS and INDUCE-seq was compared together as a means of understanding relevant differences and similarities across the two techniques’ results. Total number of reads across GSCs and differentiated cells in R10 between BLISS and INDUCE-seq demonstrated considerable variability. Table 7.4 and Figure 7.7 show the total number of detected DSB reads across BLISS and INDUCE-seq samples. Table 7.4 shows the total number of reads detected in both BLISS and INDUCE-seq for each repeat across 0 Gy-treated differentiated cells 24 hours, 10 Gy IR-treated

differentiated cells 24 hours, 0 Gy-treated GSCs 24 hours and 10 Gy IR-treated GSC 24 hours. As described, BLISS experiments were carried out in repeats of 3 and INDUCE-seq had repeats of 2. The BLISS-detected DSBs in GSCs were consistently higher than in differentiated progeny whereas this was reversed in INDUCE-seq. Whilst INDUCE-seq is a direct method of DSB measurement, differences in plating adherence between GSCs and differentiated cells may have accounted for DSB yield differences. It was also possible that this reflected a change in DSB quantity related to technical differences in measurement of DSBs in INDUCE-seq versus BLISS however this was less likely given the previously outlined plating optimisations in chapter 2. Comparisons of INDUCE-seq-detected DSBs across GSCs and differentiated cells should therefore be limited to fold change differences rather than a head-to-head comparison of total DSBs.

**Table 7.4 DSB read counts across BLISS and INDUCE-seq R10 10 Gy IR 24 hours**

| | | *BLISS DSB reads* | *INDUCE-seq DSB reads* |
|---|---|---|---|
| ***R10 DIFF r1*** | ***0 Gy*** | 472,758 | 586,423 |
| | ***IR*** | 581,376 | 2,717,324 |
| ***R10 DIFF r2*** | ***0 Gy*** | 324,468 | 909,399 |
| | ***IR*** | 623,968 | 1,072,794 |
| ***R10 DIFF r3*** | ***0 Gy*** | 437,139 | |
| | ***IR*** | 517,952 | |
| ***R10 GSC r1*** | ***0 Gy*** | 5,959,428 | 343,941 |
| | ***IR*** | 4,149,757 | 324,570 |
| ***R10 GSC r2*** | ***0 Gy*** | 4,442,180 | 477,682 |
| | ***IR*** | 4,686,535 | 363,141 |
| ***R10 GSC r3*** | ***0 Gy*** | 6,152,960 | |
| | ***IR*** | 1,404,148 | |

The total DSB read counts detected per sample following 0 Gy IR or 10 Gy IR in differentiated cells and GSC repeats. BLISS samples n=3 experimental repeats. INDUCE-seq sample n=2 experimental repeats. R10 diff r1-r3: R10 differentiated cells repeats 1-3. R10 GSC r1-r3: R10 GSCs repeats 1-3. 0 Gy: 0 Gy treated cells 24 hours. IR: 10 Gy treated cells 24 hours.

Figure 7.7 shows the mean of total DSB reads across BLISS and INDUCE-seq for R10 cells. As previously noted, DSB yield was higher in GSCs in BLISS than in INDUCE-seq (Figure 7.7). Conversely, differentiated cells had a greater number of detectable DSB reads using INDUCE-seq than with BLISS. Total cell number requirement across BLISS and INDUCE-seq should be taken into consideration with these results in addition to the differences in library preparation for both techniques. This is because BLISS utilises 2 million cells/condition/repeat and INDUCE-seq utilises 0.2 million cells/condition/repeat. Despite the differences in DSB yield, both BLISS and INDUCE-seq R10 differentiated lines showed a trend towards an increase in DSBs following 10 Gy IR at 24 hours. Conversely, GSCs did not demonstrate an increase in DSBs at 24 hours post IR. Rather, when collating the total DSBs across repeats, GSCs demonstrated a non-significant decrease in DSBs following exposure to 10 Gy at 24 hours.

Overall, INDUCE-seq and BLISS detected DSBs indicated that both could identify a similar trend in DSB yield across IR-treated populations. This was further evidence that there were true underlying differences in cell responses to IR

between GSCs and differentiated populations resulting in the contrasting DSB yields following IR.

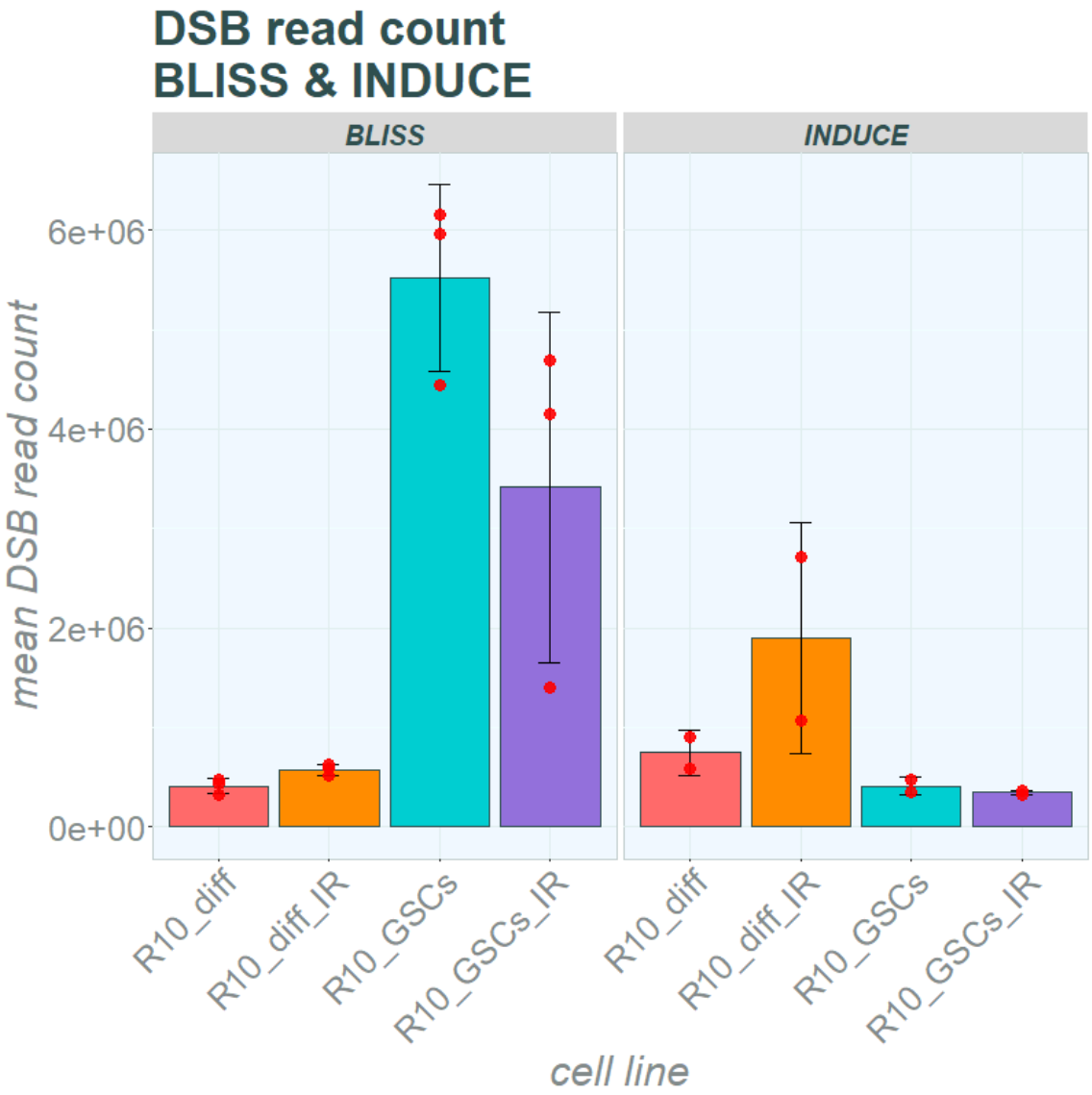



**Figure 7.7 DSB read counts across BLISS and INDUCE-seq R10 10 Gy IR 24 hours**

Bar chart of the mean DSB reads mapped by BLISS (left) and INDUCE-seq (right). Number of cells/repeat/condition in BLISS: 2 million cells. Number of cells/repeat/condition in INDUCE-seq: 0.2 million cells. Bar charts display the mean number of mapped DSB reads. Total DSB reads per repeat displayed as red dots, whiskers represent standard deviation. BLISS (left) n=3 experimental repeats, INDUCE-seq (right) n=2 experimental repeats. R10_diff: R10 differentiated 0 Gy-treated 24 hours in red, R10_diff_IR: R10 differentiated 10 Gy-treated 24 hours in orange, R10_GSCs: R10 GSC 0 Gy-treated 24 hours in turquoise, R10_GSCs_IR: R10 GSCs 10 Gy-treated 24 hours in purple.

## 7.4 Discussion and conclusions

This chapter has given an overview of DSBs in IR-treated GSCs. The quantification of DSBs in GSC and differentiated cells following mid to late DDR

timepoints have been described using IRIF. Additionally, the mapping of BLISS-detected DSBs following IR at mid to late DDR timepoints has been investigated. Finally, the DSB mapping technique INDUCE-seq, was utilised to quantify DSB frequency in GSC and differentiated cells as a validating method in exploring DSBs following IR. It is well established through IRIF data that DSBs increase significantly following IR in both GSCs and differentiated cells (Carruthers et al., 2018). However, the findings from BLISS and INDUCE-seq-detected DSBs run contrary to this established pattern of DSBs. The trend of lower DSBs in GSCs across both BLISS and INDUCE-seq data indicated an overall reduction in DSBs following exposure to IR. The contrast in trend of DSBs detected in IRIF data compared to BLISS and INDUCE-seq provides an interesting topic for discussion. Differentiated cells and GSCs demonstrate divergent DSB frequency post-IR in both BLISS and INDUCE-seq mapping methods. Given that this appears in part contradiction to established IRIF the interpretation of this IR-treated cell data is somewhat complex. This discussion will outline some of the potential explanations for this interesting divergence in GSCs and differentiated cell DSBs following IR.

### 7.4.1 Ionising radiation-induced foci as markers of DSBs

Immunofluorescent DDR DSB markers are commonly used as surrogate markers for DNA DSB damage (Banáth et al., 2004, Panier and Boulton, 2014). Using IF in R10 and E2, GBM cells were imaged using two traditional IRIF DSB marker surrogates: 53BP1 and γH2AX (Yang et al., 2015, Panier and Boulton, 2014, Valdiglesias et al., 2013, Rothkamm and Löbrich, 2003). R10 GSCs and differentiated cells treated with 10 Gy and fixed at 24 hours showed an increase in IRIF, across both 53BP1 and γH2AX, with a significant increase in differentiated lines. Similarly, R10 GSCs showed a significant increase in 53BP1 foci and a trending increase in γH2AX post-IR.

In E2 lines, the earlier 6-hour timepoint also demonstrated a significant increase in 53BP1 and γH2AX foci in differentiated cells. GSCs showed a significant increase in γH2AX foci at 6 hours though only a trend toward an increase in 53BP1 foci. The γH2AX measurement of mean cell intensity at 6 hours was able to also demonstrate a significant increase in both E2 GSCs and differentiated progeny at 6 hours but mean cell intensity appeared a poorer indicator for 53BP1

abundance. This data is broadly supportive of what is known of the GSC-differentiated cell DDR responses to IR that have been previously documented in literature (Carruthers et al., 2018). For example, Dr Ross Carruthers demonstrated in the PhD thesis “Response to ionising radiation of glioblastoma stem-like cells” that differentiated cells showed a slower resolution of γH2AX foci following IR compared to GSCs cells.

It is known that IRIF detected by confocal microscopy represent many proteins colocalising to a single site. These foci are also made up of “nano” foci which are visible using super resolution microscopy (Shibata and Jeggo, 2020, Hausmann et al., 2018). At baseline for both E2 and R10 cell lines, there were few detectable IRIF on IF staining. This might initially appear in contradiction with the previous BLISS data where there are many endogenous DSBs detected. However, there are some important factors to bear in mind when considering IRIF data. Firstly, whilst IRIF have been identified as effective methods for measuring DSBs in relation to IR, they remain an indirect means of measuring DSBs. This is because, though they colocalise to IR-induced DSBs, they are still DDR proteins. Therefore any interference to the DDR that directly impacts on γH2AX or 53BP1 activity may also interfere with results from foci imaging. In addition, confocal microscopy is limited to detecting large foci and therefore any “nano” foci that are formed at potential endogenous DSBs may not be identifiable using this type of imaging (Qian et al., 2024). Therefore, endogenous DSBs that do elicit a DDR from γH2AX or 53BP1 may not be easily picked up using this standard DSB detection method. Importantly, IRIF may not successfully detect all endogenous DSBs. As discussed in previous chapters, there is indication that a number of the endogenous DSBs detected are related to transcription and in particular TTS. This involvement of TTS implicates topoisomerases as potential culprits for inducing endogenous DSBs in these GBM lines. As discussed, TopIIß is an important player in the induction of physiological DSBs as a means of alleviating torsional stress on the double helix during transcription and replication. Type II topoisomerases do not appear to directly interact with γH2AX or 53BP1. However, when inhibited using increasing doses of etoposide γH2AX foci will form (Sunter et al., 2010). Therefore it may be that, though there is a wide pool of endogenous DSBs within these GBM cells, a high

proportion of physiologically-induced DSBs such as topoisomerase-induced DSBS may not be detectable through use of IRIF imaging.

### 7.4.2 Differential DSBs in BLISS-detected DSBs following IR

BLISS-detected DSBs were first mapped across R10 cells following IR at 24 hours to determine late timepoint changes across the genome in DSB density, since unrepaired DSBs at late timepoints are proposed to be highly lethal and are therefore of most interest to our investigations (Noda et al., 2012). Differentiated progeny R10 cells demonstrated a non-significant increase in BLISS-detected DSBs following IR 10 Gy at 24 hours. However, this was not evident in the R10 GSC lines, with two of the three repeats demonstrating a decrease in DSBs post IR. DSB distribution appeared broadly similar pre and post IR in GSCs and differentiated cells.

Though there were few locations or genes that demonstrated significant changes in DSBs in GSCs following IR this may be due to the challenges of accurately isolating IR-induced DSBs in this form of data. Repair kinetics assume that the majority of DSBs post-IR are repaired within the “immediate-early” to “early” phase repair stages (Kieffer and Lowndes, 2022). It has been estimated that 1 Gy of radiation induces roughly 40 DSBs in a cell and increases linearly with increasing IR. Therefore at 10 Gy, it would be estimated that 400 DSBs would be induced by IR (Rothkamm and Löbrich, 2003). Of these, upwards of 95% will have undergone repair by 24 hours. Therefore, the number of DSBs that would be directly induced by IR remaining unrepaired at 24 hours would be less than 5% of the detectable DSBs recorded. Additionally, BLISS appears to be capable of identifying many endogenous DSBs and therefore the sheer number of endogenous DSBs identified may make it impractical to distinguish these endogenous DSBs from IR-induced DSBs. The overwhelming number of endogenous DSBs alongside the relatively small number of IR-induced DSBs following IR makes feasibility of distinguishing between these two classes of DSBs challenging and this may not be possible with the capability of current technology. A potential means to address this could be to identify DSBs that appear specific to endogenous DSB activities such as those associated with TTS regions. These TTS DSBs could be associated with topoisomerase activity and therefore might allow for some partial separation of DSB populations. However,

this does not take into account the possible interactions of transcriptional activity and IR exposure which could confound these results.

### 7.4.3 GSC DSBs in BLISS-detected DSBs and INDUCE-seq-detected DSBs following IR and disparities with IRIF results

Given the unexpected decrease in DSBs post IR in GSCs, this raised the question of whether other DSB mapping techniques would encounter similar findings. INDUCE-seq provided an excellent opportunity to utilise another DSB sequencing method as a complementary approach to BLISS. In addition to providing an alternative mapping method for DSBs, INDUCE-seq also has the ability to give a more direct readout of absolute DSB number due to the use of flow-cell sequencing (Dobbs et al., 2022). Again, when looking at total INDUCE-seq-detected DSB reads post-IR in R10, there was a trend of DSB increase in differentiated cells but a decrease in GSC DSBs which was in line with the BLISS-detected DSB results. These initial results from INDUCE-seq supported the finding that BLISS demonstrated an increase in DSBs post-IR in differentiated cells and a contrasting decrease in GSCs, suggesting a robust biological finding, rather than a technical discrepancy.

Both BLISS and INDUCE-seq-detected DSBs GSCs following IR showed a reduction in DSB frequency in contradiction to IRIF data, which is difficult to explain in the context of current radiobiological dogma. GSCs have previously been described as having elevated levels of ATR and Chk1 expression in comparison to their differentiated progeny (Carruthers et al., 2018). Therefore GSCs, in possession of upregulated ATR and Chk1 expression, are not only better equipped to respond rapidly to IR insults but are also able to signal for cell cycle arrest. This tighter control of cell cycle checkpoints may allow these GSCs to pause the cell cycle, preventing initiation of further activity that could initiate further DSBs (Bao et al., 2006, Ahmed et al., 2015).

As described, there appear to be a number of endogenous DSBs related to transcriptional activity which may not be detectable by conventional γH2AX or 53BP1 IF. It has been demonstrated that IR can cause repression of transcription in other cell types such as in the paper by Narayanan et al. (Venkata Narayanan et al., 2017) where transcription was repressed following IR through p53 activity

in normal human fibroblasts, though p53 is commonly mutated in GSCs and there have been single nucleotide polymorphisms identified in our GSC lines (See Dr E. Clough's thesis "Investigating mechanisms and indicators of sensitivity to replication stress-targeting therapies in glioblastoma"). However, transcriptional pausing has also been seen to occur following IR in nasopharyngeal cancer and has been associated with radioresistant phenotypes (Liu et al., 2023). It may be that in response to IR, GSCs repress or pause transcriptional activity as part of the DDR. This repression of transcription could result in an overall decrease to the BLISS and INDUCE-seq-detectable DSBs given that many of these may be associated with transcription. Whilst IRIF do show an increase in DSBs post-IR, the overall decrease in BLISS or INDUCE-seq-detected DSBs may actually be a reflection of transcriptional downregulation which may be where the majority of DSBs are accounted for.

However, another explanation of the overall reduction in DSBs following IR in these GSCs is the aberrantly upregulated DDR seen in these cells (Carruthers et al., 2018, Bao et al., 2006). Whilst IRIF do not follow the pattern of DSBs in BLISS and INDUCE-seq-detected DSBs, GSCs demonstrated fewer γH2AX foci than differentiated progeny following IR in both E2 and R10 GSCs. This is in keeping with previous data demonstrating an upregulation of DDR in GSCs. The priming of the DDR through elevated levels of RS has been proposed as the mechanism for GSCs ability to survive potentially lethal IR-induced DSBs. It is possible that this priming of the DDR pathways allows for GSCs to respond to IR exposure by not only repairing the IR-induced DSBs but also efficiently addressing any endogenous DSBs by mobilising these DDR proteins.

It is feasible that this trend of decreasing DSBs post-IR in GSCs may be related to tight cell cycle control in combination with transcriptional downregulation as well as the efficient DDR, resulting in GSC survival and radioresistance.

### 7.4.4 Differentiated cell BLISS-detected DSBs and INDUCE-seq-detected DSBs following IR

However, it is interesting to note that the frequency of DSBs post-IR in GSCs and differentiated cells appeared to differ in trend. Whilst the above discussion

highlighted potential reasons for the downward trend of DSBs in GSCs, it does not address the increase in trend of DSBs in differentiated cells.

BLISS and INDUCE-seq-detected DSBs did not follow the same pattern of decreased DSB frequency following IR in R10 differentiated progeny cells but rather showed an increase in DSBs post-IR. As has previously been highlighted, these cells have a lower expression of Chk1 and ATR compared to GSCs which will impact on their capacity to effectively regulate cell cycle checkpoints compared to GSCs (Carruthers et al., 2018). Therefore these differentiated cells may not be adequately able to control the cell cycle and thereby prevent further cell activity. This may mean that, unlike GSCs, replication and transcription will continue to proceed, even in the face of sublethal or lethal damage. This damage, if left unrepaired, could lead to further DSBs from replication fork stalling, replication-transcription collisions and further DSBs being generated through cell division. Additionally, whilst DSBs are the primary means of how IR induces cell death, it is known that many more SSBs will be generated through IR exposure. Whilst these are usually easily repairable, if left unrepaired, these SSBs can be transformed into DSBs. This can occur through replication run-off, where replication machinery encounters an unrepaired SSB, leading the replication machinery to continue past the SSB, falling off the DNA and resulting in a single ended DSB (Kuzminov, 2001).

Alternatively, it may be that the global effect of IR on differentiated cells is such that extremely high levels of DSB breaks occur which differentiated cells are simply unequipped to deal with. This may be because the DDR of differentiated cells is relatively inefficient compared to GSCs. Furthermore, it is possible that a difference in DSB pathway preference could influence this. For example, DSBs targeted by NHEJ, which do not require extended resection of 3' strands may be easier to detect via BLISS or INDUCE-seq than DSBs that are undergoing extended 3' resection via HR where detection of these locations may be more difficult.

### 7.4.5 Summary of conclusions

- Immunofluorescence imaging using IRIF "DSB markers" 53BP1 and γH2AX show an increase in DSB repair foci following 10 Gy IR at 24 hours in R10

GSCs and differentiated cells and an increase in DSBs in E2 GSCs cells at 6 hours.

- BLISS and INDUCE-seq detected a trending increase in DSBs in R10 differentiated cells at 24 hours post IR.

- BLISS and INDUCE-seq detected a trending decrease in DSBs in R10 GSCs at 24 hours post IR

# Chapter 8 Discussion

## 8.1 Introduction

GBM remains a formidable oncological challenge, with universal recurrence despite aggressive combination chemo-radiotherapy following maximal debulking surgery (Stupp et al., 2005). Median overall survival remains between 12 to 18 months and is considerably worse in older populations. Few advances have been made in recent years, likely in part due to the significant heterogeneity found within these tumours. GSCs have become increasingly better characterised and established as the roots of recurrent disease surviving doses of up to 60 Gy IR (Lathia et al., 2015). GSCs demonstrate aberrant upregulated DDR pathways, in particular HR, contributing to radioresistance (Lim et al., 2012). Interestingly, GSCs have been demonstrated to have elevated RS which has been postulated as a primer to elevated DDR (Carruthers et al., 2018). The challenge of investigating RS is the double-edged sword of DNA damage. Whilst RS has been posed as a primer for key DDR pathways, it has also been observed that elevated RS levels have been associated with elevated levels of DSBs (Petermann et al., 2010, Arnaudeau et al., 2001). Given the elevated RS observed in GSCs, it brings into question whether GSCs also harbour more DSBs. These DSBs are the most deleterious type of DNA damage to healthy cells which cause mutation and genomic instability. Therefore, understanding how DSBs might contribute to GSC treatment resistance is of significant interest.

This thesis looked to use 3 primary GSC lines and their matched differentiated progeny to establish patterns of DSB damage *in vitro* to better define the relationship of DSBs and location with GBM cell survival. These lines have the advantage of having been well characterised by previous work from Dr Emily Clough and Dr Ross Carruthers and demonstrate radioresistance of GSCs compared to their differentiated progeny (referenced as GSC-bulk tumour cells in previous works). With this in mind, cell line radiosensitivity, subtypes and GSC markers were not revisited as part of this work. Firstly, the endogenous pattern of DSBs was described across the three GSC's and matched differentiated progeny with reference to endogenous DSB patterns across chromosomes as well as identifying regions of high DSB density. Patterns of DSBs across genes and gene bodies were mapped with particular interest across TSS and TTS locations.

DSB density was then described with reference to transcription and gene accessibility using previously established RNA-seq data and new ATAC-seq libraries. Finally, the influence of IR on DSB density and location was investigated in two cell lines at late and mid-to-late repair timepoints.

This chapter will summarise and discuss the overall findings from throughout this thesis and identify key features of interest. The questions and challenges raised by the findings in this thesis will also be discussed with reference to potential future investigations.

## 8.2 Characterising DSBs in GBM

Exploring the DSB landscape across GSCs was approached with the aim of identifying sites of frequent DSB locations and unifying factors within these. This was explored in the context of GSC lines as well as other publicly available data which included neural cell data and two commercial cancer cell lines K562 and MCF7. This was to provide relevant comparators to GSC data which would be investigated in the context of matched GBM differentiated cells and DSB data following exposure to IR.

### 8.2.1 Main findings and discussion

A significant consideration when embarking on this analysis was whether endogenous DSB location in GSCs was purely stochastic or even uniform across different cell lines and cell types. Our data indicated that DSB distribution in GBM cells maintained uniform DSB density across repeats, indicating consistency within cell lines. Conversely, DSB distribution across the GSC cell lines E2, G7 and R10 displayed some differences, though there were also notable shared regions of high DSB density. Comparatively, across neural cells and commercial cancer cell lines, DSB density across these cell types appeared to show unique DSB density distribution compared to GSCs, most particularly in the commercial cancer cells. Interestingly, GSCs shared a peak of DSBs in chromosome 11 with the commercial cancer cell lines. Regarding euchromatin profiles, this was variable across the three cell lines, though there were some chromosomes which appeared to have lower euchromatin enrichment compared to other chromosomes and also compared to other GSCs.

Whilst GSC cell lines E2, G7 and R10 demonstrated a broad variability in overall DSB distribution, there remained a number of high DSB density sites that were shared across these three lines. Indeed, each line shared at least half of the 100 highest DSB density 50 kbp sites with another GSC line and there were 26 sites shared across the three lines. Regarding GSCs, each line shared over half of the top 100 sites with their differentiated progeny, with G7 and R10 sharing over three quarters of sites. This was an initial indication of the similarities between GSCs and differentiated cells in their endogenous DSB density. This was perhaps surprising given the important biological differences that had previously been established between GSCs and differentiated cells having raised the question of whether endogenous DSB density had a bearing on DDR or was affected by RS (Carruthers et al., 2018).

### 8.2.2 Challenges and limitations

This initial investigation at a genomic level provided helpful context and direction to indicate locations of interest. Important considerations of this genome-wide view must be the limitations of the BLISS mapping and NGS mapping techniques in general. Given the complexity of the BLISS protocol, additional steps to control for cell cycle can be challenging to add and were not undertaken as part of this study though this may have been helpful in controlling for cell cycle-dependant DSBs. Additionally, this method relies on PCR library preparation which may influence DNA frequency, though the use of UMIs to remove PCR repeats at pre-processing steps partially mitigates this as a confounding factor. With regards to NGS mapping, it is well known that short read sequencing is at a disadvantage in mapping some locations of the genome, hence the need to exclude ‘blacklist’ regions from analysis. In fact, many of these blacklisted regions that exist at telomeres, centromeres and highly repetitive regions are also at risk of DSBs (Qiu, 2015, Iacovoni et al., 2010). DSBs at these highly repetitive regions can be frequent and difficult for cells to repair (Fumagalli et al., 2012). However, short read analysis will not detect these consistently, hence the exclusion of these sites. Whilst these regions would certainly be of interest, alternatives to short read analysis were not used for BLISS reads, however longer read sequencing might provide interesting insight into these often unexplored regions. Another important consideration regarding mapping cancer data to the “normal” human genome is the preponderance of

cancer genomes to exhibit CNV. As previously discussed in chapter 6, WGS data became available in the latter stages of this project. As such, BLISS data was not re-analysed and adjusted for CNV per DSB. WGS analysis was performed by Novogene™ and the data provided was locations of CNV with associated genes involved. The results from copy number adjustment genes with regards to DSBs in transcribed genes were reassuring in demonstrating similar patterns. However, CNV could have been utilised to adjust each DSB for copy number. A limitation of the WGS data was that the CNV regions provided were associated with whole numbers only. Cancer cells will still likely have a heterogenous population which might suggest that these results would not have been fully representative of the CNV landscape in our cells. Additionally, WGS data was only available in GSCs and not in differentiated progeny. Whilst CNV might be highly similar between lines given that differentiated cells are derived from GSCs, it is possible that CNV could diverge following differentiation. This would limit the comparisons between GSCs and differentiated cells when considering CNV. Regarding alignment considerations, BLISS data was aligned to the "normal" human genome rather than a cancer cell line-specific genome which may have implications on results interpretation. However, the *de novo* construction of a bespoke cancer genome is a challenge that is not without potential errors also, hence the pragmatic approach to "normal" genomic alignment which was taken.

## 8.3 Exploring DSBs in genes and gene length

Having identified DSB clusters occurring within gene sites in the previous chapter, DSBs within genes were taken forward for further investigation. Additionally, gene length, which had previously been identified as potentially linked to DSB frequency in neural cells was also investigated (Wei et al., 2016).

### 8.3.1 Main findings and discussion

Across all GBM GSCs and differentiated cells, *MALAT1/TALAM1* sites were identified as the most broken gene location, consistent with the 50 kbp regions with the highest DSB density as described in the previous chapter. Similarly, the genes with the highest DSB density in GSCs had a high concordance with their differentiated counterpart cells, with the majority of the top 10 genes across

GBM lines being shared between GSCs and differentiated cells. Those genes that were not shared within the top 10 genes were within the top 100 genes with the highest DSB density. Interestingly, *MALAT1/TALAM1* in neural cells was also identified as a high DSB density gene being in the top 10 genes for NPC and NEU lines and in the top 100 for NES. As discussed, the RNA structure of MALAT1 has been well described to form complex secondary structures, including G4s (McCown et al., 2019, Wang and Vasquez, 2006). It is known that G4s can interfere with replication fork stability and are thereby potential mediators of RS (Maffia et al., 2020). Whether this RNA structure reflects the potential for similar DNA structure occurrence is uncertain. What is known is that *MALAT1* is often highly expressed in many cancers and has been associated with poor prognosis in GBM as well as in other cancers including colorectal and breast cancer (Cai et al., 2018, Cervena et al., 2022, Wang et al., 2018). Furthermore, *MALAT1* has been demonstrated to have a role promoting HR in prostate cancer; when silenced, HR dysfunction occurs, resulting in PARP sensitivity (Yadav et al., 2023). The high DSB density identified in *MALAT1* may reflect expression in these cells, though the possibility of non-canonical structures existing at this gene remains. The levels of DSB formation at *MALAT1* are truly remarkable in comparison to other areas of the genome and are highly consistent between the different cell lines examined in this thesis.

As previously described, elevated RS in these GSC lines was associated with radioresistance. Therefore investigating “at-risk” sites for RS was of interest, to determine whether these regions harboured greater DSB density than others. Long neural genes have been cited as harbouring DSB clusters due to their increased risk of replication-transcription collisions resulting in RS in neural stem cells (Wei et al., 2016). With this in mind, DSBs were investigated with reference to gene length in GSCs and differentiated cells. When taken as total DSBs per gene, there was a clear positive correlation with gene length, however when DSBs were adjusted for gene length to give DSB density, this positive correlation was lost across the three GBM lines in both GSCs and differentiated cells. Furthermore, this remained the case when genes were divided into quartiles and after adjustment for gene length there was evidence of a lower DSB density in longer genes. Finally, when considered as a single group, whilst long neural genes had a higher number of DSBs per gene compared to all other genes, this

effect was lost when adjusted for gene length. This was indeed unexpected, given the findings of Wei et al (Wei et al., 2016), however on investigation of this study there were several differences that could account for the discrepancy in results. Wei et al identified recurrent DSB clusters by treating neural stem cells with aphidicolin as a means of inducing RS. Our GSCs are known to have elevated levels of RS compared to their differentiated counterparts. It is feasible that, though the difference in RS levels might have biological significance between GSCs and differentiated cells, that it does not confer the same levels of RS caused by aphidicolin treatment. It may also be that aphidicolin has an additive effect that augments DSBs within "at-risk" RS sites. Additionally, other groups have had conflicting results where induction of DSBs secondary to RS only occurred a number of days following RS induction (Michel et al., 2022, Wang et al., 2020). Furthermore, whilst there was not an apparent increase in DSB density in these longer genes, the absolute DSB density in these sites remains high and therefore may continue to retain biological significance at these locations.

As RS is an agreed mechanism of DSB generation, longer genes that are at risk of replication-transcription collisions might be expected to harbour a disproportionate number of DSBs. Regarding long neural genes in GSCs, an increase in DSBs might have been expected, given that previous research demonstrated an increase in DSBs in long neural gene in NSCs, however this was not seen in our populations. Whilst there are thought to be links in lineage of NSCs to GSCs, it is possible that GSCs do not possess the same level of fragility in long neural genes as seen in neural stem cell populations. Long neural genes did not demonstrate significant differences with other genes across GSCs however there was an overall reversal in trend of DSB density where DSB density decreased with increasing gene length. This was certainly interesting to observe given that these locations would still be expected to be at greater risk of replication-transcription collisions. It is not clear why the reverse would be the case but it is possible that long genes in these GSCs either have extended abilities to resolve DSBs at longer genes or that these locations do not undergo the same levels of replication-transcription conflicts at these sites as in neural cells. It might also be important to consider the influence of other factors on DSBs such as transcription. For example, given the clear finding that

transcription has been highly associated with DSBs in these GBM cell lines, it is possible that there is an overall greater transcriptional activity in shorter genes in GSCs than in longer genes which could also drive this pattern.

### 8.3.2 Challenges and limitations

Given the high density of DSBs occurring across *MALAT1* there implies an underlying related effect causing high DSBs. Given this finding across the GBM lines and non-GBM lines, additional data such as paired RNA-seq data in non-GBM lines would have been helpful in determining whether there were additional factors at play such as gene expression. Additionally, analysis of the commercial cancer lines was limited by data availability where there were BLISS datasets with no repeats which limited conclusions from these results. Multiple repeats of BLISS in these and other cancer cell lines could help in further investigating the occurrence of *MALAT1* as a high density DSB site. Furthermore, a better understanding of the physical structure of *MALAT1* is also interesting to consider. The lncRNA MALAT1 has a well described architecture, however determining DSB inducing structures within the DNA remains relatively poorly researched (McCown et al., 2019). Therefore, a direct investigation of the DNA structure of *MALAT1* would have been particularly interesting to do. In particular, exploring evidence of G4s and R-Loops for example using ChIP-PCR to isolate these could potentially have assisted in identifying whether these structures had a roll in the high DSB density levels. Importantly, *MALAT1* was not a named gene in the list of genes with CNV from the WGS data which could have also been a consideration in the particularly high DSB of this gene.

Regarding gene length and long neural genes, identifying whether the DSB clusters seen by Wei et al (Wei et al., 2016) were recapitulated in our GBM lines following treatment with aphidicolin would assist in making a clearer connection between DSB density and long neural genes. Additionally, direct comparison of the endogenous levels of RS in GSC and neural cell lines and also RS levels with aphidicolin treatment could help in understanding the magnitude of difference between these two groups.

## 8.4 Investigating DSBs in gene bodies and annotated genomic regions

To investigate genes in further detail, DSB density across genes and annotated genes sites were studied. Of particular interest were TSS and TTS locations with comparison of DSBs across GBM and non-GBM lines.

### 8.4.1 Main findings and discussion

Interestingly DSBs within gene bodies across the three GBM cell lines were not consistent in distribution, most notably at TSS locations where G7 cells showed an apparent increase in mean DSBs and conversely R10 showed an apparent decrease. It was also evident that TSS and TTS locations showed the greatest changes in mean DSBs across gene bodies. On investigating mean DSBs across non-GBM cell lines, TSS and TTS locations were also the locations of the greatest changes in mean DSBs in both neural cells and commercial cancer cell lines. For DSBs at TSSs, all neural cells and commercial cancer cells demonstrated an overall increase in mean DSBs at the level of gene body overview. However, when TSS regions were taken forward to investigate by absolute distance from TSS, there was a more complex pattern visible, particularly at neural cell TSSs which demonstrated a small decrease in DSB density just immediately around the TSS. The G7 GBM cells and K562 cell line also demonstrated a small decrease in DSB density peri-TSS, however this was less pronounced than in the case of neural cells. The TTS locations on the other hand were consistently similar in all GBM cell lines and showed an increase in DSB density at TTS which appeared more pronounced than in non-GBM cell lines. Interestingly, euchromatin enrichment across GBM lines did not match DSB patterns but rather, consistently showed greatest enrichment at TSS and less enrichment at TTS locations. To investigate this further, gene bodies were compartmentalised into annotated regions to compare expected and actual DSB frequency. This demonstrated that GSCs had a significant increase in DSB frequency at TTS (2.5-fold increase) when compared with neural cells (1.176). GSCs also demonstrated a significantly lower DSB frequency at intergenic regions compared to neural cells and a corresponding significant increase in DSB frequency within exons. This higher DSB frequency at gene end sites was consistent across all GBM lines and whilst neural cells and commercial cells showed some small increases, these were at

much smaller magnitudes (1.1-1.5 fold). This may suggest a particular biological effect occurring within GBM lines such as the action of topoisomerases at these sites. Topoisomerases have been established as key in acting in transcription through the generation of DSBs (Ju et al., 2006). TopIIß has been found to promote resistance to RS-inducing drugs and is upregulated in GSC lines (Kenig et al., 2016). Additionally, the depletion of TopIIß has been shown to sensitise GSCs to chemotherapies such as cisplatin and temozolomide which induce RS. Interestingly, TopIIß has also been associated with neuronal differentiation and chromatin modelling which may suggest a role of TopIIß in promoting stemness (Sano et al., 2008).

### 8.4.2 Challenges and limitations

The two commercial cancer cell lines provided an initial insight into other cancer genomes and indicated an increase in DSBs at TTS. In order to investigate cancers as a whole in greater depth, additional cancer datasets would be supportive in better understanding this finding. Focussing on cell lines that have demonstrated sensitivity to topoisomerase inhibition such as paediatric neuroblastoma or in cancer cells with topoisomerase mutations conferring topoisomerase resistance could better describe the impact of topoisomerase activity on TTS DSBs in cancers overall (Pan et al., 2021, Errington et al., 1999). Furthermore, given that topoisomerase inhibition can induce senescence, identifying the breakome in senescent subgroups with and without topoisomerase inhibition would help to determine whether DSBs at these locations are present in senescence and whether senescent cells are resistant to increases in DSBs at these locations (Taschner-Mandl et al., 2016). This could give a better indication as to whether higher DSB density at TTS in GBM was shared with other cancer cell types and therefore give some indication of whether this pattern might indicate treatment sensitivity. Upregulation of TopIIß has been seen in prostate cancer and has demonstrated some sensitivity to etoposide, a TopII inhibitor, therefore profiling the breakome with particular interest at TTS could be extremely attractive in such cancer cell types (Haffner et al., 2010). Mapping the breakome of cells with TopII mutations might also provide an interesting insight as to whether loss of functional TopII impacts DSB distribution at TTS regions and indeed across genes. For example, mapping the breakome across TTS regions in human cell lines exhibiting heritable B-cell

immunodeficiency secondary to TopIIß mutations could give insight into TopII mutated TTS DSB distribution (Broderick et al., 2019). Additionally, further quantification of the mutational state and expression of topoisomerases in these cell lines could help explain the increase in DSB density at TTS locations via western blot or RNA-seq data, compared to neural cells and other cell types. Finally, it would be of significant interest to identify what changes in DSB density occurred at TTS following inhibition of TopIIß. This, paired with cell survival data in GBM and non-GBM lines, would be very helpful in establishing to what extent topoisomerase activity impacts on cell survival in our GSC lines.

## 8.5 Gene transcription, euchromatin enrichment and differential DSBs across GBM

Having identified genes as important sites of DSBs, it followed that investigation of gene expression and accessibility would be important to provide insight into the factors which influence gene DSB density. Indeed, transcriptional activity has previously been associated with DSBs in other cell types (Brambilla et al., 2020, Michel et al., 2022). Transcriptional activity has also been highlighted as very important, particularly in the context of stem cells where hypertranscription appears important in early development (Percharde et al., 2017).

### 8.5.1 Main findings and discussion

Our findings in this study of transcription and DSB density in genes demonstrated an increase in DSB density in the top quintile of transcribed genes. This was seen across the three GSC and differentiated cell populations. Interestingly, this was not a clearly linear trend and DSB density appeared relatively less affected by transcriptional activity in the first to third quintiles. Transcription requires planned access to gene sites in order for RNA polymerase to access template strands for RNA sequencing. Whilst this activity per se does not necessarily cause physiological DSBs, the DNA during these processes is in a more vulnerable state than in other unexposed locations. In the study by Michel et al. (Michel et al., 2022) which looked at neural progenitor cells, they identified that transcription-associated DSBs were able to activate p53 which suggests that the generation of

these DSBs occurring within healthy cells is under close surveillance to maintain genome integrity.

The investigation of euchromatin in these cells also posed an area of interest given the links with transcription and gene access but also with chromatin structure being identified as important in response to DNA DSBs (Harrod et al., 2020, Brambilla et al., 2020, Bao and Shen, 2007). Our data identified that the most enriched euchromatin peaks had a higher DSB density compared to the least enriched peaks. This, however was not universal to every peak and in both regions for high and low euchromatin enrichment there were peaks with very few or no DSBs detected.

An important initial question in this thesis was whether there were meaningful differences in DSB distribution and density in GSCs compared to differentiated cells given the key differences in RS levels and radiosensitivity (Carruthers et al., 2018). DSB mapping provided a potential avenue to investigate a possible contributor to the elevated RS levels seen in the GSC populations. Endogenous DSB frequency was higher in GSCs compared to differentiated cell populations, which could potentially provide a link to elevated RS levels in these GSC lines (Tsegay et al., 2019). This was broadly informative as part of our investigations but did not provide a means of isolating specific RS sites for analysis.

RS-specific sites were investigated indirectly via long genes and highly transcribed sites. Differential DSB density was also investigated across GSCs and differentiated cells to determine if there were key changes in DSB distribution across the respective populations. At a wider 100 kbp region level, there were few differences that were identifiable between GSCs and differentiated cells and these differential 100 kbp regions still showed that DSB patterns at these sites were similar with most DSBs occurring in genes. At a gene level there was greater divergence in DSB density in GSCs and differentiated cells. Gene Ontology ORA analysis did not identify unifying themes across differential DSBs in genes, though gene function and DSB location would not necessarily be expected to be linked. Across G7 and R10 GSCs there were some shared genes with significantly lower DSB density, though, again, there were not clear unifying themes across these genes. Overall, it appeared that whilst endogenous DSB density in GSCs was higher than in differentiated cells, differential DSB density

did not vary meaningfully between the populations. This was suggestive that whilst DSB overall burden might be important, location of DSBs at a gene level may not be a driving factor in the changes between GSC and differentiated cell radiosensitivity. Rather, whilst the relative contributions of the mechanisms forming DSBs may be the same across GSCs and differentiated cells, it is the magnitude of DSB frequency that drives the changes between these two populations.

### 8.5.2 Challenges and limitations

Given this indication of high DSB density in highly transcribed genes it would be interesting to identify how inhibition of transcription impacted DSB density and the DDR in these cell lines as a means of understanding better how GSCs respond to these transcription-associated DSBs. This would be particularly pertinent in GSCs, where elevated RS has been associated with radioresistance (Carruthers et al., 2018). Since replication-transcription conflicts are a key cause of RS, the measurement of DSB density of transcription-inhibited cells and relevant RS levels could identify what role DSB density might have in activation of the DDR pathways potentially supporting GSC radioresistance.

Regarding the investigation of euchromatin enrichment, this analysis used broadPeak files in this instance as a means of identifying larger areas of overall DSB density. However this does mean that conclusions at a nucleosome level cannot be drawn as this would have benefitted from analysis using narrowPeak files in the first instance. Equally, another consideration in the analysis of ATAC-seq data is that ATAC-seq data purely identifies regions of accessible and mappable chromatin which means that heterochromatin cannot be studied directly. For this, additional investigations such as with heterochromatin ChIP-seq markers would have been beneficial and might have allowed for a more direct comparison of heterochromatin and euchromatin. However, it would be important to bear in mind that this would require cross-technique comparison. This will still not fully describe the landscape of heterochromatin and DSB density in these sites however, as there are a number of densely packed heterochromatin regions in highly repetitive blacklisted sites that are difficult to map without the use of long read sequencing. Finally, investigation of euchromatin alongside transcription becomes complicated given that highly

transcribed genes will naturally also intersect with areas of euchromatin enrichment. Some means by which to alleviate this could be to identify peaks not intersecting genes and thereby eliminating transcriptionally active areas from analysis. This could also be done in reverse by looking at genes that do not intersect with euchromatin peaks. However, this could be problematic, given that DSB density across intergenic sites is low overall and there may well be few to no DSBs to identify at many sites.

As mentioned, RS sites were not directly identified in this thesis but rather locations that have been cited as at-risk were used as indirect locations for investigation. There are however some techniques that could assist with this which were considered, though were not within the scope of this thesis. One means of identifying RS-specific sites would be in identifying known sites that pose an increased risk of RS such as G4s and R-Loops. ChIP-seq has been performed in the human genome in order to identify G4 sites, of which this data is now publicly available (Chambers et al., 2015). However, given that cancer genomes are uniquely constructed, this may not be fully representative of G4 locations within these cell lines. Regarding R-Loops, DRIPseq has been used as a means via the S9.6 antibody against R-Loops via fragmented chromatin, though there has been some debate about the specificity of the S9.6 antibody (Ginno et al., 2012, Smolka et al., 2020). Additionally, “R-ChIP” has also been described as a method for R-Loop mapping where the catalytically inactive ribonuclease H1 (RNASEH1) is used to identify the loci of interest (Chen et al., 2019a). Alternatively, another method of identifying “at risk” sites for RS would be to consider the mapping of replication forks alongside transcriptional activity using a method such as DNAscent to provide replication fork locations (Boemo, 2021).

## 8.6 Mapping DSBs before and after irradiation

Having explored differences in endogenous DSB density and frequency between GSCs and differentiated cells, the DSB density and frequency was investigated following irradiation. The IRIF γH2AX and 53BP1 were used to describe the DDR to DSBs and give context to BLISS and INDUCE-seq-detected DSBs following IR.

### 8.6.1 Main findings and discussion

As a background to this, IRIF "DSB markers" for the relevant cell lines were used to describe the DDR response following 10 Gy IR. R10 GSCs and differentiated cells demonstrated a significant increase in 53BP1 foci at 24 hours. E2 differentiated cells demonstrated a significant increase in 53BP1 foci at 6 hours post IR. Regarding γH2AX foci, R10 differentiated cells had a significant increase in foci at 24 hours but this was not seen in GSCs, however E2 differentiated cells and GSCs both showed a significant increase in foci at 6 hours following 10 Gy.

The BLISS assay showed an intriguingly different pattern. These data demonstrated that R10 differentiated cells showed an increase in DSBs compared to their IR-treated counterparts at 10 Gy 24 hours, however R10 GSCs showed a decrease in BLISS-detected DSB reads following IR in 2 of the three repeats. Furthermore, when BLISS was performed in E2 GSCs following IR at 6 hours there was a significant decrease in BLISS-detected DSB reads following IR. Interestingly, whilst this was a significant difference between treated and untreated GSCs, there were no significant differences in DSB distribution between 0 Gy and 10 Gy cells across 100 kbp sites or genes suggesting that whilst overall DSB burden was higher in differentiated cells post IR, the pattern of DSBs remained constant.

The trend in divergent BLISS-detected DSBs across GSCs and differentiated cells post-IR raised the question of whether these data could be replicated utilising alternative DSB genomic mapping techniques. Because BLISS precludes drawing conclusions about absolute DSB quantity, this was a limitation in using the BLISS technique. Therefore INDUCE-seq was employed which allowed a method of DSB mapping that could directly measure DSB frequency (Dobbs et al., 2022). Interestingly, a similar pattern was seen across R10 differentiated cells where there was an increase in INDUCE-seq-detected DSBs following IR. Furthermore, in GSC cells, there was also an overall decrease in INDUCE-seq-detected DSBs following IR, consistent with prior BLISS results.

Altogether this data is difficult to understand and interpret in the context of current canonical radiobiology given the differences in IRIF and BLISS-detected DSBs alongside the similar results of the INDUCE-seq-detected DSBs. The

consistency between BLISS and INDUCE-seq results implied a true biological effect resulting in divergent DSB frequency patterns post-IR in R10 GSCs and differentiated cells.

The decrease in R10 and E2 GSC DSBs post-IR coupled with the lack of discernible changes in DSB distribution implied that these changes might represent global processes affected by IR exposure in the cell rather than specifically IR-induced DSBs. The total number of IR-induced DSBs per cell per 1 Gy has historically been cited as 40 DSBs with >95% of these being repaired at 24 hours (Rothkamm and Löbrich, 2003). In the absence of IR, it has been estimated that a single cell cycle will produce around 50 endogenous DSBs (Vilenchik and Knudson, 2003). Given the high number of DSBs detected through BLISS at endogenous levels, the 10 Gy IR data may be saturated with endogenous DSBs, resulting in the signal of any IR-induced DSBs being significantly diminished. However, this does not address the apparent overall decrease in BLISS and INDUCE-seq-detected DSBs following IR. This is somewhat challenging to understand, however there are some potential indications of what processes could result in this overall reduction seen in BLISS and INDUCE-seq-detected DSBs in GSCs. Regarding GSCs, it is known that cell cycle checkpoint proteins including Chk1 and ATR are upregulated in these cells (Carruthers et al., 2018). This tighter control of the cell cycle might allow for a timely response to IR with activation of cell cycle checkpoints. This could potentially result in an overall pause in the activity of the cell and so any DSBs that might be endogenously generated through normal cell activities could be reduced. Additionally, it has been seen in nasopharyngeal cancer that IR can induce the pausing of transcription which can result in a radioresistant phenotype (Liu et al., 2023). Given that there appear to be a number of DSBs in GSCs associated with transcriptional activity, transcriptional pausing could explain the overall reduction in DSBs following IR and may not necessarily demonstrate a change in DSB distribution. This would also correspond to the radioresistant GSC phenotype. However, it is also possible that this reduction in DSBs following IR exposure is related to the upregulation of the DDR seen in GSCs (Bao et al., 2006). This upregulated DDR could result in efficient repair of IR-induced DSBs and could even result in an increase in the repair of endogenous DSBs as well. It may even be a combination of both transcriptional downregulation and efficient

DDR together to give this overall decrease in BLISS and INDUCE-seq-detected DSBs in GSCs. These intriguing data in mapped DSB frequency post IR in GSCs merits further investigation, however this was beyond the scope of this thesis.

In contrast to GSCs, differentiated cells demonstrated an increase in DSBs following IR. Again, there were relatively few changes in DSB distribution following IR in differentiated cells. This was potentially suggestive that it was not necessarily IR-induced DSBs that were being detected following 10 Gy treatment. Additionally it could also indicate that IR-induced breaks are random in their location on the genome. However differentiated cells indicated a divergent response to IR than the response initiated by GSCs seen in BLISS and INDUCE-seq-detected DSBs. Again, these data in the context of both IRIF and GSC data is somewhat complex to interpret. The reasons for the increase in DSBs in differentiated cells could include the following. Firstly, it is known that differentiated cells have a lower overall expression of cell cycle checkpoint proteins and therefore may be at a disadvantage to GSCs in responding to IR and subsequent damage. Where GSCs might be pausing the cell cycle to promote better repair and pause transcription, differentiated cells might instead be continuing through checkpoints, thereby accumulating further DSBs through attempted division. By continuing through the cell cycle, it is also possible that non-lethal SSBs might become converted to DSBs through replication run-off, where replication machinery attempts to traverse an unrepaired SSB and uncouples from the DNA, resulting in a single-ended DSB (Kuzminov, 2001). It is possible that differentiated cells do not experience a downregulation of transcriptional activity, but rather may continue at the same or at an accelerated rate. This, coupled with broad levels of DNA damage from SSBs and DSBs could result in replication-transcription conflicts leading to collapse and further DNA damage accumulation. Finally, it is also possible that this broad increase in DSBs represents the differences in the DDR between GSCs and differentiated cells. As described, the majority of DSBs would be expected to be successfully repaired within hours of DSB induction. However, it may be that, given the relative poorer DDR seen in differentiated cells, compared to GSCs that these cells are unable to repair DSBs in a timely manner, resulting in the divergence of DSB patterns between GSC and differentiated populations following IR.

Given these findings there appear to be a number of very interesting avenues for investigation. The potentially divergent activity of transcription following IR in GSCs and differentiated cells might provide an interesting area of study. For example, it would be interesting to measure the total RNA in GSCs and differentiated cells before and after IR to determine whether there is a definitive change that would correlate with the BLISS and INDUCE-seq data. Additionally, transcriptional inhibition in GSCs and differentiated cells in conjunction with cell survival and DSB induction might also identify whether differences in transcription activity are driving part of the radioresistant phenotype seen in GSCs. Conversely, inhibition of transcription in differentiated cells following IR could also determine whether transcriptional inhibition promotes radioresistance in differentiated cells.

### 8.6.2 Challenges and limitations

When considering DSBs detected following IR, it is important to bear in mind that, given the 6 hour and 24 hour timepoints, these timings may not allow for definitive isolation of persisting IR-induced DSBs. This may be because the percentage of IR-induced DSBs could be relatively small and significantly diluted by the large quantity of endogenous DSBs seen both before and after IR. It is also important to consider that both BLISS and INDUCE-seq have a “bulk” sequencing approach where the breakome is built from DSBs occurring in millions of cells. Therefore, it is more feasible to identify broad changes in DSBs occurring due to the cell responses to IR rather than identifying individual IR-induced DSBs. This poses the challenge that IR-induced DSBs remain difficult to confirm and the locations remain unclear.

Challenges to this are that techniques using antibodies to DSBs are not true markers of DSBs but rather markers to the DSB response. Equally “DSB markers” such as γH2AX are impractical at detecting individual DSBs at a meaningful location level given that γH2AX are known to span up to megabases in length (Noubissi et al., 2021a). Because of this, the location of true IR-induced DSBs remains elusive. IRIF have primarily been used as indicators of DSB induction following IR or damage-inducing treatments. There is little in the way of direct

information therefore on measuring DSBs endogenously using these methods. This poses the challenge of what exactly IRIF measure at baseline. It is known that both 53BP1 and γH2AX foci are made up of "nano"-foci (Hausmann et al., 2018). These are only individually detectable at super-resolution microscopy. Therefore, it is feasible that these foci, whilst not detectable using standard confocal microscopy and IF, could be visible at higher resolution. However, in this case it does imply a far greater response to IR-induced DSBs compared to endogenously occurring ones. These IRIFs also show DSBs in differing stages of repair and these differing stages are not identifiable using IRIFs alone. Furthermore, physiologically-induced DSBs such as those formed by topoisomerases to relieve torsional stress may not signal a response from 53BP1 or γH2AX. Topoisomerase-related DSBs do not appear to associate with γH2AX unless they are inhibited by treatments such as the topoisomerase inhibitor etoposide (Huang et al., 2003, Sunter et al., 2010). It is at this point when topoisomerase is inhibited that there appears to be an accumulation of γH2AX (Berger, 1998, Berger and Wang, 1996). Following on from this, DSBs that have been unable to signal the appropriate DDR may also not be identified using IRIFs (Schipler and Iliakis, 2013). Therefore, a subset of DSBs may not be accounted for in IRIFs; a subset which may harbour greater risk to the cell given the lack of signalling for appropriate repair processes.

Cell cycle checkpoints in IR would be an interesting area for further investigation, given that GSCs have been shown to promote cell cycle arrest in response to IR-induced damage (Bao et al., 2006, Tachon et al., 2018). Identifying DSB frequency following IR in other radiosensitive and radioresistant cancer cell lines could assist in identifying whether the increase in DSBs post-IR in differentiated cells is unique or whether this might also be an indicator of radiosensitivity. DSB detection in GSCs in the context of checkpoint inhibitors may also provide an insight into the underlying cause of the global reduction of DSBs following IR, particularly in GSCs. Similarly, looking at overall transcriptional activity in differentiated and GSC lines together following IR might also identify to what extent transcriptional activity is abrogated by IR in these cells.

The INDUCE-seq data provided a helpful contribution to assessment of total DSBs, though the challenges in plating adherence between GSCs and

differentiated cells limited the conclusions that could be taken from the two cell populations. Additionally, further repeats in these lines and at different IR timepoints would assist in further investigation of the difference in INDUCE-seq-DSB yield following IR. Furthermore, use of other radioresistant and radiosensitive cell lines with appropriate controls or consideration of cell cycle may expand the understanding of the overall role of DSB burden in radioresistance and what role DSB density plays.

A final significant consideration in the context of mapping DSBs using techniques such as BLISS and INDUCE-seq is that both of these methods, as with most DSB mapping techniques, require the blunting of DSB ends. The challenge here is whether IR-induced DSBs are easily identified for blunting steps. It is generally established that IR-induced DSB will often require the HR pathways to repair damage. If IR-induced DSBs are particularly complex they may require very extensive resection to give long 3' overhangs even up to 10 kbp in length (Chung et al., 2010). It is unclear whether these very uneven DNA DSBs could be successfully detected and blunted by the blunting enzymes used (Canela et al., 2016). Additionally, even successful blunting of these extended sites will not truly capture the nature of these DSBs. These asymmetric DSB locations will always be reported as a single base pair location, rather than accurately describing what may be a far more complex DSB. However, this reflects the overall current limitations of DSB sequencing as a whole and is not unique to either BLISS or INDUCE-seq.

## 8.7 Final Conclusions

Having explored the landscape of DSBs in GSCs, there were several key findings that prove interesting and helpful for directing further study. Interestingly, whilst GSCs demonstrate differences between cell lines, reflecting the heterogeneity of GBM, DSB density between GSCs and differentiated progeny appeared globally similar. Transcriptional activity appeared to be an important driver of endogenous DSB activity and interestingly these GBM lines also had an increase in DSBs at TTS locations which may warrant further investigation in the context of topoisomerases and abrogating torsional stress in transcription. Finally, whilst there were not large changes in DSB distribution following IR, the global decrease in DSBs in GSCs suggested a responsiveness in GSCs that was not

matched in differentiated lines. This leaves a number of interesting avenues for further investigation including how transcriptional activity and torsional supercoiling might be interacting in GBM cell survival. Understanding late persistent IR-induced DSB remains an area of interest and potential speculation which at this time is not yet feasible to address. However, if addressed this might give a better understanding of where and why true late IR-induced DSBs occur and their implications in radioresistant cells.

The management and treatment of GBM remains a persistent challenge given the highly aggressive and heterogenous nature of the disease. Much work has been done to better characterise GBM as a cancer in an effort to better target these highly resistant cells. This thesis has described the distribution of DSBs in GSCs in the context of neural cells, commercial cancer cell lines and differentiated progeny GBM cells. This work contributes to our understanding of the DSB landscape in GBM as a whole and highlights the potential importance of transcriptional activity on DSB distribution and frequency. Whilst the differences between GSCs and differentiated cells seen in RS levels and in radiosensitivity do not appear to be linked directly to DSB distribution, this provides a backdrop for moving forward in other avenues that could prove relevant for future GBM research.

# Appendices

## Supplementary Figures and Tables

### Chapter 3 supplementary figures and table

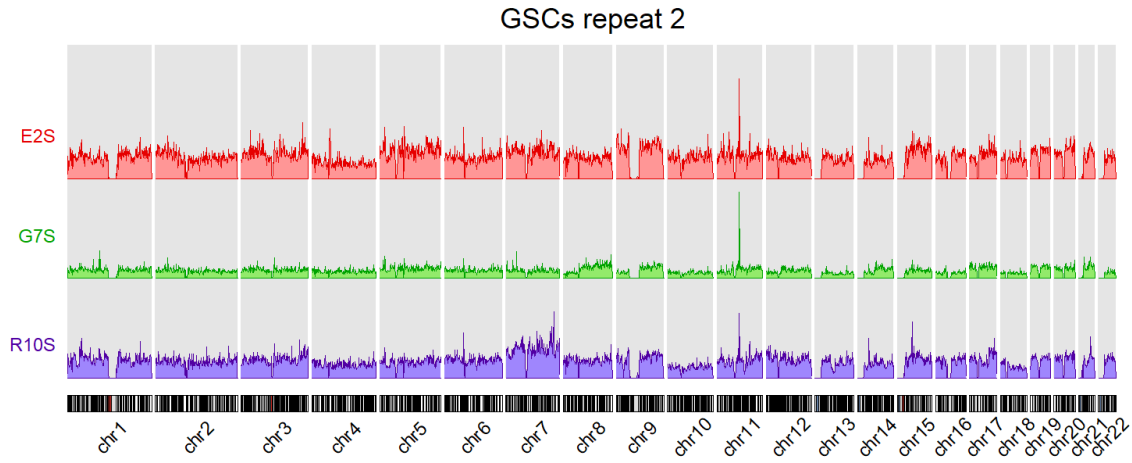

**Supplemental: repeat 2 for GSC lines E2, G7 and R10 DSBs**

Density plot of DSBs across repeat 2 for 3 GSC cell lines E2, G7 and R10. DSB density represented across chromosomes by relative chromosome length using bigWig files for DSB density. Lines from top to bottom: E2 GSC repeat 2 (red), G7 GSC repeat 2 (green), R10 GSC

repeat 2 (purple). Karyoplots are displayed below DSB density. Red markers on karyoplots represent centromeres. Chromosomes 1 to 22 displayed.

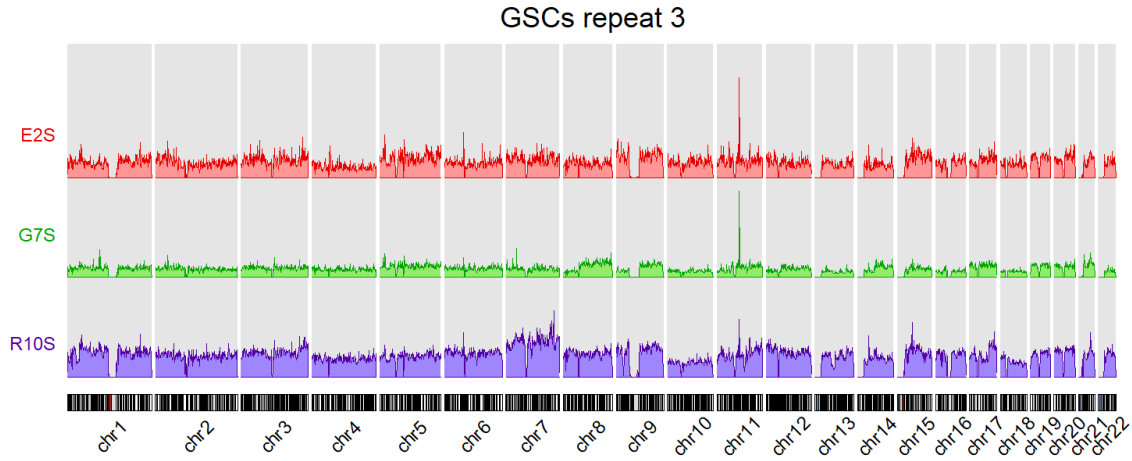

**Supplemental: repeat 3 for GSC lines G7 and R10 DSBs**

Density plot of DSBs across repeat 3 for 3 GSC cell lines G7 and R10. DSB density represented across chromosomes by relative chromosome length using bigWig files for DSB density. Lines from top to bottom: G7 GSC repeat 3 (green), R10 GSC repeat 3 (purple). Karyoplots are displayed below DSB density. Red markers on karyoplots represent centromeres. Chromosomes 1 to 22 displayed.




**Supplemental: repeat 2 for GSC lines E2, G7 and R10 DSBs against chromosomes**

Density plot of DSBs across chromosomes 1-22 in numerical order. Chromosomes from repeat 1 in order of left to right. From top to bottom: E2 GSC repeat 2 (red), G7 GSC repeat 2 (green), R10 GSC repeat 2 (purple). Karyoplot of chromosomes 1-22 displayed below.

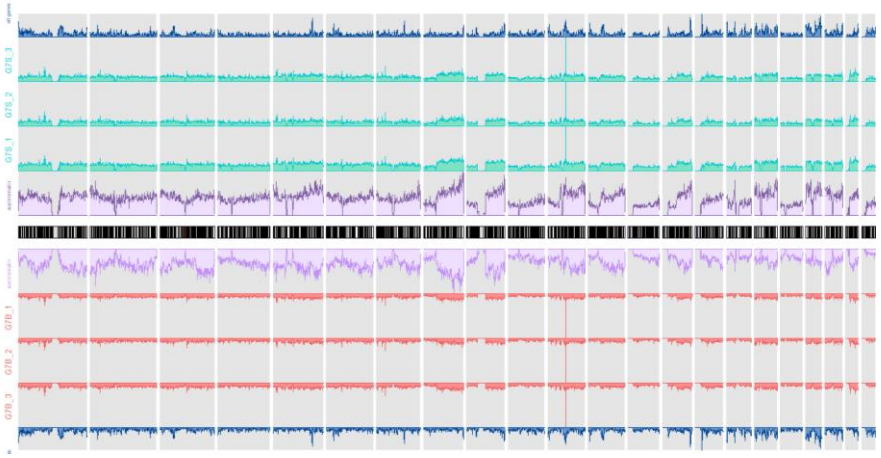



**Supplemental: repeat 3 for GSC lines G7 and R10 DSBs against chromosomes**

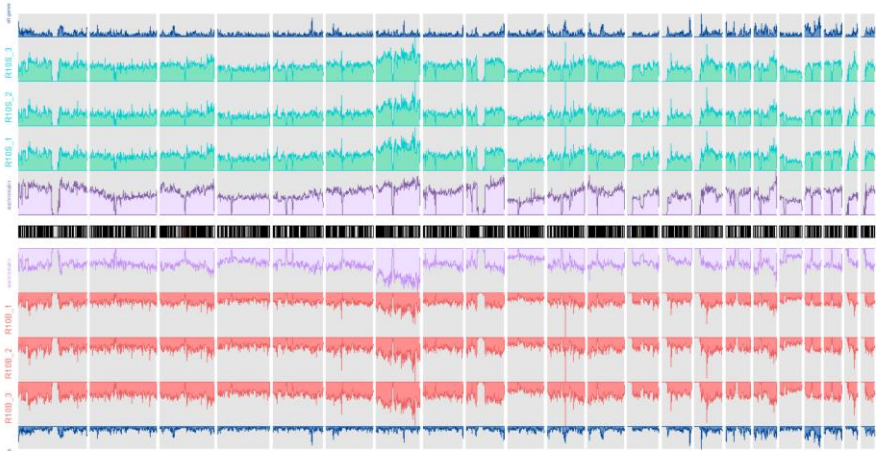
Density plot of DSBs across chromosomes 1-22 in numerical order. Chromosomes from repeat 1 in order of left to right. From top to bottom: G7 GSC repeat 3 (green), R10 GSC repeat 3 (purple). Karyoplot of chromosomes 1-22 displayed below.

(a)

(b)



(c)



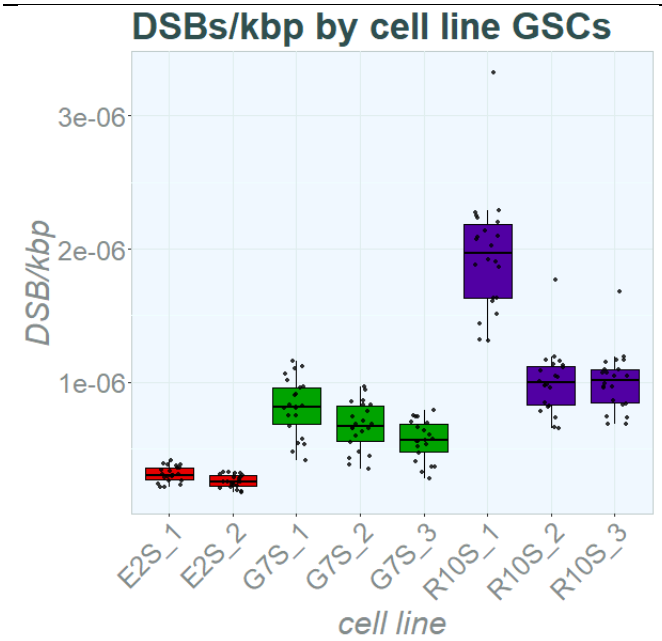
**Supplemental: DSB densities euchromatin enrichment across chromosomes 1-22 in GSCs E2, G7 and R10 GSCs and differentiated cells.**

DSB density plots of GSCs in triplicate and euchromatin enrichment across chromosomes 1-22. Chromosomes displayed below as karyoplots from chr1-chr22. Euchromatin profiles displayed immediately above and below karyoplots in lilac. Individual GSC repeats of DSB density displayed above euchromatin enrichment profiles. Gene density profiles displayed on top line in blue for reference. (a) E2 GSCs repeats and differentiated repeats 1-2. (b) G7 GSCs repeats and differentiated repeats 1-3. (c) R10 GSCs repeats and differentiated repeats 1-3.

**Supplemental Table: DSB density DSBs/kbp**

| | Median DSB density chromosomes | Interquartile range | | Range | |
|---|---|---|---|---|---|
| **E2 GSC rep 1** | $3.06e^{-7}$ | $2.70e^{-7}$ | $3.59e^{-7}$ | $2.15e^{-7}$ | $4.17e^{-7}$ |
| **E2 GSC rep 2** | $2.55e^{-7}$ | $2.24e^{-7}$ | $2.99e^{-7}$ | $1.72e^{-7}$ | $3.27e^{-7}$ |
| **G7 GSC rep 1** | $8.17e^{-7}$ | $6.91e^{-7}$ | $9.63e^{-7}$ | $4.16e^{-7}$ | $1.16e^{-6}$ |
| **G7 GSC rep 2** | $6.71e^{-7}$ | $5.62e^{-7}$ | $8.20e^{-7}$ | $3.48e^{-7}$ | $9.67e^{-7}$ |
| **G7 GSC rep 3** | $5.65e^{-7}$ | $4.80e^{-7}$ | $6.86e^{-7}$ | $2.82e^{-7}$ | $7.89e^{-7}$ |
| **R10 GSC rep 1** | $1.97e^{-6}$ | $1.63e^{-6}$ | $2.19e^{-6}$ | $1.31e^{-6}$ | $3.32e^{-6}$ |
| **R10 GSC rep 2** | $9.95e^{-7}$ | $8.28e^{-7}$ | $1.12e^{-6}$ | $6.56e^{-7}$ | $1.77e^{-6}$ |
| **R10 GSC rep 3** | $1.02e^{-6}$ | $8.47e^{-7}$ | $1.10e^{-6}$ | $6.86e^{-7}$ | $1.68e^{-6}$ |
| **K562** | $6.16e^{-7}$ | $5.59e^{-7}$ | $6.40e^{-7}$ | $3.61e^{-7}$ | $7.69e^{-7}$ |
| **MCF7** | $4.66e^{-6}$ | $4.11e^{-6}$ | $5.28e^{-6}$ | $2.68e^{-6}$ | $1.10e^{-5}$ |
| **NES rep 1** | $3.49e^{-6}$ | $3.06e^{-6}$ | $3.67e^{-6}$ | $2.47e^{-6}$ | $3.86e^{-6}$ |
| **NES rep 2** | $3.68e^{-6}$ | $3.21e^{-6}$ | $3.83e^{-6}$ | $2.60e^{-6}$ | $4.00e^{-6}$ |
| **NPC rep 1** | $1.87e^{-6}$ | $1.66e^{-6}$ | $1.98e^{-6}$ | $1.23e^{-6}$ | $2.03e^{-6}$ |
| **NPC rep 2** | $3.56e^{-6}$ | $3.11e^{-6}$ | $3.76e^{-6}$ | $2.26e^{-6}$ | $3.84e^{-6}$ |
| **NEU rep 1** | $7.41e^{-6}$ | $6.57e^{-6}$ | $7.84e^{-6}$ | $4.86e^{-6}$ | $8.12e^{-6}$ |
| **NEU rep 2** | $6.93e^{-6}$ | $6.03e^{-6}$ | $7.42e^{-6}$ | $4.45e^{-6}$ | $7.68e^{-6}$ |

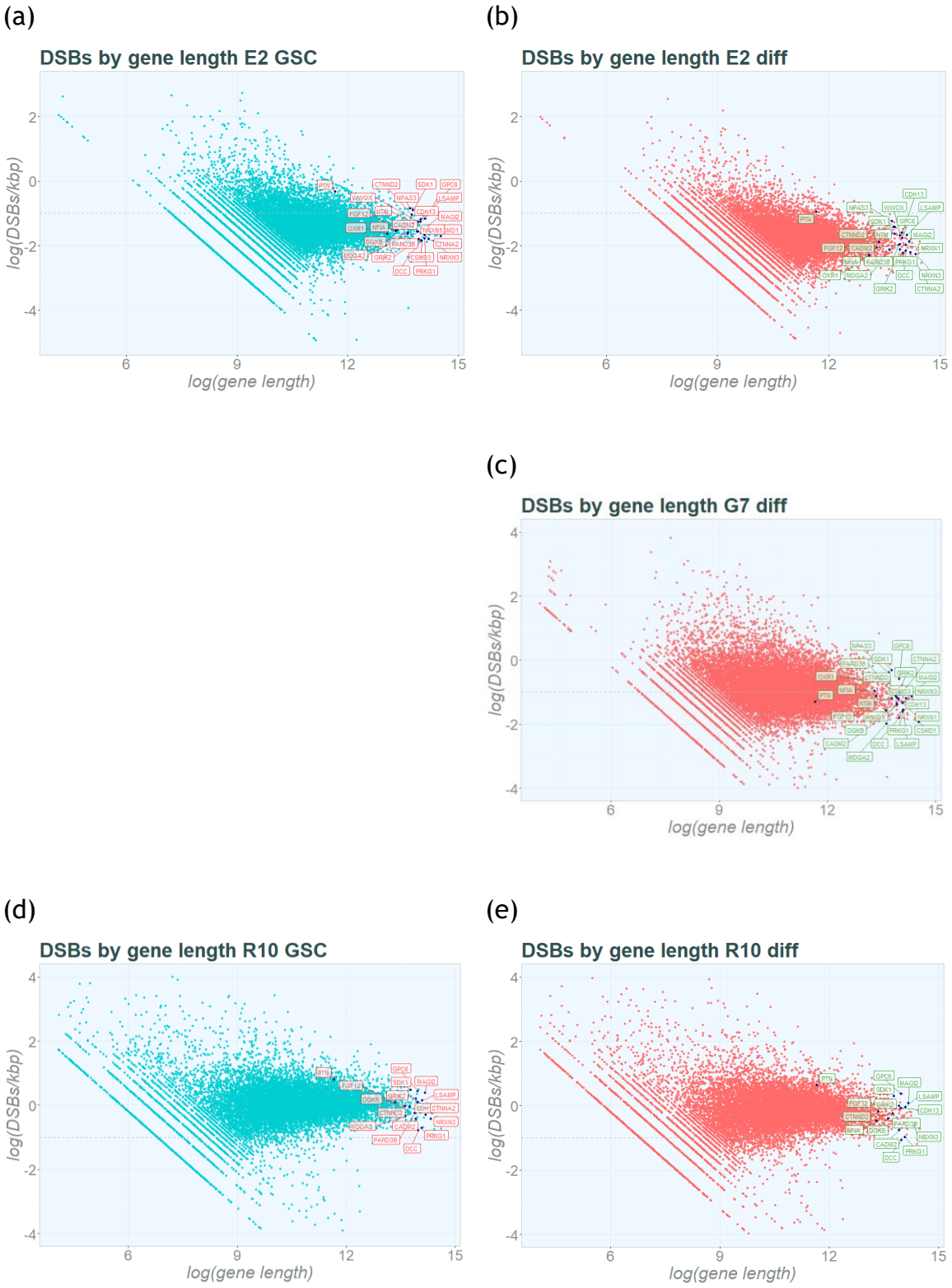
DSB density in DSBs/kbp across GSC, commercial cancer cell lines and neural cells lines calculated across chromosomes 1 to 22. Median fold change, range and standard deviation across chromosomes 1 to 22 per cell line displayed.




**Supplemental: DSB density fold change across chromosomes by cell line**

DSB/kbp per chromosomes 1-22 per cell line for GSC repeats overlaid as black dots on boxplots. GSCs plotted separately to other cell lines to also allow GSC inter cell line comparison. ES GSCs in red, G7 GSCs in green, R10 GSCs in purple.

# Chapter 4 supplementary figures



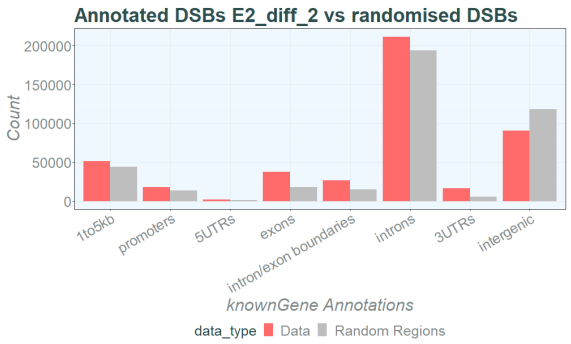

**Supplemental: DSBs per gene adjusted to gene length with annotated long neural genes: G7 GSCs**

Gene DSBs adjusted to gene length (DSBs/kbp) against length of gene. Mean DSB density across repeats 1-2 reported in E2 GSCs and differentiated cells, repeats 1-3 reported in G7 differentiated cells and repeats 1-3 reported in R10 GSCs and differentiated cells. GSCs displayed in turquoise and differentiated cells displayed in red. X-axis displays log-transformed

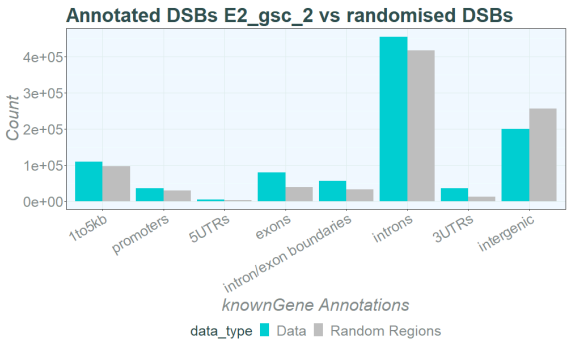
gene length, y-axis displays long-transformed gene length-adjusted DSBs (DSBs/kbp). Log-transformation performed for visualisation purposes. Turquoise dots represent individual genes. Long neural genes represented as dark blue dots and are annotated by gene name in red.

# Chapter 5 supplementary figures

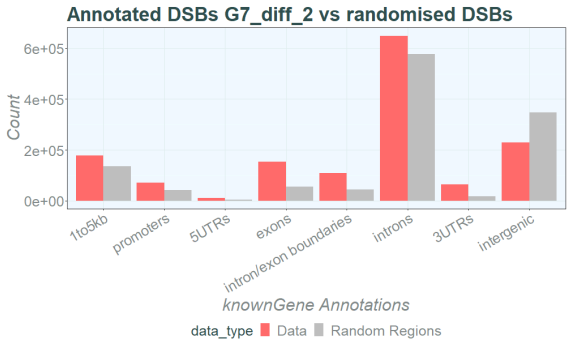
(a)



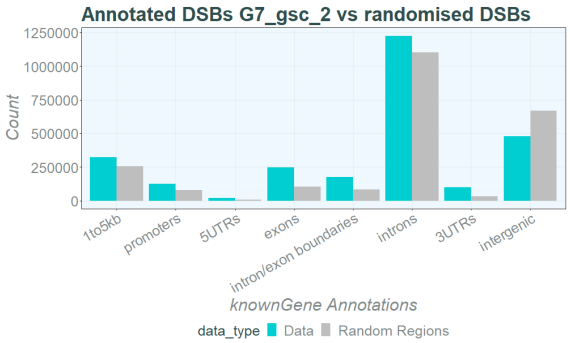

(b)



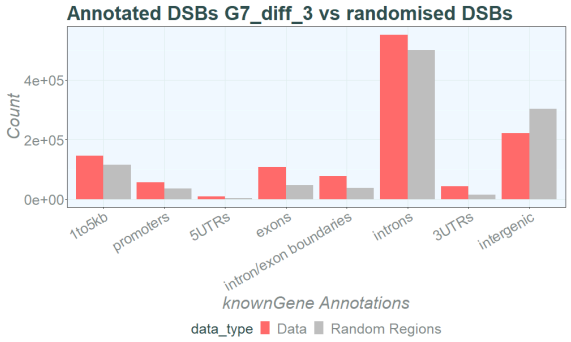

(c)



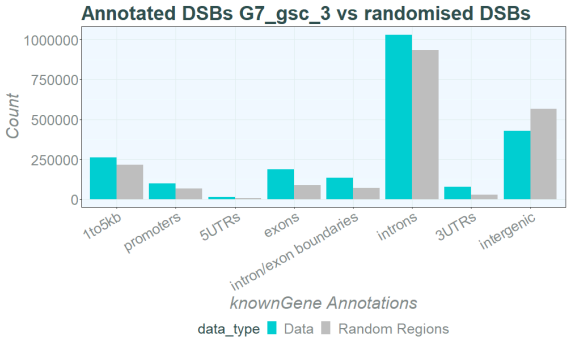

(d)



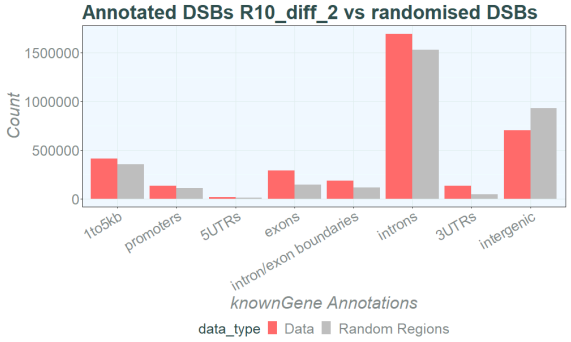

(e)



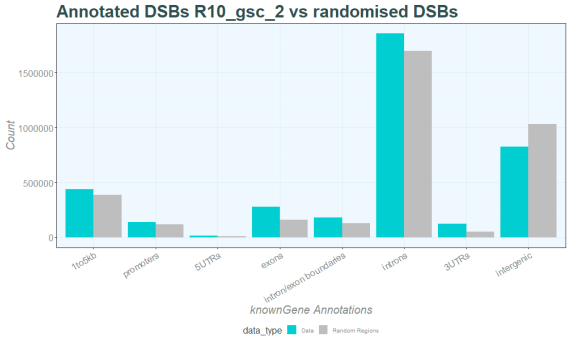

(f)



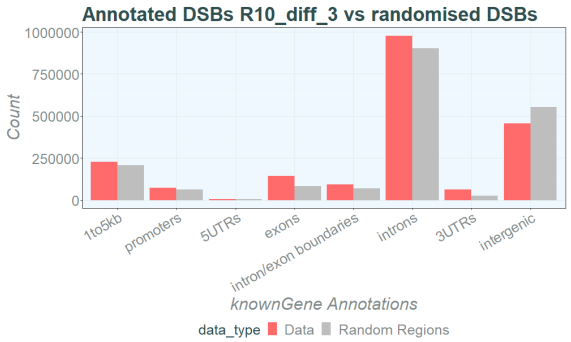

(g)



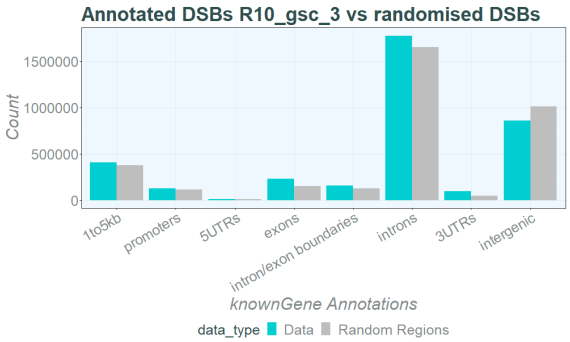

(h)



(i)

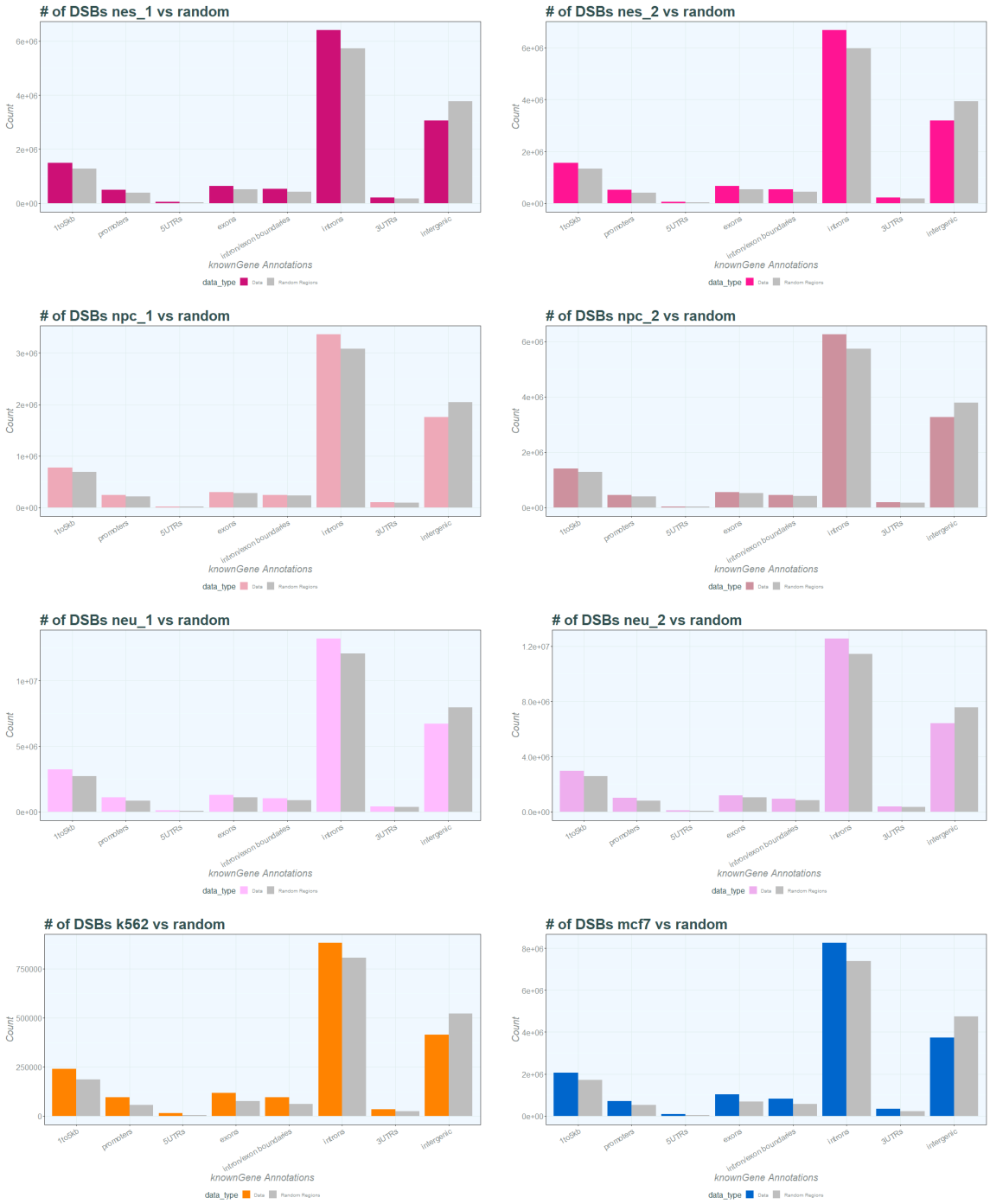



(j)




**Supplemental: Annotated DSB locations vs random expected DSB locations**

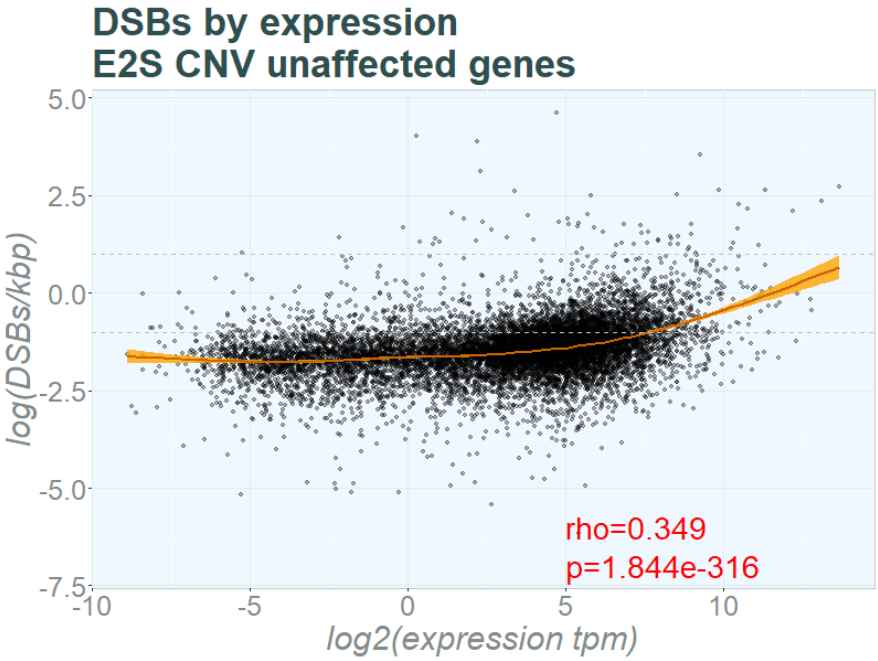
Actual versus expected DSBs per annotated genomic region in GSC and differentiated replicates 2 and 3. Actual DSBs in turquoise (GSCs) or red (differentiated cells), expected DSBs in grey. Order: regions between 1-5 kbp from transcription start sites, promotors <1 kbp from transcription start site, 5' UTRs, exons, intro/exon boundaries, introns, 3' UTRs and intergenic regions. For overlapping annotation site priority: Promoters, 1 to 5 kbp promoters, 5' UTRs, 3' UTRs, exons, introns, intron/exon boundaries, and intergenic sites. (a) E2 differentiated rep 2. (b) E2 GSC rep 2. (c) G7 differentiated rep 2. (d) G7 GSC rep 2. (e) G7 differentiated rep 3. (f) G7 GSC rep 3. (g) R10 differentiated rep 2. (h) R10 GSC rep 2. (h) R10 differentiated rep 2. (i) R10 GSC rep 2.

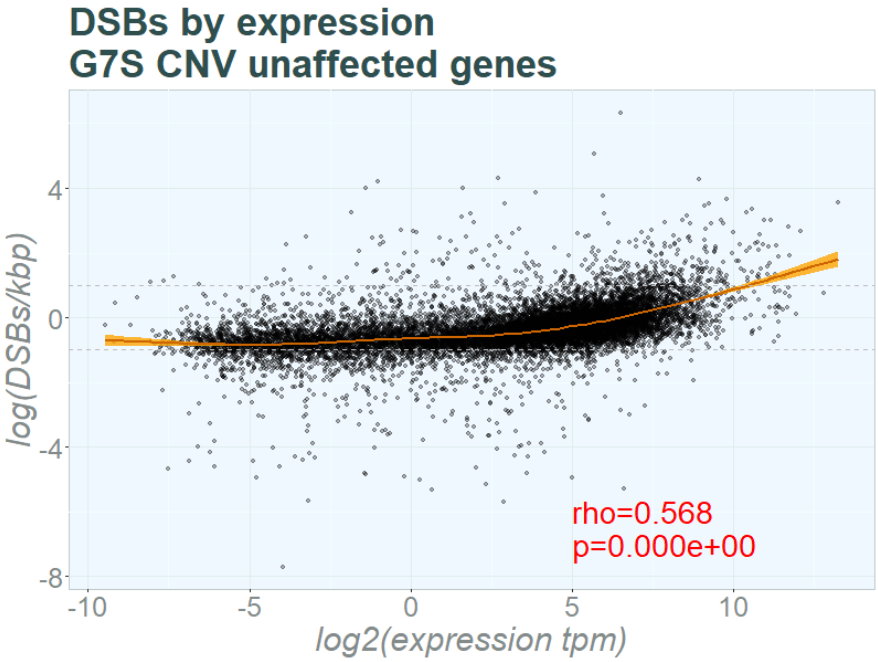


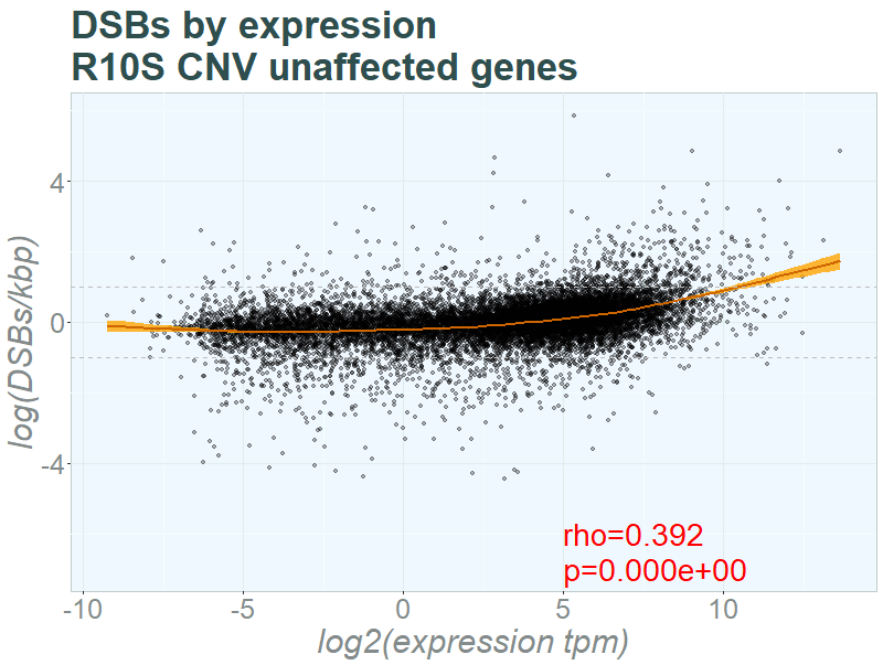

**Supplemental: Annotated DSB locations vs random expected DSB locations for neural cells and commercial cancer cell lines**

Actual versus expected DSBs per annotated genomic region in neural cell replicates NES, NPC and NEU in repeats 1 and 2 (pink) and in K562 (orange) and MCF7 cells (blue). Actual DSBs in colour, expected DSBs in grey. Order: regions between 1-5 kbp from transcription start sites, promotors <1 kbp from transcription start site, 5' UTRs, exons, intro/exon boundaries, introns, 3' UTRs and intergenic regions. For overlapping annotation site priority: Promoters, 1 to 5 kbp promoters, 5' UTRs, 3' UTRs, exons, introns, intron/exon boundaries, and intergenic sites. (a) NES rep 1. (b) NES rep 2. (c) NPC rep 1. (d) NPC rep 2. (e) NEU rep 1. (f) NEU rep 2. (g) K562 cells. (h) MCF7 cells.

## Chapter 6 supplementary figures

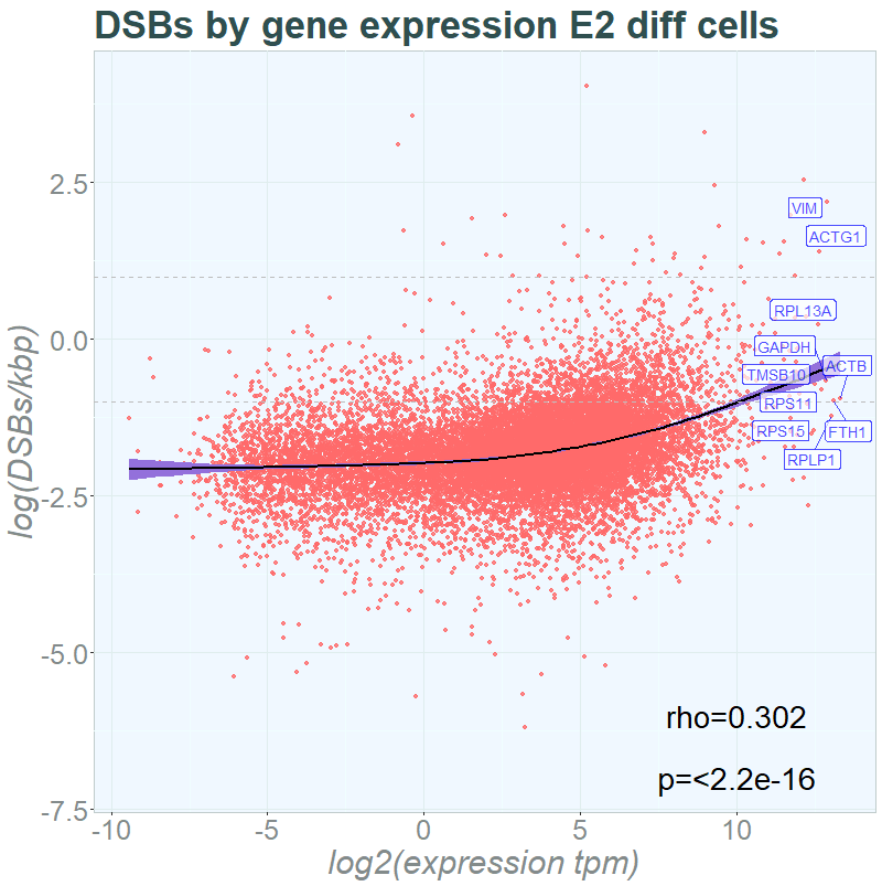


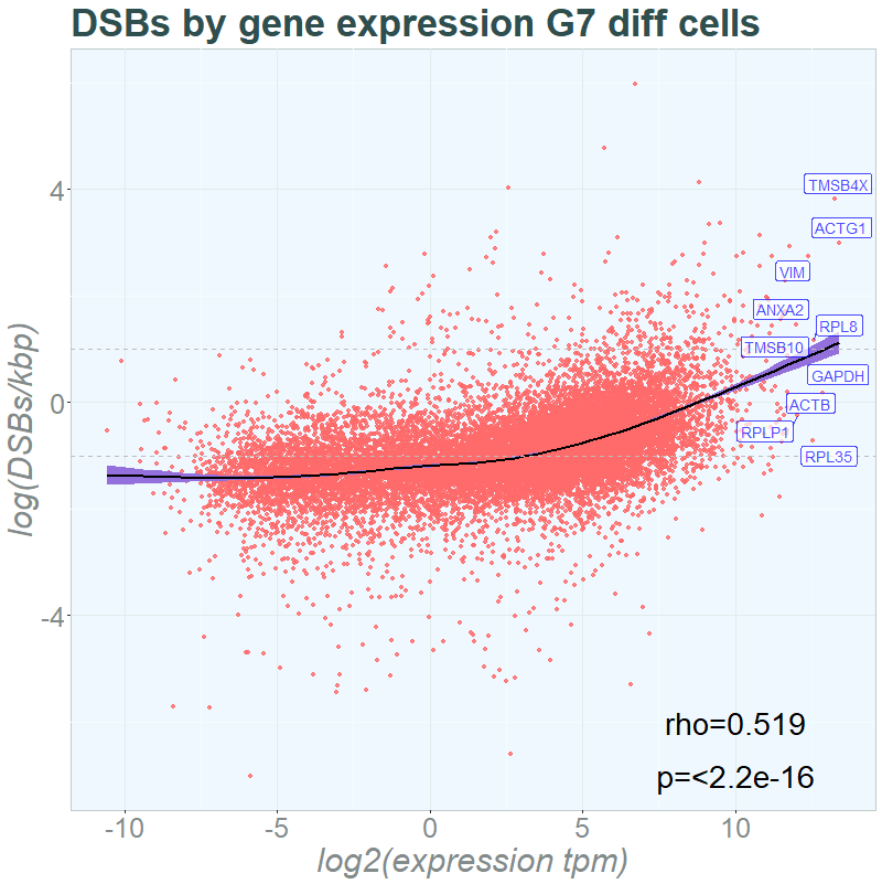



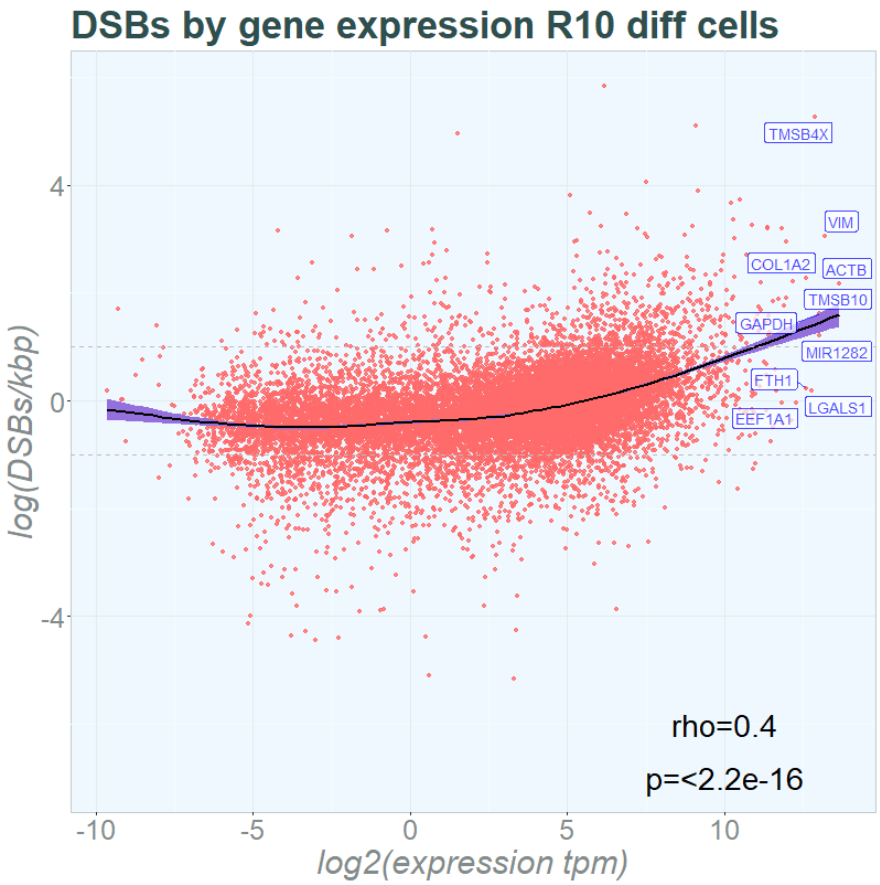



**Supplemental: Gene expression and DSB density in genes with no CNV**

Genes with no CNV were also plotted individually. Mean DSB density from BLISS data and mean gene expression (TPM) from RNAseq data are displayed before and after copy number adjustment. Orange lines represent locally estimated scatterplot smoothing (LOESS) using non-parametric local regression to estimate the shape of the data. Spearman rank correlation estimates with rho and p values displayed in the text. Individual dots mark CNV-affected genes. Data was log transformed for visualisation. (a) E2 GSC line. (b) G7 GSC line. (c) R10 GSC line.


DSBs by gene expression E2 diff cells
log(DSBs/kbp)
log2(expression tpm)
rho=0.302
p=<2.2e-16
VIM
ACTG1
RPL13A
GAPDH
TMSB10
ACTB
RPS11
RPS15
FTH1
RPLP1

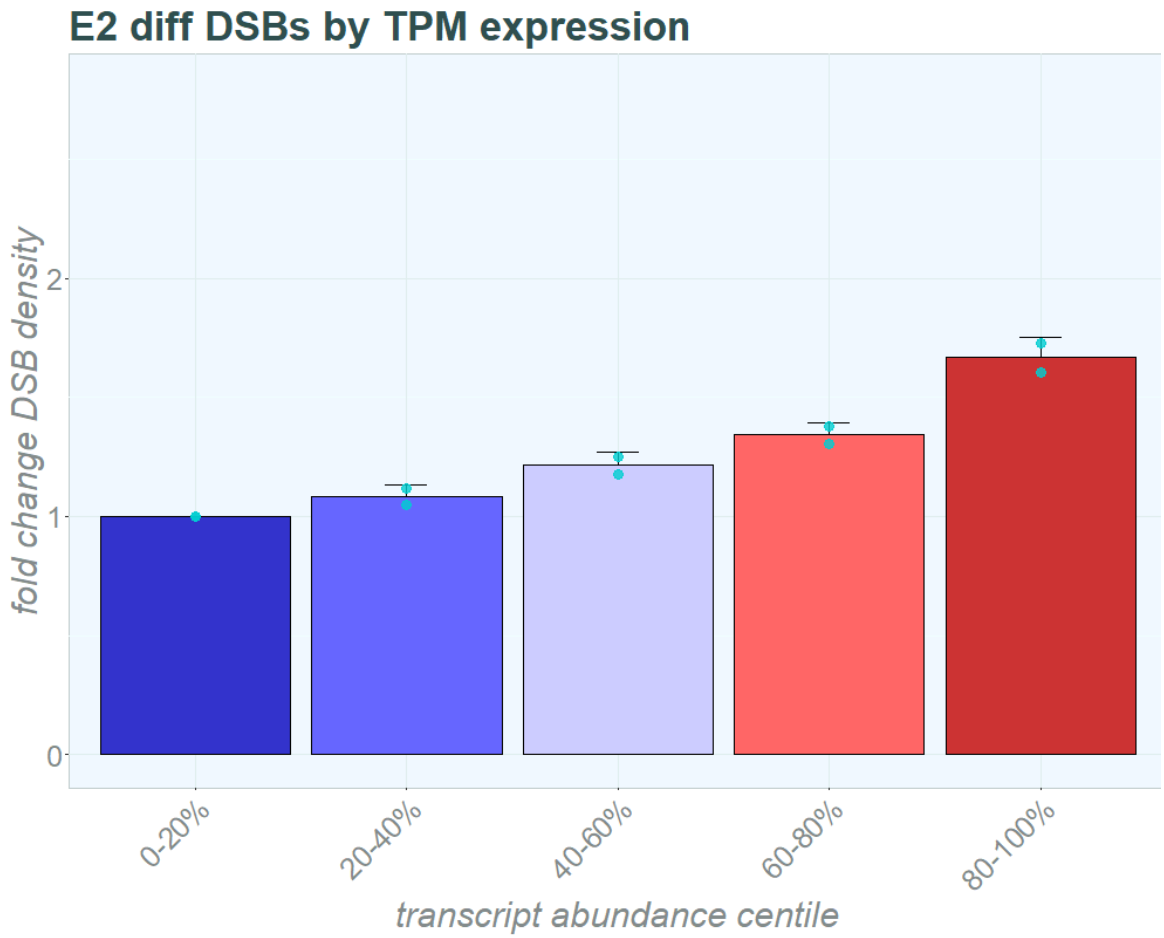

DSBs by gene expression G7 diff cells
log(DSBs/kbp)
log2(expression tpm)
rho=0.519
p=<2.2e-16
TMSB4X
ACTG1
VIM
ANXA2
RPL8
TMSB10
GAPDH
ACTB
RPLP1
RPL35




**Supplemental: Gene expression and DSB density in E2, G7 and R10 differentiated lines.**

Mean gene expression (TPM) plotted against logged mean DSB density (DSB/kbp) per gene. Expression means calculated across RNAseq repeats, DSB means calculated across BLISS repeats. DSB/kbp and TPM expression are log-transformed for visualisation and represented as a scatterplot. The top ten genes with the highest expression are represented in blue text. Horizontal dotted lines represent -1 and +1 log(DSB/kbp). The purple line represents locally estimated scatterplot smoothing (LOESS) using non-parametric local regression to estimate the shape of the data. Statistical analysis was performed by non-parametric Spearman Rank correlation with rho values and p values displayed on the graphs (a) E2 differentiated line. (b) G7 differentiated line. (c) R10 differentiated line.

(a)

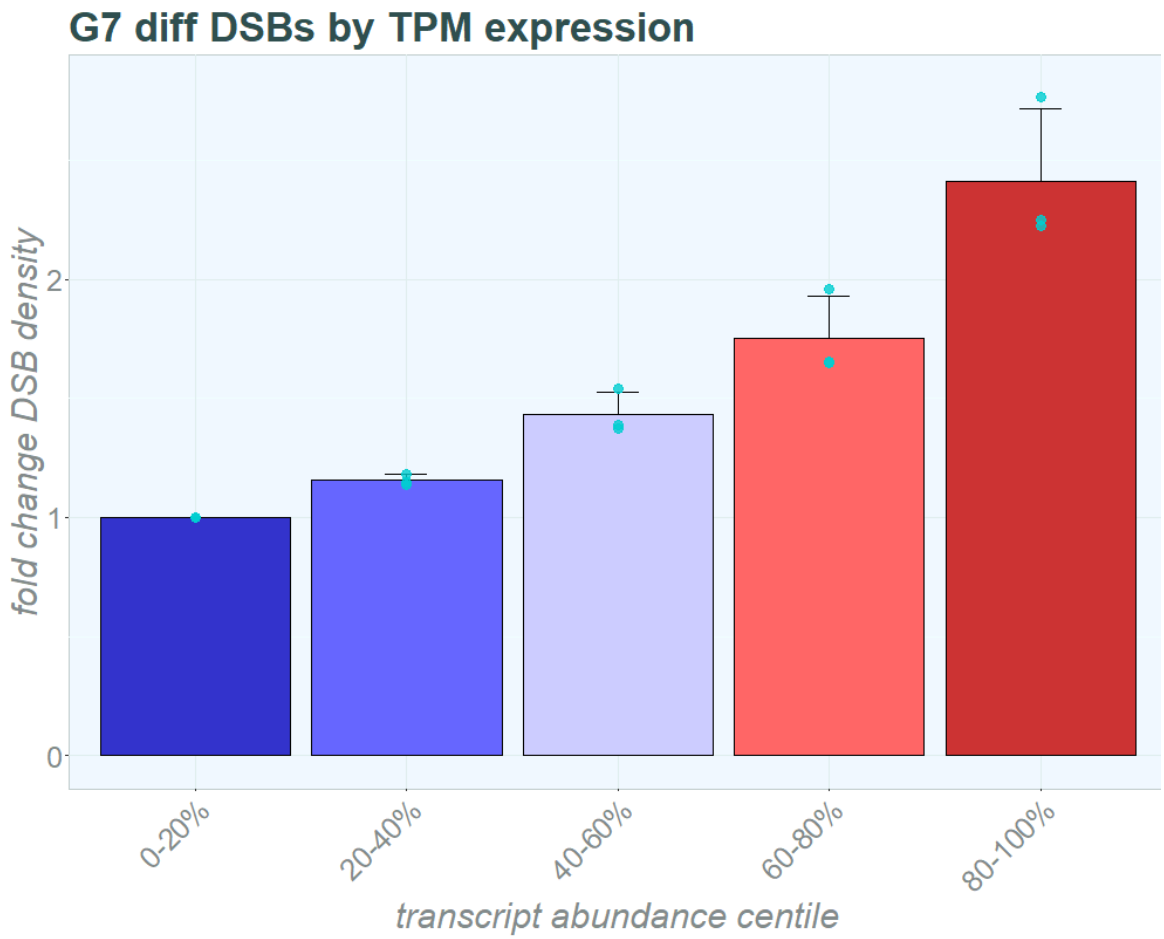



(b)

E2 differentiated exp. By 5ths p-values

| E2 GSC diff. By 5ths p-values | | ***0-20%*** | ***20-40%*** | ***40-60%*** | ***60-80%*** | ***80-100%*** |
|---|---|---|---|---|---|---|
| | ***0-20%*** | 1 | | | | |
| | ***20-40%*** | 0.60129 | 1 | | | |
| | ***40-60%*** | 0.05790 | 0.26528 | 1 | | |
| | ***60-80%*** | 0.00872 | 0.02840 | 0.28363 | 1 | |
| | ***80-100%*** | 0.0004 | 0.00074 | 0.00245 | 0.01084 | 1 |

(c)

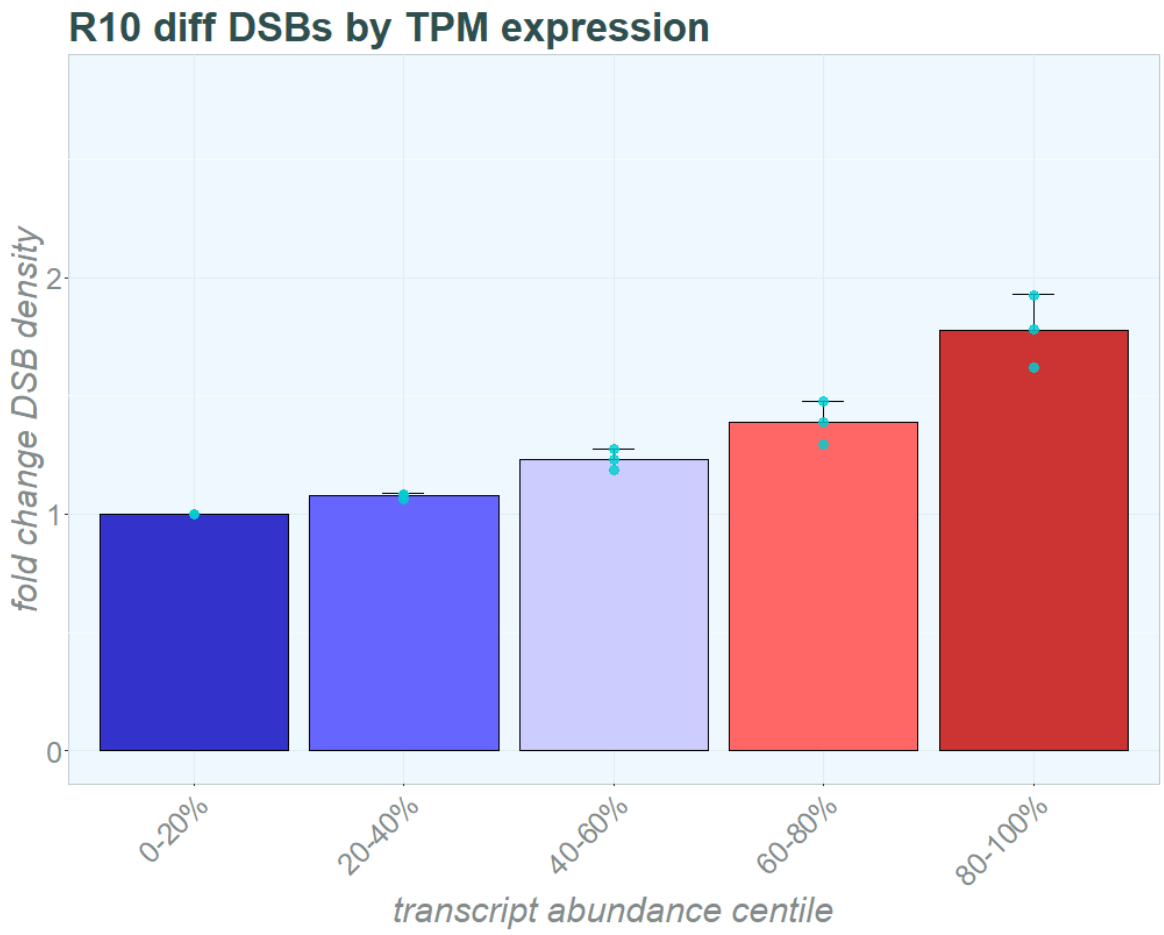



(d)

G7 differentiated exp. By 5ths p-values

| G7 diff exp. By 5ths p-values | | ***0-20%*** | ***20-40%*** | ***40-60%*** | ***60-80%*** | ***80-100%*** |
|---|---|---|---|---|---|---|
| | ***0-20%*** | 1 | | | | |
| | ***20-40%*** | 0.76931 | 1 | | | |
| | ***40-60%*** | 0.05337 | 0.29792 | 1 | | |
| | ***60-80%*** | 0.00155 | 0.00815 | 0.19405 | 1 | |
| | ***80-100%*** | 0.00001 | 0.00002 | 0.00019 | 0.00411 | 1 |

(e)




(f)

R10 differentiated exp. By 5ths p-values

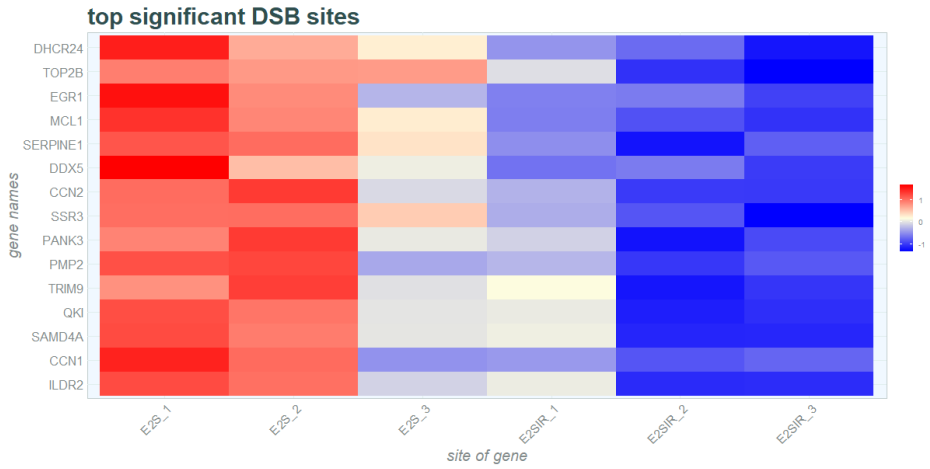
| R10 diff exp. By 5ths p-values | | ***0-20%*** | ***20-40%*** | ***40-60%*** | ***60-80%*** | ***80-100%*** |
|---|---|---|---|---|---|---|
| | ***0-20%*** | 1 | | | | |
| | ***20-40%*** | 0.77990 | 1 | | | |
| | ***40-60%*** | 0.03790 | 0.21437 | 1 | | |
| | ***60-80%*** | 0.00124 | 0.00617 | 0.20928 | 1 | |
| | ***80-100%*** | <0.00001 | 0.00001 | 0.00008 | 0.00124 | 1 |

**Supplemental: Quintile gene expression and DSB density in differentiated lines E2, G7 and R10.**

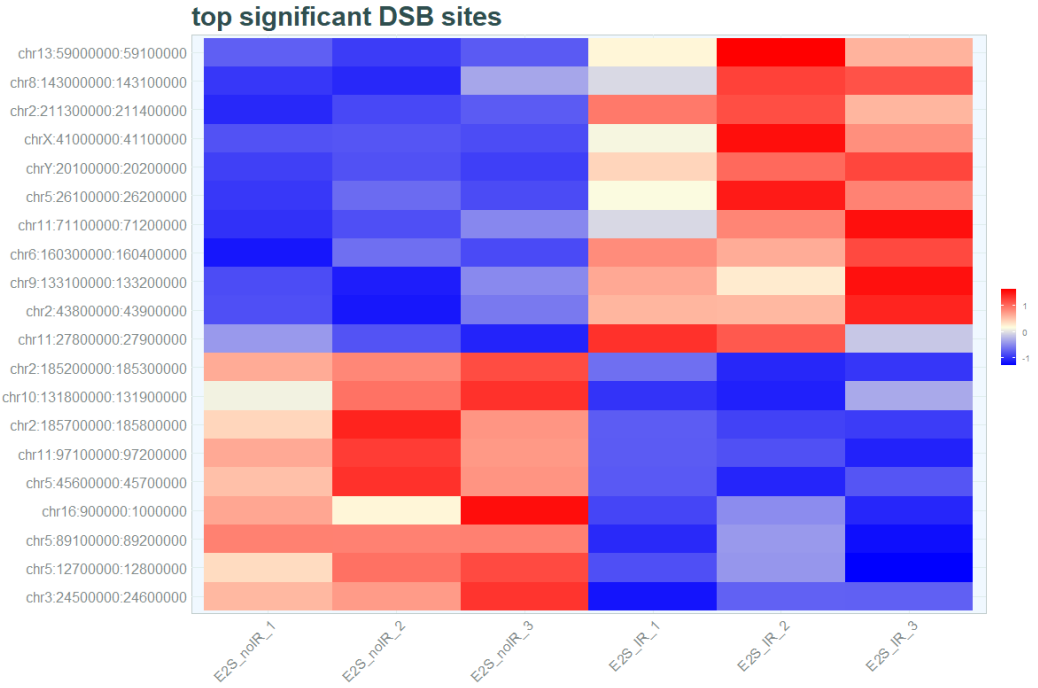
Mean DSB density per gene by expression quintiles from lowest (0-20% in blue) to highest expression (80-100% in red) displayed as a violin plot. Bar chart displays mean of medians of the fold-change in DSBs/kbp from the lowest expression quintile. Whiskers represent standard deviation. Replicate medians displayed in turquoise. Significance testing performed using ANOVA and post-hoc analysis by Tukey test. Corresponding p-values are displayed in tables below between quintiles. Red highlighted results demonstrate significant results of an adjusted p-value <0.05. (a) Mean of duplicate repeats for E2 differentiated lines using RNAseq and BLISS data. DSBs by transcription quintiles from least to most expressed genes. (b) E2 corresponding p-values generated in post-hoc pairwise testing. (c) Mean of triplicate repeats for G7 differentiated lines using RNAseq and BLISS data. DSBs by transcription quintiles from least to most expressed genes. (d) G7 corresponding p-values generated in post-hoc pairwise testing. (e) Mean of triplicate repeats for R10 differentiated lines using RNAseq and BLISS data. DSBs by transcription quintiles from least to most expressed genes. (f) R10 corresponding p-values generated in post-hoc pairwise testing.

# Chapter 7 supplementary figures

(a)




(b)




**Supplemental: Differential DSBs E2 GSCs 0 Gy vs 10 Gy IR 6 hours**

E2 GSCs differentially broken regions 0 Gy vs IR 10 Gy 6 hours. Log2 fold change >0.5/-<0.5 and adjusted p-value <0.05 (BHP corrected). (a) Heatmap of differentially broken genes. 0 Gy treated cells repeats 1-3 (E2S_1, E2S_2, E2S_3). IR-treated cells repeats 1-3 (E2SIR_1, E2SIR_2, E2SIR_3). (b) Heatmap of differentially broken 100 kbp regions. 0 Gy treated cells repeats 1-3 (E2S_noIR_1, E2S_noIR_2, E2S_noIR_3). IR-treated cells repeats 1-3 (E2S_IR_1, E2S_IR_2, E2S_IR_3).

# Supplementary Files

The following supplementary files are provided for processing fastq files for BLISS and ATAC-seq. Pattern files for BLISS are not provided within the body of the thesis but have been submitted as additional .txt files. Additionally, the immunofluorescence ImageJ foci analysis macro is also provided which was developed with Dr Mark Jackson. The RStudio scripts are not provided within the body of the thesis but the files for each chapter have been submitted as additional material as .R files and the titles for each are listed below.

# BLISS commandline files

BLISS commandline files have been submitted below and the steps for use are available in the material and methods. “My globals” refers to the reference document to call up the required files for processing. “Run bliss pipeline script” refers to the automated ordering and running of the individual .sh files as they are completed. BLISS processing was run on the Edinburgh high performance computing cluster “Eddie”.

## BLISS preprocessing pipeline

### Run bliss pipeline script

```
#!/usr/bin/env bash

# run_bliss_pipeline.sh FQNAME LABEL DIRNAME PATTERNS

FQNAME=$1  # G7-1_S23_R1_001
LABEL=$2  # G7-1
DIRNAME=$3 # G7_bliss
PATTERNS=$4 # pattfiles.list

#####################################################################
###########
# DEFINING VARIABLES
source $PROJDIR/scripts/myglobals.sh
sdir=$PROJDIR/scripts/bliss
```

```
#########################################################################
###########
# PREPARE DIRECTORY STRUCTURE
outdir=$PROJDIR/bliss/hg37/$DIRNAME/$LABEL
mkdir -p $outdir
tmpdir=$SCRATCHDIR/$DIRNAME/$LABEL
mkdir -p $tmpdir

#########################################################################
###########
# FIND DATA FILES:
find $FASTQDIR -type f -iname "*$FQNAME*.fastq.gz" | sort > filelist_"$LABEL"
# PRINT TO TERMINAL THE NAMES OF THE FASTQ FILES THAT HAVE BEEN FOUND
numb_of_files=`cat filelist_"$LABEL" | wc -l`
r1=`cat filelist_"$LABEL" | head -n1`
echo "R1 is " $r1
if [ $numb_of_files == 2 ]; then
   r2=`cat filelist_"$LABEL" | tail -n1`
   echo "R2 is " $r2
fi
rm filelist_"$LABEL"

#################################################################
# Reformat files and filter for sequence pattern
echo "qsub -V -v R1=$r1 -v R2=$r2 -v LABEL=$LABEL -v DIRNAME=$DIRNAME -v
PATTERNS=$PATTERNS $sdir/filter_fastq.sh"
qsub -V -v R1=$r1 -v R2=$r2 -v LABEL=$LABEL -v DIRNAME=$DIRNAME -v
PATTERNS=$PATTERNS $sdir/filter_fastq.sh

#################################################################
# Align the reads to the reference.
echo "qsub -V -hold_jid blissFiltfq -v DIRNAME=$DIRNAME/$LABEL
$sdir/align_reads.sh"
qsub -V -hold_jid blissFiltfq -v DIRNAME=$DIRNAME/$LABEL $sdir/align_reads.sh
```

```
#####################################################################
# Filter out duplicates and get unique UMIs
echo "qsub -V -hold_jid blissAlign -v DIRNAME=$DIRNAME/$LABEL
$sdir/filter_duplicates.sh"
qsub -V -hold_jid blissAlign -v DIRNAME=$DIRNAME/$LABEL
$sdir/filter_duplicates.sh

#####################################################################
# Get Summary for the data (number of reads after each filter, etc.)
echo "qsub -V -hold_jid blissfiltdups -v LABEL=$LABEL -v DIRNAME=$DIRNAME
$sdir/get_bliss_summary_stats.sh"
qsub -V -hold_jid blissfiltdups -v LABEL=$LABEL -v DIRNAME=$DIRNAME
$sdir/get_bliss_summary_stats.sh

#####################################################################
# Filter out blacklisted regions and cap DSB peaks
echo "qsub -V -hold_jid blissfiltdups -v DIRNAME=$DIRNAME/$LABEL
$sdir/filter_blacklist_cap.sh"
qsub -V -hold_jid blissfiltdups -v DIRNAME=$DIRNAME/$LABEL
$sdir/filter_blacklist_cap.sh

#####################################################################
# Make bigWig files from the bedgraphs, useful for visualization and
# for downstream analysis
ls $outdir/*bfilt_cap.bedgraph.gz > $outdir/$LABEL"_capbg".list

###not working
#BEDGRAPHS=/$DIRNAME/$LABEL/$LABEL"_capbg".list
#echo $BEDGRAPHS

#echo "qsub -v BEDGRAPHS=$BEDGRAPHS scripts/bedgraphs_to_bigwigs.sh"
#qsub -v BEDGRAPHS=$BEDGRAPHS scripts/bedgraphs_to_bigwigs.sh
```

qsub -v BEDGRAPH=/home/v1sderb2/bioinfsvice/GBM_BLISS/bliss/hg37/E2IR-1_bliss/E2IR-1/stem_1A.q30-umi-bfilt_cap.bedgraph.gz scripts/bedgraph_to_bigwig.sh

**My globals**

```
#  Define some global variables for the project
#
export PROJDIR=/home/v1sderb2/bioinfsvice/GBM_BLISS #my home directory
REFGENOME=/exports/igmm/software/pkg/el7/apps/bcbio/share2/genomes/Hsapiens/GRCh37/bwa/GRCh37.fa #location of reference genome on eddie
FASTQDIR=/exports/eddie/scratch/sderby/GBM_BLISS/FASTQ_2022/G7_bliss #where my fastq files are
SCRATCHDIR=/exports/eddie/scratch/sderby/bliss

#### Custom R library with more R packages in it.
R_LIBS_USER=/exports/igmm/eddie/NextGenResources/software/R/x86_64-pc-linux-gnu-library/3.2



# Annotation files
ENSGENES=$PROJDIR/hg19_annotation/ref-transcripts_merge.bed
CENTROMERE=$PROJDIR/hg19_annotation/hg19-telomere-centromere-telomere.bed
BLACKLIST=$PROJDIR/hg19_annotation/consensusBlacklist.bed
CHRLENGTHS=$PROJDIR/hg19_annotation/Homo_sapiens_assembly19_chr.lengths

##############################################
# Location of certain programs
# BLISS
# scan_for_matches - used in BLISS pipeline in filter_fastq.sh
#SCANDIR=/home/tballing/NextGenResources/software/scan_for_matches/scan_for_matches
SCANDIR=/home/v1sderb2/bioinfsvice/GBM_BLISS/scan_for_matches.git
```

```
##############################################
# ATAC-seq
# RGTDATA is necessary for footprinting analysis
RGTDATA=/home/tballing/bioinfsvice/ref_files/rgtdata



##############################################
#example for calling files

FQNAME=E2S-1_S46_L001_
LABEL=E2IR-1
DIRNAME=E2IR_bliss
PATTERNS=pattfiles.list
sdir=$PROJDIR/scripts/bliss
```

**UMI filtering**

```
# -*- coding: utf-8 -*-

import sys
import csv
import numpy as np

filename = sys.argv[1]

with open(filename, 'r') as f:
    reader = csv.reader(f)
    data = list(reader)

# sys.stderr.write("data is %d\n" % len(data))

# GROUP TOGETHER CONSECUTIVE IDENTICAL UMIS, ASSUMING THEIR PROXIMITY
# row_old = data[0]
# data_aggegated_by_umi_identity = [row_old]
# for row in data[1:]:
#     if (row_old[0]==row[0] and row_old[3]==row[3] and row_old[4]==row[4]):
```

```
#       row[5] = str(int(row[5])+int(row_old[5]))
#       if int(row_old[5]) > int(row_old[5]):
#           row[1:3] = row_old[1:3]
#       del data_aggegated_by_umi_identity[-1]
#    data_aggegated_by_umi_identity.append(row)
#    row_old = row

# GROUP TOGETHER CLOSE SPATIAL CONSECUTIVE READS WHOSE UMI DIFFERS AT
MOST BY 2 MISMATCHES
data_aggegated_by_umi_similarity = []
skipped= []
skipped_new= []
space_gap = 20
mm_gap = 2
rowi=1
oldi=0
while (rowi < len(data)) or len(skipped_new) >1:
    if rowi >= len(data):
        data_aggegated_by_umi_similarity.append(row_old)
        oldi=skipped_new.pop(0)
        rowi=skipped_new.pop(0)
        skipped=skipped_new
        skipped_new=[]
    row=data[rowi]
    row_old=data[oldi]
    s1 = row_old[4]
    s2 = row[4]
    numb_mismatches = sum(c1!=c2 for c1,c2 in zip(s1,s2))
    dist = abs(int(row[1])-int(row_old[1]))
    if (row_old[0]==row[0] and dist<=space_gap and row_old[3]==row[3] and
numb_mismatches<=mm_gap):
        if int(row[5]) > int(row_old[5]):
            row_old[1:5] = row[1:5]
        row_old[5] = str(int(row[5])+int(row_old[5]))
        if len(skipped) >0:
```

```
            rowi=skipped.pop(0)
        else:
            rowi=rowi+1
    else:
        if dist > space_gap or row_old[0] != row[0]:
            data_aggegated_by_umi_similarity.append(row_old)
            tmpskipped=skipped_new
            if len(skipped) >0:
                tmpskipped.append(rowi)
            tmpskipped.extend(skipped)
            skipped=tmpskipped
            skipped_new=[]
            if len(skipped) >0:
                oldi=skipped.pop(0)
                if len(skipped) >0:
                    rowi=skipped.pop(0)
            else:
                oldi=rowi
                rowi=rowi+1
        else:
            skipped_new.append(rowi)
            if len(skipped) >0:
                rowi=skipped.pop(0)
            else:
                rowi=rowi+1

data_aggegated_by_umi_similarity.append(row_old)
if len(skipped_new) == 1:
    row=data[skipped_new[0]]
    data_aggegated_by_umi_similarity.append(row)



thefile = open(sys.argv[2], 'w')
for item in data_aggegated_by_umi_similarity:
    thefile.write('\t'.join(item)+'\n')
```

```
print('Done with filtering UMIs!')
```

**Filter fastq**

```
#!/usr/bin/env bash

# THIS SCRIPT CAN BE CALLED AS
# qsub -v R1=file.fastq.gz -v R2=file.fastq.gz -v LABEL=label -v DIRNAME=X_bliss -
v PATTERNS=patternfiles.list filter_fastq.sh

#$ -N blissFiltfq
#$ -cwd
#$ -j y
#$ -l h_rt=02:30:00
#$ -l h_vmem=3G
#$ -pe sharedmem 4
#  $ -o
/exports/eddie/scratch/tballing/errorlogs/$JOB_NAME.o$JOB_ID.$TASK_ID

unset MODULEPATH
. /etc/profile.d/modules.sh
module load igmm/apps/bcbio/1.0.8
source $PROJDIR/scripts/myglobals.sh

#####################################################################
# DEFINING VARIABLES and PREPARE DIRECTORY STRUCTURE
outdir=$PROJDIR/bliss/hg37/$DIRNAME/$LABEL
mkdir -p $outdir
tmpdir=$SCRATCHDIR/$DIRNAME/$LABEL
mkdir -p $tmpdir
pattdir=$PROJDIR

#####################################################################
# Reformat fastq and filter for sequence pattern
r1=$R1
```

```
r2=$R2
zcat -dc $R1 | paste - - - - | cut -f 1,2 | sed 's/^@/>/' | tr "\t" "\n" >
$tmpdir/r1.fa
zcat -dc $R1 | paste - - - - | LC_ALL=C sort --temporary-directory=$tmpdir -k1,1
> $tmpdir/r1oneline.fq

label=$LABEL
totreads=`wc -l < $tmpdir/r1oneline.fq`
echo "pattern          count  percent" > $label.stats
echo "none $totreads" | awk '{p=$2/tot; print $1"\t"$2"\t"p*100}' tot=$totreads >>
$label.stats

while read -r pf; do
        # Filter for the pattern
        pname=`basename $pf .txt | sed 's/pattern_//'`
        pf=$pattdir/$pf
        echo "pattern file is $pf, pname is $pname"
        cat $tmpdir/r1.fa \
        | parallel --tmpdir $tmpdir --block 100M -k --pipe -L 2 \
        "$SCANDIR/scan_for_matches $pf - " > $tmpdir/$pname.r1.fa

        cat $tmpdir/$pname.r1.fa | tr '>' '@' \
        | cut -d'[' -f1 | sed 's/:$//' | paste - - \
        | awk '{print $1,$NF}' \
        | LC_ALL=C sort --temporary-directory=$tmpdir -k1,1 >
$tmpdir/$pname.ID_genomic

        LC_ALL=C join $tmpdir/$pname.ID_genomic $tmpdir/r1oneline.fq \
        | awk '{print $1,$2,"+",substr($6, length($6)-length($2)+1, length($6))}' \
        | tr " " "\n" > $tmpdir/$pname.r1.2b.aln.fq

#########################################################
# If it's paired end, then get the matching second read for the filtered ones.
        if [ ! -z $R2 ]; then
```

```
        zcat -dc $R2 | paste - - - - | LC_ALL=C sort --temporary-directory=$tmpdir
-k1,1 > $tmpdir/r2oneline.fq
   # Filter for the read1s with the pattern
        LC_ALL=C join $tmpdir/$pname.ID_genomic $tmpdir/r2oneline.fq | cut -
d' ' -f 1,4-6 | tr " " "\n" > $tmpdir/$pname.r2.2b.aln.fq
        fi

        nfilt=`wc -l < $tmpdir/$pname.ID_genomic`
        echo "$pname $nfilt" | awk '{p=$2/tot; print $1"\t"$2"\t"p*100}'
tot=$totreads >> $label.stats

done < $PATTERNS
```

**Align reads**

```
#!/usr/bin/env bash

# THIS SCRIPT CAN BE CALLED AS
# qsub -v DIRNAME=directory align_reads.sh
# DIRNAME should be something like E2_bliss/E2-1

#$ -N blissAlign
#$ -cwd
#$ -j y
#$ -l h_rt=06:30:00
#$ -l h_vmem=3G
#$ -pe sharedmem 3
#  $ -o
/exports/eddie/scratch/tballing/errorlogs/$JOB_NAME.o$JOB_ID.$TASK_ID

unset MODULEPATH
. /etc/profile.d/modules.sh
module load igmm/apps/bcbio/1.0.8
module load igmm/apps/samtools/1.6
source $PROJDIR/scripts/myglobals.sh
```

```
######################################################################
######
# DEFINING VARIABLES and DIRECTORIES
numproc=3          # number of threads to be used

# source myglobals.sh
outdir=$PROJDIR/bliss/hg37/$DIRNAME
tmpdir=$SCRATCHDIR/$DIRNAME
mkdir -p $tmpdir

#####################################################################
# Align the reads to the reference.
exps=""
for f1 in `ls $tmpdir/*_t.r1.2b.aln.fq`; do
        exp=`basename $f1 _t.r1.2b.aln.fq`
        exps=$exps" "$exp
        f2=$tmpdir/$exp"_t.r2.2b.aln.fq"
        if [ -s $f2 ]; then
                echo "bwa mem -t $numproc $REFGENOME $f1 $f2 >
$tmpdir/$exp.sam"
                bwa mem -t $numproc $REFGENOME $f1 $f2 > $tmpdir/$exp.sam
        else
                echo "bwa mem -t $numproc $REFGENOME $f1 > $tmpdir/$exp.sam"
                bwa mem -t $numproc $REFGENOME $f1 > $tmpdir/$exp.sam
        fi
done

#####################################################################
# filter out reads with low quality mapping
# (quality less than 30 is what they use in published BLISS processing).
for experiment in $exps; do
        if [ -s $f2 ]; then
                samtools view -Sb -f 66 -q 30 $tmpdir/$experiment.sam >
$tmpdir/$experiment.q30.bam
        else
```

```
                samtools view -Sb -q 30 $tmpdir/$experiment.sam >
$tmpdir/$experiment.q30.bam
        fi
        samtools sort $tmpdir/$experiment.q30.bam -o
$tmpdir/$experiment.q30.sorted.bam
        samtools index $tmpdir/$experiment.q30.sorted.bam
        cp $tmpdir/$experiment.q30.sorted.bam $outdir
        cp $tmpdir/$experiment.q30.sorted.bam.bai $outdir
done
```

**Filter duplicates**

```
#!/usr/bin/env bash

# THIS SCRIPT CAN BE CALLED AS
# qsub -v DIRNAME=directory filter_duplicates.sh
# DIRNAME is something like E2_bliss/E2-1

#$ -N blissfiltdups
#$ -cwd
#$ -j y
#$ -l h_rt=02:30:00
#$ -l h_vmem=3G
#$ -pe sharedmem 3
#  $ -o
/exports/eddie/scratch/tballing/errorlogs/$JOB_NAME.o$JOB_ID.$TASK_ID

unset MODULEPATH
. /etc/profile.d/modules.sh
module load igmm/apps/bcbio/1.0.8
source $PROJDIR/scripts/myglobals.sh

#####################################################################
######
# DEFINING VARIABLES
outdir=$PROJDIR/bliss/hg37/$DIRNAME
```

```
tmpdir=$SCRATCHDIR/$DIRNAME
mkdir -p $tmpdir

numproc=3          # number of threads to be used

exps=""
for f in `ls $tmpdir/*_t.ID_genomic`; do
        exp=`basename $f _t.ID_genomic`
        exps=$exps" "$exp
done

#########################################################################
######
# Need to get the UMIs for the aligned reads to filter out
# PCR duplicates
for exp in $exps; do
        samtools view -F 0x10 $tmpdir/$exp.q30.sorted.bam \
        | awk '{print $1"\t"$3"\t"$4"\t+"}' > $tmpdir/forward

        samtools view -f 0x10 $tmpdir/$exp.q30.bam \
        | awk '{print $1"\t"$3"\t"$4+length($10)"\t-"}' > $tmpdir/reverse

        cat $tmpdir/forward $tmpdir/reverse \
        | LC_ALL=C sort --parallel=$numproc -T $tmpdir -k1,1 >
$tmpdir/id.chr.loc.strand

        cat $tmpdir/$exp"_t".r1.fa | tr -d ">" | cut -d'[' -f1 | sed 's/:$//' | paste -
- \
        | awk '{if (NF==4) print; else if (NF==5) print $1,$3,$4,$5}' \
        | LC_ALL=C sort --parallel=$numproc -T $tmpdir -k1,1 >
$tmpdir/id.umi.barcode.genomic

        LC_ALL=C join $tmpdir/id.chr.loc.strand $tmpdir/id.umi.barcode.genomic
\
        | cut -d' ' -f2- \
```

```
        | LC_ALL=C sort --parallel=$numproc -T $tmpdir -t' ' -k1,1 -k2,2n -k3,3 |
uniq -c \
        | awk 'BEGIN{OFS="\t"}{print "chr"$2,$3,$3+1,$4,$5,$1}' >
$tmpdir/$exp.q30.umi.bed
done

#################################################################
# Filter out PCR duplicates by getting unique UMIs
for exp in $exps; do
        cat $tmpdir/$exp.q30.umi.bed \
        | tr "\t" "," > $tmpdir/$exp.q30.umi.csv
        python $PROJDIR/scripts/bliss/umi_filtering_tjb.py
$tmpdir/$exp.q30.umi.csv $tmpdir/$exp.umi_filt1.txt
        cut -f1-3 $tmpdir/$exp.umi_filt1.txt \
                | LC_ALL=C sort -k1,1 -k2,2n | uniq -c \
        | awk '{OFS="\t"; print $2,$3,$4,$1}' > $tmpdir/$exp.q30_chr-loc-
uniqueUMIcount.bedgraph
        gzip $tmpdir/$exp.q30_chr-loc-uniqueUMIcount.bedgraph
        mv $tmpdir/$exp.q30_chr-loc-uniqueUMIcount.bedgraph.gz $outdir
done
```

**Filter blacklists**

```
#!/usr/bin/env bash

# THIS SCRIPT CAN BE CALLED AS
# qsub -v DIRNAME=directory filter_blacklist_cap.sh
# DIRNAME is something like E2_bliss/E2-1

#$ -N blissfilt
#$ -cwd
#$ -j y
#$ -l h_rt=01:30:00
#$ -l h_vmem=3G
# $ -pe sharedmem 3
```

```
#  $ -o
/exports/eddie/scratch/tballing/errorlogs/$JOB_NAME.o$JOB_ID.$TASK_ID

unset MODULEPATH
. /etc/profile.d/modules.sh
module load igmm/apps/bcbio/1.0.8
source $PROJDIR/scripts/myglobals.sh

#########################################################################
######
# DEFINING VARIABLES
outdir=$PROJDIR/bliss/hg37/$DIRNAME

for f in `ls $outdir/*.q30_chr-loc-uniqueUMIcount.bedgraph.gz`; do
        exp=`basename $f .q30_chr-loc-uniqueUMIcount.bedgraph.gz`
        gzip -dc $f | \
                bedtools intersect -v -a stdin -b $CENTROMERE \
                | bedtools intersect -v -a stdin -b $BLACKLIST \
                | awk 'BEGIN{OFS="\t"}{if ($4>50) $4=50; print $1,$2,$3,$4}' \
                | gzip > $outdir/$exp.q30-umi-bfilt_cap.bedgraph.gz
done
```

**Get bliss summary stats**

```
#!/usr/bin/env bash

# THIS SCRIPT CAN BE CALLED AS
# qsub -v LABEL=label -v DIRNAME=directory get_bliss_summary_stats.sh

#$ -N summarystats
#$ -cwd
#$ -j y
#$ -l h_rt=00:30:00
#$ -l h_vmem=2G
# $ -pe sharedmem 3
```

```
#  $ -o
/exports/eddie/scratch/tballing/errorlogs/$JOB_NAME.o$JOB_ID.$TASK_ID

unset MODULEPATH
. /etc/profile.d/modules.sh
module load igmm/apps/bcbio/1.0.8
source $PROJDIR/scripts/myglobals.sh

#########################################################################
######
# PREPARE DIRECTORY STRUCTURE
outdir=$PROJDIR/bliss/hg37/$DIRNAME/$LABEL
tmpdir=$SCRATCHDIR/$DIRNAME/$LABEL

#########################################################################
# Get Summary for the data (number of reads after each filter, etc.)
summaryfile=$tmpdir/$LABEL"_summary.txt"
echo "Number of fragments:" > $summaryfile
cat $tmpdir/r1oneline.fq | wc -l >> $summaryfile

exps=""
for f in `ls $tmpdir/*_t.ID_genomic`; do
        exp=`basename $f _t.ID_genomic`
        exps=$exps" "$exp
done
echo "exps is $exps"

for exp in $exps; do
        echo $exp >> $summaryfile
        echo "Number of fragments with prefix:" >> $summaryfile
        cat $tmpdir/$exp"_t".ID_genomic | wc -l >> $summaryfile
        echo "Alignment statistics:" >> $summaryfile
        samtools flagstat $tmpdir/$exp.sam >> $summaryfile
        echo "Number of left and right cuts:" >> $summaryfile
```

```
	cat $tmpdir/$exp.umi_filt1.txt | grep -v "_" | cut -f4 | sort | uniq -c >>
$summaryfile
	echo "Number of DSB locations:" >> $summaryfile
	gzip -dc $outdir/$exp.q30_chr-loc-uniqueUMIcount.bedgraph.gz | wc -l >>
$summaryfile
	#cat $outdir/$exp.q30_chr-loc-uniqueUMIcount.bed | grep -v "_" | wc -l
>> $summaryfile
done

mv $summaryfile $outdir
```

**Make bigwigs**

```
#!/usr/bin/env bash

#$ -N bwa_index
#$ -cwd
#$ -j y
#$ -l h_rt=06:30:00
#$ -l h_vmem=3G
#$ -pe sharedmem 3

#  $ -o
/exports/eddie/scratch/tballing/errorlogs/$JOB_NAME.o$JOB_ID.$TASK_ID

unset MODULEPATH
. /etc/profile.d/modules.sh
module load igmm/apps/bcbio/1.0.8
module load igmm/apps/samtools/1.6
source $PROJDIR/scripts/myglobals.sh

##########################################################################
######
# DEFINING VARIABLES and DIRECTORIES
numproc=3          # number of threads to be used
```

```
# source myglobals.sh
outdir=$PROJDIR/bliss/hg37/$DIRNAME
tmpdir=$SCRATCHDIR/$DIRNAME
mkdir -p $tmpdir


# Make bigWig files from the bedgraphs, useful for visualization and
# for downstream analysis
ls $outdir/$LABEL/*bfilt_cap.bedgraph.gz > $outdir/$LABEL"_capbg".list
echo "qsub -V -v BEDGRAPHS=$outdir/$LABEL"_capbg".list"
qsub -V -hold_jid blissfilt -v BEDGRAPHS=$outdir/$LABEL"_capbg".list
```

# ATAC-seq

ATAC-seq files were run on commandline via WSL and were not run on the Edinburgh high performance computing cluster “Eddie”. Pipeline analysis was performed through Nextflow nf-core ATAC-seq. The .sh files are provided below and refer to each sample.

## ATAC-seq Nextflow nf-core ATAC-seq .sh files

### E2 samples

```
nextflow run nf-core/atacseq -profile docker -r 2.1.1 -resume \
    --max_cpus 8 --max_memory 30.GB \
    --input E2_ATACseq_samplesheet.csv \
    --outdir results/nf-atacseq \
    --fasta $(pwd)/resources/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
\
    --gtf $(pwd)/resources/Homo_sapiens.GRCh38.110.gtf.gz \
    --blacklist $(pwd)/resources/ENCFF356LFX_exclusion_lists.bed.gz \
    --macs_gsize 2700000000
```

### G7 samples

```
nextflow run nf-core/atacseq -profile docker -r 2.1.1 -resume \
    --max_cpus 8 --max_memory 30.GB \
```

```
    --input G7_ATACseq_samplesheet.csv \
    --outdir results/nf-atacseq \
    --fasta $(pwd)/resources/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
\
    --gtf $(pwd)/resources/Homo_sapiens.GRCh38.110.gtf.gz \
    --blacklist $(pwd)/resources/ENCFF356LFX_exclusion_lists.bed.gz \
    --macs_gsize 2700000000
```

**R10 samples**

```
nextflow run nf-core/atacseq -profile docker -r 2.1.1 -resume \
    --max_cpus 8 --max_memory 30.GB \
    --input R10_ATACseq_samplesheet.csv \
    --outdir results/nf-atacseq \
    --fasta $(pwd)/resources/Homo_sapiens.GRCh38.dna.primary_assembly.fa.gz
\
    --gtf $(pwd)/resources/Homo_sapiens.GRCh38.110.gtf.gz \
    --blacklist $(pwd)/resources/ENCFF356LFX_exclusion_lists.bed.gz \
    --macs_gsize 2700000000
```

## Immunofluorescence ImageJ analysis

Below is a copy of the ImageJ auto-counting foci tool developed with help from Dr Mark Jackson. The tool was designed to analyse the total foci per nucleus and the integrated density per cell nucleus. In addition this tool was also used to identify overlapping foci however this was not used as part of the thesis.

### Macro auto counting tool

```
macro "Foci overlap Action Tool -
C000D11D12D13D14D15D16D17D18D19D1aD1bD1cD1dD21D2dD31D3dD41D4dD51D
5dD61D6dD71D7dD81D8dD91D9dDa1DadDb1DbdDc1DcdDd1DddDe1De2De3De4De5
De6De7De8De9DeaDebDecDedCf00D55D56D57D58D65D66D67D68D75D76D85D86Cf
f0D77D78D87D88C0f0D79D7aD89D8aD97D98D99D9aDa7Da8Da9Daa"{

// user dialogue
```

```
        minA=3;
        maxA=50;
        minG=60;
        minR=50;
        Dialog.create("Select foci parameters");
        Dialog.addNumber("Min. focus area (pixel^2):", minA);
        Dialog.addNumber("Max. focus area (pixel^2):", maxA);
        Dialog.addNumber("Intensity threshold green (0-255):", minG);
        Dialog.addNumber("Intensity threshold red (0-255):", minR);
        Dialog.show;
        minA=Dialog.getNumber();
        maxA=Dialog.getNumber();
        minG=Dialog.getNumber();
        minR=Dialog.getNumber();
print("Settings:\nMin. area="+minA+"\nMax.
area="+maxA+"\nThresholdGreen="+minG+"\nThresholdRed="+minR+"\n");
mainDir = getDirectory("Choose folder containing images");
mainList = getFileList(mainDir);
setBackgroundColor(0, 0, 0);
run("Set Measurements...", "area integrated redirect=None decimal=2");

allfile = newArray();
allnuc = newArray();
allarea = newArray();
allintint = newArray();
allred = newArray();
allredint = newArray();
allredpannuc = newArray();
allgreen = newArray();
allgreenint = newArray();
allgreenpannuc = newArray();
alloverlap = newArray();
setBatchMode(true);
for (i=0; i<mainList.length; i++)
                {
```

```
		if(endsWith(mainList[i], ".tif"))
		{
			open(mainList[i]);
			filename = getTitle();
			print(filename);
			run("Set Scale...", "distance=0 known=0 pixel=1 unit=pixel");
			run("Split Channels");
			run("8-bit");
			selectWindow(filename +" (blue)");
			setBatchMode("show");
			// nuclei detect
			//run("Brightness/Contrast...");
			//waitForUser("Waiting for user. Press Okay to
continue....");
			run("Threshold...");
			waitForUser("Waiting for user. Press Okay to continue....");
			//setThreshold(5, 255);

			run("Analyze Particles...", "size=1000-Infinity show=Masks
exclude include");

			//check watershedding and correct
			run("Watershed");
			//run("Invert LUT");
			setTool("Paintbrush Tool");
			//run("Color Picker...");
			setForegroundColor(0, 0, 0);
			selectWindow("Mask of " + filename + " (blue)");
			setBatchMode("show");
			waitForUser("Click OK to proceed when masks appropriate");
			run("Analyze Particles...", "size=1000-Infinity show=Nothing
include add");
			maskname = getTitle();
			numbnuc=roiManager("count");
			roiManager("reset");
```

```
fileArray=newArray(numbnuc);
nucArray=newArray(numbnuc);
areaArray=newArray(numbnuc);
intintArray=newArray(numbnuc);
redArray=newArray(numbnuc);
redintArray=newArray(numbnuc);
redpannucArray=newArray(numbnuc);
greenArray=newArray(numbnuc);
greenintArray=newArray(numbnuc);
greenpannucArray=newArray(numbnuc);
overlapArray=newArray(numbnuc);
for (j=0; j<numbnuc; j++)
{
	fileArray[j]=filename;
	selectWindow(maskname);
	run("Analyze Particles...", "size=1000-Infinity
show=Nothing include add");
	//DAPI measurements
	selectWindow(filename +" (blue)");
	//roiManager("Select", j);
	roiManager("Measure");
	areaArray[j] = getResult("Area");
	intintArray[j] = getResult("IntDen");
	//red int den
	selectWindow(filename +" (red)");
	run("8-bit");
	roiManager("Select", j);
	roiManager("Measure");
	redintArray[j] = getResult("IntDen");
	if (redintArray[j]>70000)
	{
		nPNred = "Pan nuclear";
	}
	if (redintArray[j]<70000)
	{
```

```
        nPNred = "Not pan nuclear";
}
redpannucArray[j] = nPNred;
print("53BP1 staining?");
print(nPNred);
//green int density
selectWindow(filename +" (green)");
run("8-bit");
roiManager("Select", j);
roiManager("Measure");
greenintArray[j] = getResult("IntDen");
if (greenintArray[j]>70000)
{
        nPNgreen = "Pan nuclear";
}
if (greenintArray[j]<70000)
{
        nPNgreen = "Not pan nuclear";
}
greenpannucArray[j] = nPNgreen;
print("green staining?");
print(nPNgreen);
//red foci
selectWindow(filename +" (red)");
run("Select All");
roiManager("Select", j);
run("Copy");
run("Internal Clipboard");
rename("nuc red");
masknameC1 = "nuc red";
//green foci
selectWindow(filename +" (green)");
run("Select All");
roiManager("Select", j);
run("Copy");
```

```
                        run("Internal Clipboard");
                        rename("nuc green");
                        masknameC2 = "nuc green";
                        roiManager("reset");
                        //red foci counts
                        selectWindow(masknameC1);
                        setThreshold(minR, 255);
                        run("Convert to Mask");
                        run("Watershed");
                        run("Analyze Particles...", "size="+minA+"-"+maxA+"
include add");
                        close(masknameC1);
                        currRedCount=roiManager("count");
                        print("redCount="+currRedCount);
                        redArray[j]= currRedCount;
                        //green foci counts
                        selectWindow(masknameC2);
                        setThreshold(minG, 255);
                        run("Convert to Mask");
                        run("Watershed");
                        run("Analyze Particles...", "size="+minA+"-"+maxA+"
include add");
                        close(masknameC2);
                        currGreenCount=(roiManager("count")-currRedCount);
                        print("greenCount="+currGreenCount);
                        greenArray[j]= currGreenCount;
                        //overlap
                        overlapCount=0;
                        for (k=0;k<roiManager('count');k++){
                                for (h=0;h<roiManager('count');h++){
                                        if (k!=h) {

        roiManager('select',newArray(k,h));
                                                roiManager("AND");
                                        }
```

```
                                        if ((k!=h)&(selectionType()>-1)) {
                                                overlapCount=overlapCount+1;
                                        }
                                }
                        }
                        overlapArray[j]=overlapCount/2;
                        //print(overlapCount/2);
                        nucArray[j]=j+1;
                        roiManager("reset");
                        }
        allfile=Array.concat(allfile, fileArray);
        allnuc=Array.concat(allnuc, nucArray);
        allarea=Array.concat(allarea, areaArray);
        allintint=Array.concat(allintint, intintArray);
        allredint=Array.concat(allredint, redintArray);
        allredpannuc=Array.concat(allredpannuc, redpannucArray);
        allred=Array.concat(allred, redArray);
        allgreen=Array.concat(allgreen, greenArray);
        allgreenint=Array.concat(allgreenint, greenintArray);
        allgreenpannuc=Array.concat(allgreenpannuc, greenpannucArray);
        alloverlap=Array.concat(alloverlap, overlapArray);
        run("Close All");
                }
        }
outTab = Table.create("Foci counts");
        Table.setColumn("File", allfile);
        Table.setColumn("Nucleus", allnuc);
        Table.setColumn("Nuc area (pixel2)", allarea);
        Table.setColumn("IntDensity", allintint);
        Table.setColumn("RedIntDensity", allredint);
        Table.setColumn("GreenIntDensity", allgreenint);
        Table.setColumn("RedPanNuc", allredpannuc);
        Table.setColumn("GreenPanNuc", allgreenpannuc);
        Table.setColumn("Red foci", allred);
        Table.setColumn("Green foci", allgreen);
```

```
        Table.setColumn("Overlap foci", alloverlap);
Table.save(mainDir+"/foci_counts.csv");
run("Close All");
print("\ndone");
}
```

# RStudio chapter scripts

RStudio chapter script titles are attached below in addition to the packages, functions and custom theme that were used for the downstream analysis.

## Packages, functions and themes

A list of packages, functions and theme used are detailed below.

### Packages

R version 4.3.0 (2023-04-21 ucrt)

Platform: x86_64-w64-mingw32/x64 (64-bit)

Running under: Windows 11 x64 (build 26100)

Matrix products: default

locale:

[1] LC_COLLATE=English_United Kingdom.utf8

[2] LC_CTYPE=English_United Kingdom.utf8

[3] LC_MONETARY=English_United Kingdom.utf8

[4] LC_NUMERIC=C

[5] LC_TIME=English_United Kingdom.utf8

time zone: Europe/London

tzcode source: internal

attached base packages:

[1] stats4 stats graphics grDevices utils datasets

[7] methods base

other attached packages:

[1] ggrepel_0.9.4

[2] annotatr_1.28.0

[3] FSA_0.9.5

[4] ggsignif_0.6.4

[5] plyr_1.8.9

[6] rlang_1.1.2

[7] plyranges_1.22.0

[8] circlize_0.4.15

[9] DESeq2_1.42.0

[10] SummarizedExperiment_1.32.0

[11] MatrixGenerics_1.14.0

[12] matrixStats_1.0.0

[13] valr_0.7.0

[14] org.Hs.eg.db_3.18.0

[15] biomaRt_2.58.0

[16] biomartr_1.0.6

[17] rstatix_0.7.2

[18] ggpubr_0.6.0

[19] lubridate_1.9.3

[20] forcats_1.0.0

[21] purrr_1.0.2

[22] readr_2.1.4

[23] tidyr_1.3.0

[24] tibble_3.2.1

[25] tidyverse_2.0.0

[26] dplyr_1.1.4

[27] stringr_1.5.1

[28] plotly_4.10.3

[29] data.table_1.14.8

[30] karyoploteR_1.28.0

[31] TxDb.Hsapiens.UCSC.hg38.knownGene_3.18.0

[32] regioneR_1.34.0

[33] ChIPseeker_1.38.0

[34] MotifDb_1.44.0

[35] GenomicFeatures_1.54.1

[36] AnnotationDbi_1.64.1

[37] Biobase_2.60.0

[38] Rsamtools_2.18.0

[39] BSgenome.Hsapiens.UCSC.hg38_1.4.5

[40] BSgenome_1.70.1

[41] rtracklayer_1.62.0

[42] BiocIO_1.12.0

[43] Biostrings_2.70.1

[44] XVector_0.40.0

[45] GenomicRanges_1.52.0

[46] GenomeInfoDb_1.38.1

[47] IRanges_2.34.0

[48] S4Vectors_0.38.1

[49] BiocGenerics_0.48.1

[50] vctrs_0.6.4

[51] pathview_1.42.0

[52] clusterProfiler_4.10.0

[53] amap_0.8-19

[54] reshape2_1.4.4

[55] ggplot2_3.4.4

loaded via a namespace (and not attached):

[1] fs_1.6.3

[2] ProtGenerics_1.34.0

[3] bitops_1.0-7

[4] enrichplot_1.22.0

[5] HDO.db_0.99.1

[6] httr_1.4.7

[7] RColorBrewer_1.1-3

[8] Rgraphviz_2.46.0

[9] tools_4.3.0

[10] backports_1.4.1

[11] utf8_1.2.4

[12] R6_2.5.1

[13] lazyeval_0.2.2

[14] withr_2.5.2

[15] prettyunits_1.2.0

[16] gridExtra_2.3

[17] cli_3.6.1

[18] scatterpie_0.2.1

[19] KEGGgraph_1.62.0

[20] yulab.utils_0.1.0

[21] gson_0.1.0

[22] foreign_0.8-86

[23] DOSE_3.28.1

[24] dichromat_2.0-0.1

[25] plotrix_3.8-4

[26] rstudioapi_0.15.0

[27] RSQLite_2.3.3

[28] shape_1.4.6

[29] generics_0.1.3

[30] gridGraphics_0.5-1

[31] TxDb.Hsapiens.UCSC.hg19.knownGene_3.2.2

[32] gtools_3.9.5

[33] car_3.1-2

[34] GO.db_3.18.0

[35] Matrix_1.6-3

[36] fansi_1.0.5

[37] abind_1.4-5

[38] lifecycle_1.0.4

[39] yaml_2.3.7

[40] carData_3.0-5

[41] gplots_3.1.3

[42] qvalue_2.34.0

[43] SparseArray_1.2.2

[44] BiocFileCache_2.10.1

[45] grid_4.3.0

[46] blob_1.2.4

[47] promises_1.2.1

[48] crayon_1.5.2

[49] lattice_0.22-5

[50] cowplot_1.1.1

[51] KEGGREST_1.42.0

[52] pillar_1.9.0

[53] knitr_1.45

[54] fgsea_1.28.0

[55] rjson_0.2.21

[56] boot_1.3-28.1

[57] codetools_0.2-19

[58] fastmatch_1.1-4

[59] glue_1.6.2

[60] ggfun_0.1.3

[61] png_0.1-8

[62] treeio_1.26.0

[63] gtable_0.3.4

[64] cachem_1.0.8

[65] xfun_0.41

[66] S4Arrays_1.2.0

[67] mime_0.12

[68] tidygraph_1.2.3

[69] interactiveDisplayBase_1.40.0

[70] ellipsis_0.3.2

[71] nlme_3.1-164

[72] ggtree_3.10.0

[73] bit64_4.0.5

[74] progress_1.2.2

[75] filelock_1.0.2

[76] KernSmooth_2.23-22

[77] rpart_4.1.21

[78] splitstackshape_1.4.8

[79] colorspace_2.1-0

[80] DBI_1.1.3

[81] Hmisc_5.1-1

[82] nnet_7.3-19

[83] tidyselect_1.2.0

[84] bit_4.0.5

[85] compiler_4.3.0

[86] curl_5.1.0

[87] graph_1.80.0

[88] htmlTable_2.4.2

[89] bezier_1.1.2

[90] xml2_1.3.5

[91] DelayedArray_0.28.0

[92] shadowtext_0.1.2

[93] checkmate_2.3.0

[94] scales_1.3.0

[95] caTools_1.18.2

[96] rappdirs_0.3.3

[97] digest_0.6.33

[98] rmarkdown_2.25

[99] htmltools_0.5.7

[100] pkgconfig_2.0.3

[101] base64enc_0.1-3

[102] dbplyr_2.4.0

[103] fastmap_1.1.1

[104] ensembldb_2.26.0

[105] GlobalOptions_0.1.2

[106] htmlwidgets_1.6.3

[107] shiny_1.8.0

[108] farver_2.1.1

[109] jsonlite_1.8.7

[110] BiocParallel_1.36.0

[111] GOSemSim_2.28.0

[112] VariantAnnotation_1.48.1

[113] RCurl_1.98-1.12

[114] magrittr_2.0.3

[115] Formula_1.2-5

[116] GenomeInfoDbData_1.2.11

[117] ggplotify_0.1.2

[118] patchwork_1.1.3

[119] munsell_0.5.0

[120] Rcpp_1.0.11

[121] ape_5.7-1

[122] bamsignals_1.34.0

[123] viridis_0.6.4

[124] stringi_1.8.2

[125] ggraph_2.1.0

[126] zlibbioc_1.46.0

[127] MASS_7.3-60

[128] AnnotationHub_3.10.0

[129] parallel_4.3.0

[130] graphlayouts_1.0.2

[131] splines_4.3.0

[132] hms_1.1.3

[133] locfit_1.5-9.8

[134] igraph_1.5.1

[135] BiocVersion_3.18.1

[136] XML_3.99-0.16

[137] evaluate_0.23

[138] biovizBase_1.50.0

[139] BiocManager_1.30.22

[140] tzdb_0.4.0

[141] tweenr_2.0.2

[142] httpuv_1.6.12

[143] polyclip_1.10-6

[144] ggforce_0.4.1

[145] broom_1.0.5

[146] xtable_1.8-4

[147] restfulr_0.0.15

[148] AnnotationFilter_1.26.0

[149] tidytree_0.4.5

[150] later_1.3.1

[151] viridisLite_0.4.2

[152] aplot_0.2.2

[153] memoise_2.0.1

[154] GenomicAlignments_1.38.0

[155] cluster_2.1.5

[156] timechange_0.2.0

[157] EdgeR_4.0.2

[158] car 3.1-3

**Functions**

```
#####Functions####
####saves plot
#make %notin% function
`%notin%` <- Negate(`%in%`)
save_plot= function(plot, plot.path, plot.height, plot.width) {
  png(plot.path,width = plot.width, height = plot.height)
  print(plot)
  dev.off()
}
####read bed
read_bed <- function(filename, n_fields = 3, col_types = bed12_coltypes,
                     sort = TRUE, ...) {
  coltypes <- col_types[1:n_fields]
  colnames <- names(coltypes)

  out <- readr::read_tsv(
    filename,
    col_names = colnames,
    col_types = coltypes, ...
  )

  if (sort) out <- bed_sort(out)

  out
}
####reads bigwig
read_bigwig <- function(path, set_strand = "+") {
  # note that rtracklayer will produce a one-based GRanges object
  res <- rtracklayer::import(path)
  res <- dplyr::as_tibble(res)
  res <- dplyr::mutate(res,
                       chrom = as.character(seqnames),
                       start = start - 1L,
                       strand = set_strand)
```

```
  dplyr::select(res, chrom, start, end, score, strand)}
####reads broadpeak
read_broadpeak <- function(filename, ...) {
  coltypes <- peak_coltypes[1:length(peak_coltypes) - 1]
  colnames <- names(coltypes)
  out <- readr::read_tsv(filename, col_names = colnames, col_types = coltypes)
  out
}
####reads narrowpeak
read_narrowpeak <- function(filename, ...) {
  colnames <- names(peak_coltypes)
  out <- readr::read_tsv(
    filename,
    col_types = peak_coltypes,
    col_names = colnames
  )
  out
}
####peak column names
peak_coltypes <- list(
  chrom = readr::col_character(),
  start = readr::col_integer(),
  end = readr::col_integer(),
  name = readr::col_character(),
  score = readr::col_integer(),
  strand = readr::col_character(),
  signal = readr::col_double(),
  pvalue = readr::col_double(),
  qvalue = readr::col_double(),
  peak = readr::col_integer()
)

bed12_coltypes <- list(
  chrom = readr::col_character(),
  start = readr::col_integer(),
```

```
  end = readr::col_integer(),
  name = readr::col_character(),
  score = readr::col_character(),
  strand = readr::col_character(),
  cds_start = readr::col_integer(),
  cds_end = readr::col_integer(),
  item_rgb = readr::col_character(),
  exon_count = readr::col_integer(),
  exon_sizes = readr::col_character(),
  exon_starts = readr::col_character()
)

#transcripts per million function
tpm3 <- function(counts,len) {

  x <- counts/len

  return(t(t(x)*1e6/colSums(x)))

}
make_volcano = function(de_table, fold_cutoff, p_cutoff, plot_name)
{
  # get the sig genes
  sig = row.names(subset(de_table, p.adj < p_cutoff & abs(log2fold) >
fold_cutoff))
  de_table_sig = de_table[sig,]
  sig_names = de_table_sig[1:20,]

  # make the plot
  ggp = ggplot(de_table ,aes(x=log2fold, y=mlog10p)) +
    geom_point(colour = "black") +
    geom_point(data = de_table_sig, colour ="red") +
    labs(title = plot_name, x = "log2fold change", y = "-log10fold of p value") +
    theme_classic () +
```

```
    geom_vline(xintercept = -1, linetype="dashed", colour = "grey", linewidth =
0.5) +
    geom_vline(xintercept = 1, linetype="dashed", colour = "grey", linewidth = 0.5)
+
    geom_hline(yintercept = -log10(0.05), linetype="dashed", colour = "grey",
linewidth = 0.5) +
    my_theme #+
  # geom_text(data = sig_names,
  #          aes(label = sig_names$symbol))

  # return the plot
  return(ggp)
}
#ggp.volcano = make_volcano(master, 1, 0.05, "R10S vs IR volcano plot")
####Make save plot function
save_plot= function(plot, plot.path, plot.height, plot.width)
{
  png(plot.path,width = plot.width, height = plot.height)
  print(plot)
  dev.off()
}

#to use function
#save_plot(ggp.volcano, "Signatures/data_d1/volcano",500,600)
####makes ma plot
make_ma = function(de, p_cutoff, fold_cutoff)
{
  # get the sig genes
  sig = row.names(subset(de, p.adj < p_cutoff & abs(log2fold) > fold_cutoff))
  de_table_sig = de[sig,]

  #make MAplot
  MAplot= ggplot(de, aes(x=log10(mean), y=log2fold)) +
    geom_point(size = 2, colour="black") +
    geom_point (data = de_table_sig, colour ="red", alpha = 0.6) +
```

```
    labs(title="MA plot", x="Mean expresssion", y="log2fold change") +
    geom_hline(yintercept = -0.5, linetype="dashed",color="grey",size=0.5) +
    geom_hline(yintercept = 0.5, linetype="dashed",color="grey",size=0.5) +
    geom_vline(xintercept = 0.0, linetype="dashed",color="grey",size=0.5) +
    my_theme

  # return the plot
  return(MAplot)
}

#creates MA plot
#ggp.MA = make_ma(master, 0.05, 1)

#save plot
#save_plot(ggp.MA, "Signatures/data_d1/MA_plot",500,600)

#### PCA

####makes PCA
make_pca = function(em, sample_groups, plot_name)
{
  # do the PCA
  xx = prcomp(t(em_scaled))
  pca_coordinates = data.frame(xx$x)

  # get % variation
  vars = apply(xx$x, 2, var)
  prop_x = round(vars["PC1"] / sum(vars),4) * 100
  prop_y = round(vars["PC2"] / sum(vars),4) * 100

  x_axis_label = paste("PC1 ", " (",prop_x,"%)",sep="")
  y_axis_label = paste("PC2 ", " (",prop_y,"%)",sep="")

  # plot
```

```
  gg.pca = ggplot(pca_coordinates, aes(x= PC1, y= PC2, colour=
ss$SAMPLE_GROUP)) +
    geom_point (aes(label=ss$SAMPLE_GROUP), size = 3) +
    ##add in nice extra bits
    scale_colour_manual(values=c("green", "purple","blue")) +
    geom_text(aes(label=ss$SAMPLE),position= position_nudge(y = 2), colour =
"black") +
    ##uses the labels we made above
    labs(title = plot_name, x= x_axis_label, y= y_axis_label) +
    my_theme +
    theme(axis.text.y = element_blank(),axis.ticks = element_blank(),
legend.title = element_blank(),
          legend.spacing.x = unit(0.25,'cm'))

  return(gg.pca)
}

#gg.pca = make_pca(em,ss)
#save_plot(gg.pca, "Signatures/data_d1/pca_plot",500,600)

####make heatmap

#em_sorted_order = order(master[,"p.adj"],decreasing = FALSE)
#master = master[em_sorted_order,]
##sorts in the order requested
#em_scaled_sig = em_scaled_sig[em_sorted_order,]
##pull out top 100
#top_em_scaled_sig = em_scaled_sig[1:100,]

#makes function without gene list
make_heatmap = function(em_symbols)
{

  hm.matrix = as.matrix(em_symbols)
  y.dist = Dist(hm.matrix,method="spearman")
```

```
  y.cluster = hclust(y.dist, method="average")
  y.did =as.dendrogram(y.cluster)
  y.did.reorder = reorder(y.did,0,FUN="average")
  y.order = order.dendrogram(y.did.reorder)
  hm.matrix_clustered = hm.matrix[y.order,]

  #makes the colour palette
  #can add more colours but three is about right
  colours = c("blue", "lightyellow", "red")
  palette = colorRampPalette(colours)(100) ##number of colours in palette

  hm.matrix_clustered = melt(hm.matrix_clustered)
  gg.heatmap = ggplot(hm.matrix_clustered,aes(x = Var2, y = Var1, fill = value))
+
    geom_tile () +
    ##adds the desired colours
    scale_fill_gradientn(colours = palette) +

    ##adds titles
    ylab("gene names") +
    xlab("site of gene") +
    labs(title = "top significant DSB sites") +
    my_theme +
    theme(axis.ticks = element_blank(), legend.title = element_blank(),
          legend.spacing.x = unit(0.25,'cm')) +
    theme(axis.text.x = element_text(angle = 45, hjust = 1))

  return(gg.heatmap)
}
#run function and save heatmap
#gg.heatmap = make_heatmap(top_em_scaled_sig)
#save_plot(gg.heatmap, "Signatures/data_d1/gg.heatmap",500,600)


#### Function that gets the default gg plot colours
```

```
#gg_color_hue = function(n) {
# hues = seq(15, 375, length = n + 1)
#  hcl(h = hues, l = 65, c = 100)[1:n]
#}
# uses the function to get the colours
#number_of_groups = 3
#colours = gg_color_hue(number_of_groups)
## we can change the order of the colours, if needed, e.g.
#colours = c(colours[3],colours[2], colours[1])
####Heatmap rug function
make_heatmap_rug = function(groups, colours)
{
 #gives the generic colours for the variables like in the PCA (blue, green and
red)
 hm.palette = colorRampPalette(colours)(3)

 #makes plot
 heatrug.plot = ggplot(groups_data, aes(x = Var1, y = Var2, fill = value)) +
  geom_tile(linetype="blank") +
  scale_fill_gradientn(colours = hm.palette) +
  #removes labels
  labs(x = "", y = "") +
  #removes ticks and legend
  theme(legend.position="none", legend.title = element_blank(),
       axis.text.x = element_blank(), axis.text.y = element_blank(),
axis.ticks=element_blank())

 #print plot
 return(heatrug.plot)
}

#heatrug.plot = make_heatmap_rug(groups_data, colours)
#save_plot(heatrug.plot,
"Signatures/data_d1/figures_day/heatrug.plot.png",100,600)
```

```
####multigene boxplot
make_multi_gene_boxplot = function(scaled_gene_data, desired_gene_list,
sample_sheet, plot_name)
{
  gene_data = scaled_gene_data[desired_gene_list,]
  gene_data = data.frame(t(gene_data))
  gene_data$sample_group = sample_sheet$SAMPLE_GROUP

  ##need to melt the multiple columns to allow simultaneous processing
  ##can also sort by sample group
  gene_data.m = melt(gene_data,id.vars="sample_group")

  ##renames the column names to allow easier sorting
  names(gene_data.m) = c("sample_group", "gene_name", "expression")

  ##allows identity via levels in each gene (gut, duct, node)
  gene_data$sample_group = factor(gene_data$sample_group, levels = c("no_IR",
"IR"))
  levels(gene_data$"sample_group")

  ##makes the boxplot
  gg.multi_box = ggplot(gene_data.m,aes(x=gene_name,y=expression,
fill=sample_group, )) +
    geom_boxplot (alpha = 1, colour = "black") +
    labs(title = plot_name, x= "gene", y= "expression") +
    my_theme +
    ##adjusts the x axis names to fit them in
    theme(axis.text.x = element_text(angle = 45, hjust = 1))
  gg.multi_box
}
####get pathway analysis function
get_pathways = function(gene_list, organism_db)
{
  sig_genes_entrez = bitr(gene_list, fromType = "SYMBOL", toType =
c("ENTREZID"),OrgDb = organism_db)
```

```
  go_enrich = enrichGO(gene = sig_genes_entrez$ENTREZID, OrgDb =
organism_db, readable = T, ont = "BP",
                       pvalueCutoff = 0.05, qvalueCutoff = 0.10)

  pathways = go_enrich$Description
  pathways_list_signature = as.data.frame(pathways)

  return(pathways_list_signature)
}
#get_pathways(signature_1, org.Mm.eg.db)
library(tidyverse)
```

**Theme**

```
my_theme = theme(
  plot.title = element_text(size=35, face="bold", family="Calibri", colour =
"darkslategrey"),
  axis.text.x = element_text(size=25, colour = "azure4", family="Calibri"),
  axis.text.y = element_text(size=25, colour = "azure4", family="Calibri"),
  axis.title.x = element_text(size=30, face="italic", colour = "azure4",
family="Calibri"),
  axis.title.y = element_text(size=30, face="italic", colour = "azure4",
family="Calibri"),
  panel.background = element_rect(colour = "azure3", fill = "aliceblue"),
  element_line(colour = "azure3"),
  legend.box.background = element_rect(colour = "aliceblue", fill = "aliceblue",
linetype = "blank"),
  legend.text = element_text(size = 25, colour = "azure4", family="Calibri"),
  legend.title = element_text(size =25, colour ="darkslategrey", family="Calibri"),
  legend.key = element_rect(fill = "white"),
  panel.grid.major = element_line(colour ="azure2"),
  panel.grid.minor = element_line(colour ="azure1"),
  strip.text.x = element_text(size = 20, color = "darkslategrey", face =
"bold.italic"),
  strip.text.y = element_text(size = 20, color = "darkslategrey", face =
"bold.italic")
```

)

## RStudio script chapters

The following commandline and RStudio scripts chapters are available:

- Chapter 2: ch_2_CMDLNE_bigwig_hg38_liftover_cmdline.txt
- Chapter 3: ch_3_thesis_script_dsb_overview
    - ch_3.1_CMDLNE_bedtools_intersect_50kb_cmdline
- Chapter 4: ch_4_thesis_script_genes_genelength
- Chapter 5: ch_5_thesis_script_gene_bodies_annotated_genic_regions
- Chapter 6: ch_6_thesis_script_transcription_euchr_differential
    - ch_6.1_CMDLINE_bedtools_intersect_foldchange_ATACpeaks
- Chapter 7: ch_7_thesis_script_IR_BLISS_INDUCEseq

# List of References

AH-PINE, F., CASAS, D., MENEI, P., BOISSELIER, B., GARCION, E. & ROUSSEAU, A. 2021. RNA-sequencing of IDH-wild-type glioblastoma with chromothripsis identifies novel gene fusions with potential oncogenic properties. *Transl Oncol,* 14**,** 100884.

AHMED, S. U., CARRUTHERS, R., GILMOUR, L., YILDIRIM, S., WATTS, C. & CHALMERS, A. J. 2015. Selective Inhibition of Parallel DNA Damage Response Pathways Optimizes Radiosensitization of Glioblastoma Stem-like Cells. *Cancer Research,* 75**,** 4416.

ALEKSANDROV, R., HRISTOVA, R., STOYNOV, S. & GOSPODINOV, A. 2020. The Chromatin Response to Double-Strand DNA Breaks and Their Repair. *Cells,* 9.

ALLEN, C. P., HIRAKAWA, H., NAKAJIMA, N. I., MOORE, S., NIE, J., SHARMA, N., SUGIURA, M., HOKI, Y., ARAKI, R., ABE, M., OKAYASU, R., FUJIMORI, A. & NICKOLOFF, J. A. 2017. Low- and High-LET Ionizing Radiation Induces Delayed Homologous Recombination that Persists for Two Weeks before Resolving. *Radiat Res,* 188**,** 82-93.

AMEMIYA, H. M., KUNDAJE, A. & BOYLE, A. P. 2019. The ENCODE Blacklist: Identification of Problematic Regions of the Genome. *Sci Rep,* 9**,** 9354.

ANDRES, S. N., MODESTI, M., TSAI, C. J., CHU, G. & JUNOP, M. S. 2007. Crystal structure of human XLF: a twist in nonhomologous DNA end-joining. *Mol Cell,* 28**,** 1093-101.

ARNAUDEAU, C., LUNDIN, C. & HELLEDAY, T. 2001. DNA double-strand breaks associated with replication forks are predominantly repaired by homologous recombination involving an exchange mechanism in mammalian cells. *J Mol Biol,* 307**,** 1235-45.

ASHBURNER, M., BALL, C. A., BLAKE, J. A., BOTSTEIN, D., BUTLER, H., CHERRY, J. M., DAVIS, A. P., DOLINSKI, K., DWIGHT, S. S., EPPIG, J. T., HARRIS, M. A., HILL, D. P., ISSEL-TARVER, L., KASARSKIS, A., LEWIS, S., MATESE, J. C., RICHARDSON, J. E., RINGWALD, M., RUBIN, G. M. & SHERLOCK, G. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet,* 25**,** 25-9.

AVERBECK, N. B., RINGEL, O., HERRLITZ, M., JAKOB, B., DURANTE, M. & TAUCHER-SCHOLZ, G. 2014. DNA end resection is needed for the repair of complex lesions in G1-phase human cells. *Cell cycle (Georgetown, Tex.),* 13**,** 2509-2516.

AYMARD, F., AGUIRREBENGOA, M., GUILLOU, E., JAVIERRE, B. M., BUGLER, B., ARNOULD, C., ROCHER, V., IACOVONI, J. S., BIERNACKA, A., SKRZYPCZAK, M., GINALSKI, K., ROWICKA, M., FRASER, P. & LEGUBE, G. 2017. Genome-wide mapping of long-range contacts unveils clustering of DNA double-strand breaks at damaged active genes. *Nat Struct Mol Biol,* 24**,** 353-361.

AYMARD, F., BUGLER, B., SCHMIDT, C. K., GUILLOU, E., CARON, P., BRIOIS, S., IACOVONI, J. S., DABURON, V., MILLER, K. M., JACKSON, S. P. & LEGUBE, G. 2014. Transcriptionally active chromatin recruits homologous recombination at DNA double-strand breaks. *Nature Structural & Molecular Biology,* 21**,** 366-374.

AZVOLINSKY, A., GIRESI, P. G., LIEB, J. D. & ZAKIAN, V. A. 2009. Highly transcribed RNA polymerase II genes are impediments to replication fork progression in Saccharomyces cerevisiae. *Mol Cell,* 34**,** 722-34.

BALIC, M., LIN, H., YOUNG, L., HAWES, D., GIULIANO, A., MCNAMARA, G., DATAR, R. H. & COTE, R. J. 2006. Most early disseminated cancer cells

detected in bone marrow of breast cancer patients have a putative breast cancer stem cell phenotype. *Clin Cancer Res,* 12**,** 5615-21.

BALLARINO, R., BOUWMAN, B. A. M., AGOSTINI, F., HARBERS, L., DIEKMANN, C., WERNERSSON, E., BIENKO, M. & CROSETTO, N. 2022. An atlas of endogenous DNA double-strand breaks arising during human neural cell fate determination. *Sci Data,* 9**,** 400.

BALLINGER, T. J., BOUWMAN, B. A. M., MIRZAZADEH, R., GARNERONE, S., CROSETTO, N. & SEMPLE, C. A. 2019. Modeling double strand break susceptibility to interrogate structural variation in cancer. *Genome Biology,* 20**,** 28.

BALZANO, E., DI TOMMASO, E., ANTOCCIA, A., PELLICCIA, F. & GIUNTA, S. 2021. Characterization of Chromosomal Instability in Glioblastoma. *Front Genet,* 12**,** 810793.

BANÁTH, J. P., MACPHAIL, S. H. & OLIVE, P. L. 2004. Radiation Sensitivity, H2AX Phosphorylation, and Kinetics of Repair of DNA Strand Breaks in Irradiated Cervical Cancer Cell Lines. *Cancer Research,* 64**,** 7144-7149.

BAO, S., WU, Q., LI, Z., SATHORNSUMETEE, S., WANG, H., MCLENDON, R. E., HJELMELAND, A. B. & RICH, J. N. 2008. Targeting cancer stem cells through L1CAM suppresses glioma growth. *Cancer Res,* 68**,** 6043-8.

BAO, S., WU, Q., MCLENDON, R. E., HAO, Y., SHI, Q., HJELMELAND, A. B., DEWHIRST, M. W., BIGNER, D. D. & RICH, J. N. 2006. Glioma stem cells promote radioresistance by preferential activation of the DNA damage response. *Nature,* 444**,** 756-60.

BAO, Y. & SHEN, X. 2007. Chromatin remodeling in DNA double-strand break repair. *Current Opinion in Genetics & Development,* 17**,** 126-131.

BARNETT, D. W., GARRISON, E. K., QUINLAN, A. R., STRÖMBERG, M. P. & MARTH, G. T. 2011. BamTools: a C++ API and toolkit for analyzing and managing BAM files. *Bioinformatics,* 27**,** 1691-2.

BARTKOVA, J., HAMERLIK, P., STOCKHAUSEN, M. T., EHRMANN, J., HLOBILKOVA, A., LAURSEN, H., KALITA, O., KOLAR, Z., POULSEN, H. S., BROHOLM, H., LUKAS, J. & BARTEK, J. 2010. Replication stress and oxidative damage contribute to aberrant constitutive activation of DNA damage signalling in human gliomas. *Oncogene,* 29**,** 5095-102.

BECK, S., BERNSTEIN, B. E., CAMPBELL, R. M., COSTELLO, J. F., DHANAK, D., ECKER, J. R., GREALLY, J. M., ISSA, J. P., LAIRD, P. W., POLYAK, K., TYCKO, B. & JONES, P. A. 2012. A blueprint for an international cancer epigenome consortium. A report from the AACR Cancer Epigenome Task Force. *Cancer Res,* 72**,** 6319-24.

BEIER, D., HAU, P., PROESCHOLDT, M., LOHMEIER, A., WISCHHUSEN, J., OEFNER, P. J., AIGNER, L., BRAWANSKI, A., BOGDAHN, U. & BEIER, C. P. 2007. CD133(+) and CD133(-) glioblastoma-derived cancer stem cells show differential growth characteristics and molecular profiles. *Cancer Res,* 67**,** 4010-5.

BELYAEV, I. Y. 2010. Radiation-induced DNA repair foci: spatio-temporal aspects of formation, application for assessment of radiosensitivity and biological dosimetry. *Mutat Res,* 704**,** 132-41.

BERGER, J. M. 1998. Type II DNA topoisomerases. *Current Opinion in Structural Biology,* 8**,** 26-32.

BERGER, J. M. & WANG, J. C. 1996. Recent developments in DNA topoisomerase II structure and mechanism. *Current Opinion in Structural Biology,* 6**,** 84-90.

BERMEJO, R., LAI, M. S. & FOIANI, M. 2012. Preventing replication stress to maintain genome stability: resolving conflicts between replication and transcription. *Mol Cell,* 45**,** 710-8.

BERMUDEZ, V. P., LINDSEY-BOLTZ, L. A., CESARE, A. J., MANIWA, Y., GRIFFITH, J. D., HURWITZ, J. & SANCAR, A. 2003. Loading of the human 9-1-1 checkpoint complex onto DNA by the checkpoint clamp loader hRad17-replication factor C complex in vitro. *Proc Natl Acad Sci U S A,* 100**,** 1633-8.

BHARGAVA, R., ONYANGO, D. O. & STARK, J. M. 2016. Regulation of Single-Strand Annealing and its Role in Genome Maintenance. *Trends in Genetics,* 32**,** 566-575.

BIERNACKA, A., SKRZYPCZAK, M., ZHU, Y., PASERO, P., ROWICKA, M. & GINALSKI, K. 2021. High-resolution, ultrasensitive and quantitative DNA double-strand break labeling in eukaryotic cells using i-BLESS. *Nature Protocols,* 16**,** 1034-1061.

BIERNACKA, A., ZHU, Y., SKRZYPCZAK, M., FOREY, R., PARDO, B., GRZELAK, M., NDE, J., MITRA, A., KUDLICKI, A., CROSETTO, N., PASERO, P., ROWICKA, M. & GINALSKI, K. 2018. i-BLESS is an ultra-sensitive method for detection of DNA double-strand breaks. *Commun Biol,* 1**,** 181.

BIFFI, G., TANNAHILL, D., MILLER, J., HOWAT, W. J. & BALASUBRAMANIAN, S. 2014. Elevated levels of G-quadruplex formation in human stomach and liver cancer tissues. *PLoS One,* 9**,** e102711.

BIRCH, J. L., STRATHDEE, K., GILMOUR, L., VALLATOS, A., MCDONALD, L., KOUZELI, A., VASAN, R., QAISI, A. H., CROFT, D. R., CRIGHTON, D., GILL, K., GRAY, C. H., KONCZAL, J., MEZNA, M., MCARTHUR, D., SCHÜTTELKOPF, A. W., MCCONNELL, P., SIME, M., HOLMES, W. M., BOWER, J., MCKINNON, H. J., DRYSDALE, M., OLSON, M. F. & CHALMERS, A. J. 2018. A Novel Small-Molecule Inhibitor of MRCK Prevents Radiation-Driven Invasion in Glioblastoma. *Cancer Res,* 78**,** 6509-6522.

BLAISDELL, J. O., HARRISON, L. & WALLACE, S. S. 2001. Base excision repair processing of radiation-induced clustered DNA lesions. *Radiat Prot Dosimetry,* 97**,** 25-31.

BOEMO, M. A. 2021. DNAscent v2: detecting replication forks in nanopore sequencing data with deep learning. *BMC Genomics,* 22**,** 430.

BOQUE-SASTRE, R., SOLER, M., OLIVEIRA-MATEOS, C., PORTELA, A., MOUTINHO, C., SAYOLS, S., VILLANUEVA, A., ESTELLER, M. & GUIL, S. 2015. Head-to-head antisense transcription and R-loop formation promotes transcriptional activation. *Proceedings of the National Academy of Sciences,* 112**,** 5785.

BORREGO-SOTO, G., ORTIZ-LÓPEZ, R. & ROJAS-MARTÍNEZ, A. 2015. Ionizing radiation-induced DNA injury and damage detection in patients with breast cancer. *Genet Mol Biol,* 38**,** 420-32.

BOTEVA, L., NOZAWA, R.-S., NAUGHTON, C., SAMEJIMA, K., EARNSHAW, W. C. & GILBERT, N. 2020. Common Fragile Sites Are Characterized by Faulty Condensin Loading after Replication Stress. *Cell Reports,* 32.

BOTHMER, A., ROBBIANI, D. F., FELDHAHN, N., GAZUMYAN, A., NUSSENZWEIG, A. & NUSSENZWEIG, M. C. 2010. 53BP1 regulates DNA resection and the choice between classical and alternative end joining during class switch recombination. *J Exp Med,* 207**,** 855-65.

BOUGEARD, G., RENAUX-PETEL, M., FLAMAN, J. M., CHARBONNIER, C., FERMEY, P., BELOTTI, M., GAUTHIER-VILLARS, M., STOPPA-LYONNET, D., CONSOLINO, E., BRUGIÈRES, L., CARON, O., BENUSIGLIO, P. R., BRESSAC-

DE PAILLERETS, B., BONADONA, V., BONAÏTI-PELLIÉ, C., TINAT, J., BAERT-DESURMONT, S. & FREBOURG, T. 2015. Revisiting Li-Fraumeni Syndrome From TP53 Mutation Carriers. *J Clin Oncol,* 33**,** 2345-52.

BOUWMAN, B. A. M., AGOSTINI, F., GARNERONE, S., PETROSINO, G., GOTHE, H. J., SAYOLS, S., MOOR, A. E., ITZKOVITZ, S., BIENKO, M., ROUKOS, V. & CROSETTO, N. 2020. Genome-wide detection of DNA double-strand breaks by in-suspension BLISS. *Nat Protoc,* 15**,** 3894-3941.

BOUWMAN, B. A. M. & CROSETTO, N. 2018. Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling. *Genes (Basel),* 9.

BRAMBILLA, F., GARCIA-MANTEIGA, J. M., MONTELEONE, E., HOELZEN, L., ZOCCHI, A., AGRESTI, A. & BIANCHI, M. E. 2020. Nucleosomes effectively shield DNA from radiation damage in living cells. *Nucleic Acids Res,* 48**,** 8993-9006.

BRANDSMA, I. & GENT, D. C. 2012. Pathway choice in DNA double strand break repair: observations of a balancing act. *Genome Integr,* 3**,** 9.

BRESCIA, P., RICHICHI, C. & PELICCI, G. 2012. Current strategies for identification of glioma stem cells: adequate or unsatisfactory? *J Oncol,* 2012**,** 376894.

BRILL, S. J., DINARDO, S., VOELKEL-MEIMAN, K. & STERNGLANZ, R. 1987. Need for DNA topoisomerase activity as a swivel for DNA replication for transcription of ribosomal RNA. *Nature,* 326**,** 414-6.

BRODERICK, L., YOST, S., LI, D., MCGEOUGH, M. D., BOOSHEHRI, L. M., GUADERRAMA, M., BRYDGES, S. D., KUCHAROVA, K., PATEL, N. C., HARR, M., HAKONARSON, H., ZACKAI, E., COWELL, I. G., AUSTIN, C. A., HÜGLE, B., GEBAUER, C., ZHANG, J., XU, X., WANG, J., CROKER, B. A., FRAZER, K. A., PUTNAM, C. D. & HOFFMAN, H. M. 2019. Mutations in topoisomerase IIß result in a B cell immunodeficiency. *Nature Communications,* 10**,** 3644.

BROWN, E. J. & BALTIMORE, D. 2003. Essential and dispensable roles of ATR in cell cycle arrest and genome maintenance. *Genes Dev,* 17**,** 615-28.

BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C., CHANG, H. Y. & GREENLEAF, W. J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods,* 10**,** 1213-8.

BULSTRODE, H., JOHNSTONE, E., MARQUES-TORREJON, M. A., FERGUSON, K. M., BRESSAN, R. B., BLIN, C., GRANT, V., GOGOLOK, S., GANGOSO, E., GAGRICA, S., ENDER, C., FOTAKI, V., SPROUL, D., BERTONE, P. & POLLARD, S. M. 2017. Elevated FOXG1 and SOX2 in glioblastoma enforces neural stem cell identity through transcriptional control of cell cycle and epigenetic regulators. *Genes Dev,* 31**,** 757-773.

BUNCH, H., LAWNEY, B. P., LIN, Y. F., ASAITHAMBY, A., MURSHID, A., WANG, Y. E., CHEN, B. P. & CALDERWOOD, S. K. 2015. Transcriptional elongation requires DNA break-induced signalling. *Nat Commun,* 6**,** 10191.

BURMA, S., CHEN, B. P., MURPHY, M., KURIMASA, A. & CHEN, D. J. 2001. ATM phosphorylates histone H2AX in response to DNA double-strand breaks. *J Biol Chem,* 276**,** 42462-7.

CADORET, J. C., MEISCH, F., HASSAN-ZADEH, V., LUYTEN, I., GUILLET, C., DURET, L., QUESNEVILLE, H. & PRIOLEAU, M. N. 2008. Genome-wide studies highlight indirect links between human replication origins and gene regulation. *Proc Natl Acad Sci U S A,* 105**,** 15837-42.

CAI, T., LIU, Y. & XIAO, J. 2018. Long noncoding RNA MALAT1 knockdown reverses chemoresistance to temozolomide via promoting microRNA-101 in glioblastoma. *Cancer Med,* 7**,** 1404-1415.

CAI, Z., VALLIS, K. A. & REILLY, R. M. 2009. Computational analysis of the number, area and density of gamma-H2AX foci in breast cancer cells exposed to (111)In-DTPA-hEGF or gamma-rays using Image-J software. *Int J Radiat Biol,* 85**,** 262-71.

CAMARILLO, R., JIMENO, S. & HUERTAS, P. 2021. The Effect of Atypical Nucleic Acids Structures in DNA Double Strand Break Repair: A Tale of R-loops and G-Quadruplexes. *Front Genet,* 12**,** 742434.

CANELA, A., SRIDHARAN, S., SCIASCIA, N., TUBBS, A., MELTZER, P., SLECKMAN, B. P. & NUSSENZWEIG, A. 2016. DNA Breaks and End Resection Measured Genome-wide by End Sequencing. *Mol Cell,* 63**,** 898-911.

CARRUTHERS, R. D., AHMED, S. U., RAMACHANDRAN, S., STRATHDEE, K., KURIAN, K. M., HEDLEY, A., GOMEZ-ROMAN, N., KALNA, G., NEILSON, M., GILMOUR, L., STEVENSON, K. H., HAMMOND, E. M. & CHALMERS, A. J. 2018. Replication Stress Drives Constitutive Activation of the DNA Damage Response and Radioresistance in Glioblastoma Stem-like Cells. *Cancer Res,* 78**,** 5060-5071.

CAVALCANTE, R. G. & SARTOR, M. A. 2017. Annotatr: Genomic regions in context. *Bioinformatics,* 33**,** 2381-2383.

CERAMI, E., GAO, J., DOGRUSOZ, U., GROSS, B. E., SUMER, S. O., AKSOY, B. A., JACOBSEN, A., BYRNE, C. J., HEUER, M. L., LARSSON, E., ANTIPIN, Y., REVA, B., GOLDBERG, A. P., SANDER, C. & SCHULTZ, N. 2012. The cBio cancer genomics portal: an open platform for exploring multidimensional cancer genomics data. *Cancer Discov,* 2**,** 401-4.

CERVENA, K., VODENKOVA, S. & VYMETALKOVA, V. 2022. MALAT1 in colorectal cancer: Its implication as a diagnostic, prognostic, and predictive biomarker. *Gene,* 843**,** 146791.

CHADWICK, K. H. & LEENHOUTS, H. P. 1973. A molecular theory of cell survival. *Physics in Medicine and Biology,* 18**,** 78-87.

CHAMBERS, V. S., MARSICO, G., BOUTELL, J. M., DI ANTONIO, M., SMITH, G. P. & BALASUBRAMANIAN, S. 2015. High-throughput sequencing of DNA G-quadruplex structures in the human genome. *Nat Biotechnol,* 33**,** 877-81.

CHAUDHURI, J. & ALT, F. W. 2004. Class-switch recombination: interplay of transcription, DNA deamination and DNA repair. *Nat Rev Immunol,* 4**,** 541-52.

CHAURASIA, P., SEN, R., PANDITA, T. K. & BHAUMIK, S. R. 2012. Preferential repair of DNA double-strand break at the active gene in vivo. *J Biol Chem,* 287**,** 36414-22.

CHAWLA, A., NAGY, C. & TURECKI, G. 2021. Chromatin Profiling Techniques: Exploring the Chromatin Environment and Its Contributions to Complex Traits. *Int J Mol Sci,* 22.

CHEN, H. M., NIKOLIC, A., SINGHAL, D. & GALLO, M. 2022. Roles of Chromatin Remodelling and Molecular Heterogeneity in Therapy Resistance in Glioblastoma. *Cancers (Basel),* 14.

CHEN, J. Y., ZHANG, X., FU, X. D. & CHEN, L. 2019a. R-ChIP for genome-wide mapping of R-loops by using catalytically inactive RNASEH1. *Nat Protoc,* 14**,** 1661-1685.

CHEN, W., XU, X. K., LI, J. L., KONG, K. K., LI, H., CHEN, C., HE, J., WANG, F., LI, P., GE, X. S. & LI, F. C. 2017. MALAT1 is a prognostic factor in glioblastoma multiforme and induces chemoresistance to temozolomide

through suppressing miR-203 and promoting thymidylate synthase expression. *Oncotarget,* 8**,** 22783-22799.

CHEN, Y. H., JONES, M. J., YIN, Y., CRIST, S. B., COLNAGHI, L., SIMS, R. J., 3RD, ROTHENBERG, E., JALLEPALLI, P. V. & HUANG, T. T. 2015. ATR-mediated phosphorylation of FANCI regulates dormant origin firing in response to replication stress. *Mol Cell,* 58**,** 323-38.

CHEN, Y. H., KEEGAN, S., KAHLI, M., TONZI, P., FENYÖ, D., HUANG, T. T. & SMITH, D. J. 2019b. Transcription shapes DNA replication initiation and termination in human cells. *Nat Struct Mol Biol,* 26**,** 67-77.

CHIARLE, R., ZHANG, Y., FROCK, RICHARD L., LEWIS, SUSANNA M., MOLINIE, B., HO, Y.-J., MYERS, DARIENNE R., CHOI, VIVIAN W., COMPAGNO, M., MALKIN, DANIEL J., NEUBERG, D., MONTI, S., GIALLOURAKIS, COSMAS C., GOSTISSA, M. & ALT, FREDERICK W. 2011. Genome-wide Translocation Sequencing Reveals Mechanisms of Chromosome Breaks and Rearrangements in B Cells. *Cell,* 147**,** 107-119.

CHIOLO, I., MINODA, A., COLMENARES, S. U., POLYZOS, A., COSTES, S. V. & KARPEN, G. H. 2011. Double-strand breaks in heterochromatin move outside of a dynamic HP1a domain to complete recombinational repair. *Cell,* 144**,** 732-744.

CHOU, D. M., ADAMSON, B., DEPHOURE, N. E., TAN, X., NOTTKE, A. C., HUROV, K. E., GYGI, S. P., COLAIÁCOVO, M. P. & ELLEDGE, S. J. 2010. A chromatin localization screen reveals poly (ADP ribose)-regulated recruitment of the repressive polycomb and NuRD complexes to sites of DNA damage. *Proc Natl Acad Sci U S A,* 107**,** 18475-80.

CHUNG, W. H., ZHU, Z., PAPUSHA, A., MALKOVA, A. & IRA, G. 2010. Defective resection at DNA double-strand breaks leads to de novo telomere formation and enhances gene targeting. *PLoS Genet,* 6**,** e1000948.

COHEN, A., SATO, M., ALDAPE, K., MASON, C. C., ALFARO-MUNOZ, K., HEATHCOCK, L., SOUTH, S. T., ABEGGLEN, L. M., SCHIFFMAN, J. D. & COLMAN, H. 2015. DNA copy number analysis of Grade II–III and Grade IV gliomas reveals differences in molecular ontogeny including chromothripsis associated with IDH mutation status. *Acta Neuropathologica Communications,* 3**,** 34.

CORTÉS-CIRIANO, I., LEE, J. J., XI, R., JAIN, D., JUNG, Y. L., YANG, L., GORDENIN, D., KLIMCZAK, L. J., ZHANG, C. Z., PELLMAN, D. S. & PARK, P. J. 2020. Comprehensive analysis of chromothripsis in 2,658 human cancers using whole-genome sequencing. *Nat Genet,* 52**,** 331-341.

CROSETTO, N., MITRA, A., SILVA, M. J., BIENKO, M., DOJER, N., WANG, Q., KARACA, E., CHIARLE, R., SKRZYPCZAK, M., GINALSKI, K., PASERO, P., ROWICKA, M. & DIKIC, I. 2013. Nucleotide-resolution DNA double-strand break mapping by next-generation sequencing. *Nat Methods,* 10**,** 361-5.

CRUZ DA SILVA, E., MERCIER, M. C., ETIENNE-SELLOUM, N., DONTENWILL, M. & CHOULIER, L. 2021. A Systematic Review of Glioblastoma-Targeted Therapies in Phases II, III, IV Clinical Trials. *Cancers (Basel),* 13.

CUI, X. & MEEK, K. 2007. Linking double-stranded DNA breaks to the recombination activating gene complex directs repair to the nonhomologous end-joining pathway. *Proceedings of the National Academy of Sciences,* 104**,** 17046-17051.

CUVERTINO, S., STUART, H. M., CHANDLER, K. E., ROBERTS, N. A., ARMSTRONG, R., BERNARDINI, L., BHASKAR, S., CALLEWAERT, B., CLAYTON-SMITH, J., DAVALILLO, C. H., DESHPANDE, C., DEVRIENDT, K., DIGILIO, M. C., DIXIT, A., EDWARDS, M., FRIEDMAN, J. M., GONZALEZ-MENESES, A., JOSS, S.,

KERR, B., LAMPE, A. K., LANGLOIS, S., LENNON, R., LOGET, P., MA, D. Y. T., MCGOWAN, R., DES MEDT, M., O'SULLIVAN, J., ODENT, S., PARKER, M. J., PEBREL-RICHARD, C., PETIT, F., STARK, Z., STOCKLER-IPSIROGLU, S., TINSCHERT, S., VASUDEVAN, P., VILLA, O., WHITE, S. M., ZAHIR, F. R., WOOLF, A. S. & BANKA, S. 2017. ACTB Loss-of-Function Mutations Result in a Pleiotropic Developmental Disorder. *Am J Hum Genet,* 101**,** 1021-1033.

DALEY, J. M. & SUNG, P. 2014. 53BP1, BRCA1, and the choice between recombination and end joining at DNA double-strand breaks. *Mol Cell Biol,* 34**,** 1380-8.

DARLING, J. L. & THOMAS, D. G. 2001. Response of short-term cultures derived from human malignant glioma to aziridinylbenzoquinone, etoposide and doxorubicin: an in vitro phase II trial. *Anticancer Drugs,* 12**,** 753-60.

DE ALMEIDA SASSI, F., LUNARDI BRUNETTO, A., SCHWARTSMANN, G., ROESLER, R. & ABUJAMRA, A. L. 2012. Glioma revisited: from neurogenesis and cancer stem cells to the epigenetic regulation of the niche. *J Oncol,* 2012**,** 537861.

DITULLIO, R. A., JR., MOCHAN, T. A., VENERE, M., BARTKOVA, J., SEHESTED, M., BARTEK, J. & HALAZONETIS, T. D. 2002. 53BP1 functions in an ATM-dependent checkpoint pathway that is constitutively activated in human cancer. *Nat Cell Biol,* 4**,** 998-1002.

DOBBS, F. M., VAN EIJK, P., FELLOWS, M. D., LOIACONO, L., NITSCH, R. & REED, S. H. 2022. Precision digital mapping of endogenous and induced genomic DNA breaks by INDUCE-seq. *Nat Commun,* 13**,** 3989.

DROPCHO, E. J. 1991. Central nervous system injury by therapeutic irradiation. *Neurol Clin,* 9**,** 969-88.

DUNNICK, W., HERTZ, G. Z., SCAPPINO, L. & GRITZMACHER, C. 1993. DNA sequences at immunoglobulin switch region recombination sites. *Nucleic Acids Res,* 21**,** 365-72.

DURAND-DUBIEF, M., PERSSON, J., NORMAN, U., HARTSUIKER, E. & EKWALL, K. 2010. Topoisomerase I regulates open chromatin and controls gene expression in vivo. *Embo j,* 29**,** 2126-34.

DURAND-DUBIEF, M., SVENSSON, J. P., PERSSON, J. & EKWALL, K. 2011. Topoisomerases, chromatin and transcription termination. *Transcription,* 2**,** 66-70.

DURKIN, S. G., RAGLAND, R. L., ARLT, M. F., MULLE, J. G., WARREN, S. T. & GLOVER, T. W. 2008. Replication stress induces tumor-like microdeletions in FHIT/FRA3B. *Proc Natl Acad Sci U S A,* 105**,** 246-51.

DUTTA, A., ECKELMANN, B., ADHIKARI, S., AHMED, K. M., SENGUPTA, S., PANDEY, A., HEGDE, P. M., TSAI, M.-S., TAINER, J. A., WEINFELD, M., HEGDE, M. L. & MITRA, S. 2017. Microhomology-mediated end joining is activated in irradiated human cells due to phosphorylation-dependent formation of the XRCC1 repair complex. *Nucleic Acids Research,* 45**,** 2585-2599.

EDDY, J., VALLUR, A. C., VARMA, S., LIU, H., REINHOLD, W. C., POMMIER, Y. & MAIZELS, N. 2011. G4 motifs correlate with promoter-proximal transcriptional pausing in human genes. *Nucleic Acids Research,* 39**,** 4975-4983.

EL-KHAMISY, S. F., MASUTANI, M., SUZUKI, H. & CALDECOTT, K. W. 2003. A requirement for PARP-1 for the assembly or stability of XRCC1 nuclear foci at sites of oxidative DNA damage. *Nucleic Acids Res,* 31**,** 5526-33.

EMLET, D. R., GUPTA, P., HOLGADO-MADRUGA, M., DEL VECCHIO, C. A., MITRA, S. S., HAN, S. Y., LI, G., JENSEN, K. C., VOGEL, H., XU, L. W., SKIRBOLL, S. S. & WONG, A. J. 2014. Targeting a glioblastoma cancer stem-cell population defined by EGF receptor variant III. *Cancer Res,* 74**,** 1238-49.

ERRINGTON, F., WILLMORE, E., TILBY, M. J., LI, L., LI, G., LI, W., BAGULEY, B. C. & AUSTIN, C. A. 1999. Murine transgenic cells lacking DNA topoisomerase IIbeta are resistant to acridines and mitoxantrone: analysis of cytotoxicity and cleavable complex formation. *Mol Pharmacol,* 56**,** 1309-16.

EWELS, P., MAGNUSSON, M., LUNDIN, S. & KÄLLER, M. 2016. MultiQC: summarize analysis results for multiple tools and samples in a single report. *Bioinformatics,* 32**,** 3047-8.

EWELS, P. A., PELTZER, A., FILLINGER, S., PATEL, H., ALNEBERG, J., WILM, A., GARCIA, M. U., DI TOMMASO, P. & NAHNSEN, S. 2020. The nf-core framework for community-curated bioinformatics pipelines. *Nature Biotechnology,* 38**,** 276-278.

FACCHINO, S., ABDOUH, M., CHATOO, W. & BERNIER, G. 2010. BMI1 confers radioresistance to normal and cancerous neural stem cells through recruitment of the DNA damage response machinery. *J Neurosci,* 30**,** 10096-111.

FAEL AL-MAYHANI, T. M., BALL, S. L., ZHAO, J. W., FAWCETT, J., ICHIMURA, K., COLLINS, P. V. & WATTS, C. 2009. An efficient method for derivation and propagation of glioblastoma cell lines that conserves the molecular profile of their original tumours. *J Neurosci Methods,* 176**,** 192-9.

FALK, A., KOCH, P., KESAVAN, J., TAKASHIMA, Y., LADEWIG, J., ALEXANDER, M., WISKOW, O., TAILOR, J., TROTTER, M., POLLARD, S., SMITH, A. & BRÜSTLE, O. 2012. Capture of neuroepithelial-like stem cells from pluripotent stem cells provides a versatile system for in vitro production of human neurons. *PLoS One,* 7**,** e29597.

FEATHERSTONE, C. & JACKSON, S. P. 1999. DNA double-strand break repair. *Current Biology,* 9**,** R759-R761.

FERNANDEZ-VIDAL, A., VIGNARD, J. & MIREY, G. 2017. Around and beyond 53BP1 Nuclear Bodies. *Int J Mol Sci,* 18.

FORAY, N., CHARVET, A.-M., DUCHEMIN, D., FAVAUDON, V. & LAVALETTE, D. 2005. The repair rate of radiation-induced DNA damage: A stochastic interpretation based on the Gamma function. *Journal of Theoretical Biology,* 236**,** 448-458.

FOURNIER, L. A., KUMAR, A. & STIRLING, P. C. 2018. Chromatin as a Platform for Modulating the Replication Stress Response. *Genes (Basel),* 9.

FRATTINI, C., PROMONET, A., ALGHOUL, E., VIDAL-EYCHENIE, S., LAMARQUE, M., BLANCHARD, M.-P., URBACH, S., BASBOUS, J. & CONSTANTINOU, A. 2021. TopBP1 assembles nuclear condensates to switch on ATR signaling. *Molecular Cell,* 81**,** 1231-1245.e8.

FRIEDMANN-MORVINSKI, D., BUSHONG, E. A., KE, E., SODA, Y., MARUMOTO, T., SINGER, O., ELLISMAN, M. H. & VERMA, I. M. 2012. Dedifferentiation of neurons and astrocytes by oncogenes can induce gliomas in mice. *Science,* 338**,** 1080-4.

FRIESNER, J. D., LIU, B., CULLIGAN, K. & BRITT, A. B. 2005. Ionizing radiation-dependent gamma-H2AX focus formation requires ataxia telangiectasia mutated and ataxia telangiectasia mutated and Rad3-related. *Mol Biol Cell,* 16**,** 2566-76.

FUMAGALLI, M., ROSSIELLO, F., CLERICI, M., BAROZZI, S., CITTARO, D., KAPLUNOV, J. M., BUCCI, G., DOBREVA, M., MATTI, V., BEAUSEJOUR, C. M., HERBIG, U., LONGHESE, M. P. & D'ADDA DI FAGAGNA, F. 2012. Telomeric DNA damage is irreparable and causes persistent DNA-damage-response activation. *Nat Cell Biol,* 14**,** 355-65.

GASSER, S. M., LAROCHE, T., FALQUET, J., BOY DE LA TOUR, E. & LAEMMLI, U. K. 1986. Metaphase chromosome structure. Involvement of topoisomerase II. *J Mol Biol,* 188**,** 613-29.

GEL, B., DÍEZ-VILLANUEVA, A., SERRA, E., BUSCHBECK, M., PEINADO, M. A. & MALINVERNI, R. 2016. regioneR: an R/Bioconductor package for the association analysis of genomic regions based on permutation tests. *Bioinformatics,* 32**,** 289-91.

GEL, B. & SERRA, E. 2017. karyoploteR: an R/Bioconductor package to plot customizable genomes displaying arbitrary data. *Bioinformatics,* 33**,** 3088-3090.

GELLERT, M., LIPSETT, M. N. & DAVIES, D. R. 1962. Helix formation by guanylic acid. *Proc Natl Acad Sci U S A,* 48**,** 2013-8.

GEORGAKILAS, A. G., O'NEILL, P. & STEWART, R. D. 2013. Induction and repair of clustered DNA lesions: what do we know so far? *Radiat Res,* 180**,** 100-9.

GHOSH, A., PANDEY, SATYA P., JOSHI, DHEERAJ C., RANA, P., ANSARI, ASGAR H., SUNDAR, JENNIFER S., SINGH, P., KHAN, Y., EKKA, MARY K., CHAKRABORTY, D. & MAITI, S. 2023. Identification of G-quadruplex structures in MALAT1 lncRNA that interact with nucleolin and nucleophosmin. *Nucleic Acids Research,* 51**,** 9415-9431.

GIMPLE, R. C., BHARGAVA, S., DIXIT, D. & RICH, J. N. 2019. Glioblastoma stem cells: lessons from the tumor hierarchy in a lethal cancer. *Genes Dev,* 33**,** 591-609.

GINJALA, V., NACERDDINE, K., KULKARNI, A., OZA, J., HILL, S. J., YAO, M., CITTERIO, E., VAN LOHUIZEN, M. & GANESAN, S. 2011. BMI1 is recruited to DNA breaks and contributes to DNA damage-induced H2A ubiquitination and repair. *Mol Cell Biol,* 31**,** 1972-82.

GINNO, P. A., LOTT, P. L., CHRISTENSEN, H. C., KORF, I. & CHÉDIN, F. 2012. R-loop formation is a distinctive characteristic of unmethylated human CpG island promoters. *Mol Cell,* 45**,** 814-25.

GIRASOL, M. J., KRASILNIKOVA, M., MARQUES, C. A., DAMASCENO, J. D., LAPSLEY, C., LEMGRUBER, L., BURCHMORE, R., BERALDI, D., CARRUTHERS, R., BRIGGS, E. M. & MCCULLOCH, R. 2023. RAD51-mediated R-loop formation acts to repair transcription-associated DNA breaks driving antigenic variation in Trypanosoma brucei. *Proc Natl Acad Sci U S A,* 120**,** e2309306120.

GLÜCKSMANN, A. & SPEAR, F. G. 1939. The Effect of Gamma Radiation on Cells in Vivo Part II. *British Journal of Radiology,* 12**,** 486-498.

GOEHRING, L., HUANG, T. T. & SMITH, D. J. 2023. Transcription–Replication Conflicts as a Source of Genome Instability. *Annual Review of Genetics,* 57**,** 157-179.

GOFFART, N., KROONEN, J. & ROGISTER, B. 2013. Glioblastoma-initiating cells: relationship with neural stem cells and the micro-environment. *Cancers (Basel),* 5**,** 1049-71.

GOMES, C. P., NÓBREGA-PEREIRA, S., DOMINGUES-SILVA, B., REBELO, K., ALVES-VALE, C., MARINHO, S. P., CARVALHO, T., DIAS, S. & BERNARDES DE JESUS, B. 2019. An antisense transcript mediates MALAT1 response in human breast cancer. *BMC Cancer,* 19**,** 771.

GONZÁLEZ-BARRERA, S., GARCÍA-RUBIO, M. & AGUILERA, A. 2002. Transcription and double-strand breaks induce similar mitotic recombination events in Saccharomyces cerevisiae. *Genetics,* 162**,** 603-14.

GOODARZI, A. A., NOON, A. T., DECKBAR, D., ZIV, Y., SHILOH, Y., LÖBRICH, M. & JEGGO, P. A. 2008. ATM signaling facilitates repair of DNA double-strand breaks associated with heterochromatin. *Mol Cell,* 31**,** 167-77.

GRANDI, F. C., MODI, H., KAMPMAN, L. & CORCES, M. R. 2022. Chromatin accessibility profiling by ATAC-seq. *Nature Protocols,* 17**,** 1518-1552.

GU, Z., GU, L., EILS, R., SCHLESNER, M. & BRORS, B. 2014. circlize implements and enhances circular visualization in R. *Bioinformatics,* 30**,** 2811-2812.

HACIOGLU, C., KAR, F., DAVRAN, F. & TUNCER, C. 2023. Borax regulates iron chaperone- and autophagy-mediated ferroptosis pathway in glioblastoma cells. *Environ Toxicol,* 38**,** 1690-1701.

HAFFNER, M. C., ARYEE, M. J., TOUBAJI, A., ESOPI, D. M., ALBADINE, R., GUREL, B., ISAACS, W. B., BOVA, G. S., LIU, W., XU, J., MEEKER, A. K., NETTO, G., DE MARZO, A. M., NELSON, W. G. & YEGNASUBRAMANIAN, S. 2010. Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. *Nature Genetics,* 42**,** 668-675.

HAMPERL, S., BOCEK, M. J., SALDIVAR, J. C., SWIGUT, T. & CIMPRICH, K. A. 2017. Transcription-Replication Conflict Orientation Modulates R-Loop Levels and Activates Distinct DNA Damage Responses. *Cell,* 170**,** 774-786.e19.

HANAHAN, D. 2022. Hallmarks of Cancer: New Dimensions. *Cancer Discovery,* 12**,** 31-46.

HAO, Z., LV, D., GE, Y., SHI, J., WEIJERS, D., YU, G. & CHEN, J. 2020. RIdeogram: drawing SVG graphics to visualize and map genome-wide data on the idiograms. *PeerJ Comput. Sci.,* 6**,** e251.

HARROD, A., LANE, K. A. & DOWNS, J. A. 2020. The role of the SWI/SNF chromatin remodelling complex in the response to DNA double strand breaks. *DNA Repair (Amst),* 93**,** 102919.

HAUSMANN, M., WAGNER, E., LEE, J. H., SCHROCK, G., SCHAUFLER, W., KRUFCZIK, M., PAPENFUß, F., PORT, M., BESTVATER, F. & SCHERTHAN, H. 2018. Super-resolution localization microscopy of radiation-induced histone H2AX-phosphorylation in relation to H3K9-trimethylation in HeLa cells. *Nanoscale,* 10**,** 4320-4331.

HEGI, M. E., DISERENS, A.-C., GORLIA, T., HAMOU, M.-F., DE TRIBOLET, N., WELLER, M., KROS, J. M., HAINFELLNER, J. A., MASON, W., MARIANI, L., BROMBERG, J. E. C., HAU, P., MIRIMANOFF, R. O., CAIRNCROSS, J. G., JANZER, R. C. & STUPP, R. 2005. MGMT Gene Silencing and Benefit from Temozolomide in Glioblastoma. *New England Journal of Medicine,* 352**,** 997-1003.

HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell,* 38**,** 576-89.

HELMRICH, A., BALLARINO, M. & TORA, L. 2011. Collisions between replication and transcription complexes cause common fragile site instability at the longest human genes. *Mol Cell,* 44**,** 966-77.

HOLLIDAY, R. 1964. A mechanism for gene conversion in fungi. *Genetical Research,* 5**,** 282-304.

HOLLOMAN, W. K., WIEGAND, R., HOESSLI, C. & RADDING, C. M. 1975. Uptake of homologous single-stranded fragments by superhelical DNA: a possible mechanism for initiation of genetic recombination. *Proc Natl Acad Sci U S A,* 72, 2394-8.

HU, J., MEYERS, R. M., DONG, J., PANCHAKSHARI, R. A., ALT, F. W. & FROCK, R. L. 2016. Detecting DNA double-stranded breaks in mammalian genomes by linear amplification-mediated high-throughput genome-wide translocation sequencing. *Nat Protoc,* 11, 853-71.

HUANG, X., TRAGANOS, F. & DARZYNKIEWICZ, Z. 2003. DNA damage induced by DNA topoisomerase I- and topoisomerase II-inhibitors detected by histone H2AX phosphorylation in relation to the cell cycle phase and apoptosis. *Cell Cycle,* 2, 614-9.

HUARTE, M. 2015. The emerging role of lncRNAs in cancer. *Nature Medicine,* 21, 1253-1261.

HUMPHRIES, A., CERESER, B., GAY, L. J., MILLER, D. S., DAS, B., GUTTERIDGE, A., ELIA, G., NYE, E., JEFFERY, R., POULSOM, R., NOVELLI, M. R., RODRIGUEZ-JUSTO, M., MCDONALD, S. A., WRIGHT, N. A. & GRAHAM, T. A. 2013. Lineage tracing reveals multipotent stem cells maintain human adenomas and the pattern of clonal expansion in tumor evolution. *Proc Natl Acad Sci U S A,* 110, E2490-9.

IACOVONI, J. S., CARON, P., LASSADI, I., NICOLAS, E., MASSIP, L., TROUCHE, D. & LEGUBE, G. 2010. High-resolution profiling of gammaH2AX around DNA double strand breaks in the mammalian genome. *Embo j,* 29, 1446-57.

IRONY-TUR SINAI, M. & KEREM, B. 2018. DNA replication stress drives fragile site instability. *Mutat Res,* 808, 56-61.

ISMAIL, I. H., ANDRIN, C., MCDONALD, D. & HENDZEL, M. J. 2010. BMI1-mediated histone ubiquitylation promotes DNA double-strand break repair. *J Cell Biol,* 191, 45-60.

IVANOV, V. N. & HEI, T. K. 2014. Radiation-induced glioblastoma signaling cascade regulates viability, apoptosis and differentiation of neural stem cells (NSC). *Apoptosis,* 19, 1736-54.

IYER, M. K., NIKNAFS, Y. S., MALIK, R., SINGHAL, U., SAHU, A., HOSONO, Y., BARRETTE, T. R., PRENSNER, J. R., EVANS, J. R., ZHAO, S., POLIAKOV, A., CAO, X., DHANASEKARAN, S. M., WU, Y. M., ROBINSON, D. R., BEER, D. G., FENG, F. Y., IYER, H. K. & CHINNAIYAN, A. M. 2015. The landscape of long noncoding RNAs in the human transcriptome. *Nat Genet,* 47, 199-208.

JAIN, S. K., COX, M. M. & INMAN, R. B. 1995. Occurrence of three-stranded DNA within a RecA protein filament. *J Biol Chem,* 270, 4943-9.

JU, B. G., LUNYAK, V. V., PERISSI, V., GARCIA-BASSETS, I., ROSE, D. W., GLASS, C. K. & ROSENFELD, M. G. 2006. A topoisomerase IIbeta-mediated dsDNA break required for regulated transcription. *Science,* 312, 1798-802.

JUNG, D. & ALT, F. W. 2004. Unraveling V(D)J recombination; insights into gene regulation. *Cell,* 116, 299-311.

KAWASHIMA, Y., YAMAGUCHI, N., TESHIMA, R., NARAHARA, H., YAMAOKA, Y., ANAI, H., NISHIDA, Y. & HANADA, K. 2017. Detection of DNA double-strand breaks by pulsed-field gel electrophoresis. *Genes Cells,* 22, 84-93.

KENIG, S., FAORO, V., BOURKOULA, E., PODERGAJS, N., IUS, T., VINDIGNI, M., SKRAP, M., LAH, T., CESSELLI, D., STORICI, P. & VINDIGNI, A. 2016. Topoisomerase IIß mediates the resistance of glioblastoma stem cells to replication stress-inducing drugs. *Cancer Cell Int,* 16, 58.

KHAN, F. A. & ALI, S. O. 2017. Physiological Roles of DNA Double-Strand Breaks. *J Nucleic Acids,* 2017, 6439169.

KHARCHENKO, P. V., ALEKSEYENKO, A. A., SCHWARTZ, Y. B., MINODA, A., RIDDLE, N. C., ERNST, J., SABO, P. J., LARSCHAN, E., GORCHAKOV, A. A., GU, T., LINDER-BASSO, D., PLACHETKA, A., SHANOWER, G., TOLSTORUKOV, M. Y., LUQUETTE, L. J., XI, R., JUNG, Y. L., PARK, R. W., BISHOP, E. P., CANFIELD, T. K., SANDSTROM, R., THURMAN, R. E., MACALPINE, D. M., STAMATOYANNOPOULOS, J. A., KELLIS, M., ELGIN, S. C., KURODA, M. I., PIRROTTA, V., KARPEN, G. H. & PARK, P. J. 2011. Comprehensive analysis of the chromatin landscape in Drosophila melanogaster. *Nature,* 471**,** 480-5.

KIEFFER, S. R. & LOWNDES, N. F. 2022. Immediate-Early, Early, and Late Responses to DNA Double Stranded Breaks. *Front Genet,* 13**,** 793884.

KING, H. O., BREND, T., PAYNE, H. L., WRIGHT, A., WARD, T. A., PATEL, K., EGNUNI, T., STEAD, L. F., PATEL, A., WURDAK, H. & SHORT, S. C. 2017. RAD51 Is a Selective DNA Repair Target to Radiosensitize Glioma Stem Cells. *Stem Cell Reports,* 8**,** 125-139.

KOYANAGI, E., KAKIMOTO, Y., MINAMISAWA, T., YOSHIFUJI, F., NATSUME, T., HIGASHITANI, A., OGI, T., CARR, A. M., KANEMAKI, M. T. & DAIGAKU, Y. 2022. Global landscape of replicative DNA polymerase usage in the human genome. *Nat Commun,* 13**,** 7221.

KUMAGAI, A., LEE, J., YOO, H. Y. & DUNPHY, W. G. 2006. TopBP1 activates the ATR-ATRIP complex. *Cell,* 124**,** 943-55.

KUMAR, R., NAGPAL, G., KUMAR, V., USMANI, S. S., AGRAWAL, P. & RAGHAVA, G. P. S. 2019. HumCFS: a database of fragile sites in human chromosomes. *BMC Genomics,* 19**,** 985.

KUMARI, N., VARTAK, S. V., DAHAL, S., KUMARI, S., DESAI, S. S., GOPALAKRISHNAN, V., CHOUDHARY, B. & RAGHAVAN, S. C. 2019. G-quadruplex Structures Contribute to Differential Radiosensitivity of the Human Genome. *iScience,* 21**,** 288-307.

KUO, L. J. & YANG, L. X. 2008. Gamma-H2AX - a novel biomarker for DNA double-strand breaks. *In Vivo,* 22**,** 305-9.

KUZMINOV, A. 2001. Single-strand interruptions in replicating chromosomes cause double-strand breaks. *Proc Natl Acad Sci U S A,* 98**,** 8241-6.

LATHIA, J. D., GALLAGHER, J., HEDDLESTON, J. M., WANG, J., EYLER, C. E., MACSWORDS, J., WU, Q., VASANJI, A., MCLENDON, R. E., HJELMELAND, A. B. & RICH, J. N. 2010. Integrin alpha 6 regulates glioblastoma stem cells. *Cell Stem Cell,* 6**,** 421-32.

LATHIA, J. D., MACK, S. C., MULKEARNS-HUBERT, E. E., VALENTIM, C. L. & RICH, J. N. 2015. Cancer stem cells in glioblastoma. *Genes Dev,* 29**,** 1203-17.

LAWLOR, K., MARQUES-TORREJON, M. A., DHARMALINGHAM, G., EL-AZHAR, Y., SCHNEIDER, M. D., POLLARD, S. M. & RODRÍGUEZ, T. A. 2020. Glioblastoma stem cells induce quiescence in surrounding neural stem cells via Notch signaling. *Genes Dev,* 34**,** 1599-1604.

LEE, J., KOTLIAROVA, S., KOTLIAROV, Y., LI, A., SU, Q., DONIN, N. M., PASTORINO, S., PUROW, B. W., CHRISTOPHER, N., ZHANG, W., PARK, J. K. & FINE, H. A. 2006. Tumor stem cells derived from glioblastomas cultured in bFGF and EGF more closely mirror the phenotype and genotype of primary tumors than do serum-cultured cell lines. *Cancer Cell,* 9**,** 391-403.

LEE, J. H., LEE, J. E., KAHNG, J. Y., KIM, S. H., PARK, J. S., YOON, S. J., UM, J. Y., KIM, W. K., LEE, J. K., PARK, J., KIM, E. H., CHUNG, W. S., JU, Y. S., PARK, S. H., CHANG, J. H. & KANG, S. G. 2018. Human glioblastoma arises

from subventricular zone cells with low-level driver mutations. *Nature,* 560**,** 243-247.

LEE, J. S., LEE, H. J., MOON, B. H., SONG, S. H., LEE, M. O., SHIM, S. H., KIM, H. S., LEE, M. C., KWON, J. T., FORNACE, A. J., JR., KIM, S. U. & CHA, H. J. 2012. Generation of cancerous neural stem cells forming glial tumor by oncogenic stimulation. *Stem Cell Rev Rep,* 8**,** 532-45.

LENSING, S. V., MARSICO, G., HÄNSEL-HERTSCH, R., LAM, E. Y., TANNAHILL, D. & BALASUBRAMANIAN, S. 2016. DSBCapture: in situ capture and sequencing of DNA breaks. *Nature Methods,* 13**,** 855-857.

LEONARD, A. & WOLFF, J. E. 2013. Etoposide improves survival in high-grade glioma: a meta-analysis. *Anticancer Res,* 33**,** 3307-15.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G. & DURBIN, R. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIANG, F. & JASIN, M. 1996. Ku80-deficient cells exhibit excess degradation of extrachromosomal DNA. *J Biol Chem,* 271**,** 14405-11.

LIANG, L., DENG, L., CHEN, Y., LI, G. C., SHAO, C. & TISCHFIELD, J. A. 2005. Modulation of DNA End Joining by Nuclear Proteins*. *Journal of Biological Chemistry,* 280**,** 31442-31449.

LIAU, B. B., SIEVERS, C., DONOHUE, L. K., GILLESPIE, S. M., FLAVAHAN, W. A., MILLER, T. E., VENTEICHER, A. S., HEBERT, C. H., CAREY, C. D., RODIG, S. J., SHAREEF, S. J., NAJM, F. J., VAN GALEN, P., WAKIMOTO, H., CAHILL, D. P., RICH, J. N., ASTER, J. C., SUVÀ, M. L., PATEL, A. P. & BERNSTEIN, B. E. 2017. Adaptive Chromatin Remodeling Drives Glioblastoma Stem Cell Plasticity and Drug Tolerance. *Cell Stem Cell,* 20**,** 233-246.e7.

LIM, Y. C., ROBERTS, T. L., DAY, B. W., HARDING, A., KOZLOV, S., KIJAS, A. W., ENSBEY, K. S., WALKER, D. G. & LAVIN, M. F. 2012. A role for homologous recombination and abnormal cell-cycle progression in radioresistance of glioma-initiating cells. *Mol Cancer Ther,* 11**,** 1863-72.

LIM, Y. C., ROBERTS, T. L., DAY, B. W., STRINGER, B. W., KOZLOV, S., FAZRY, S., BRUCE, Z. C., ENSBEY, K. S., WALKER, D. G., BOYD, A. W. & LAVIN, M. F. 2014. Increased sensitivity to ionizing radiation by targeting the homologous recombination pathway in glioma initiating cells. *Mol Oncol,* 8**,** 1603-15.

LIMBO, O., CHAHWAN, C., YAMADA, Y., DE BRUIN, R. A., WITTENBERG, C. & RUSSELL, P. 2007. Ctp1 is a cell-cycle-regulated protein that functions with Mre11 complex to control double-strand break repair by homologous recombination. *Mol Cell,* 28**,** 134-46.

LINKE, R., LIMMER, M., JURANEK, S. A., HEINE, A. & PAESCHKE, K. 2021. The Relevance of G-Quadruplexes for DNA Repair. *Int J Mol Sci,* 22.

LIU, H., FU, H., YU, C., ZHANG, N., HUANG, C., LV, L., HU, C., CHEN, F., XIAO, Z., ZHANG, Z., LU, H. & YUAN, K. 2023. Transcriptional pausing induced by ionizing radiation enables the acquisition of radioresistance in nasopharyngeal carcinoma. *J Mol Cell Biol*.

LOPES, J., PIAZZA, A., BERMEJO, R., KRIEGSMAN, B., COLOSIO, A., TEULADE-FICHOU, M. P., FOIANI, M. & NICOLAS, A. 2011. G-quadruplex-induced instability during leading-strand replication. *Embo j,* 30**,** 4033-46.

LOUIS, D. N., PERRY, A., WESSELING, P., BRAT, D. J., CREE, I. A., FIGARELLA-BRANGER, D., HAWKINS, C., NG, H. K., PFISTER, S. M., REIFENBERGER, G., SOFFIETTI, R., VON DEIMLING, A. & ELLISON, D. W. 2021. The 2021 WHO

Classification of Tumors of the Central Nervous System: a summary. *Neuro Oncol,* 23**,** 1231-1251.

LOVE, M. I., HUBER, W. & ANDERS, S. 2014. Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biology,* 15**,** 550.

MA, A. & DAI, X. 2018. The relationship between DNA single-stranded damage response and double-stranded damage response. *Cell Cycle,* 17**,** 73-79.

MA, Y., PANNICKE, U., SCHWARZ, K. & LIEBER, M. R. 2002. Hairpin opening and overhang processing by an Artemis/DNA-dependent protein kinase complex in nonhomologous end joining and V(D)J recombination. *Cell,* 108**,** 781-94.

MACK, S. C., SINGH, I., WANG, X., HIRSCH, R., WU, Q., VILLAGOMEZ, R., BERNATCHEZ, J. A., ZHU, Z., GIMPLE, R. C., KIM, L. J. Y., MORTON, A., LAI, S., QIU, Z., PRAGER, B. C., BERTRAND, K. C., MAH, C., ZHOU, W., LEE, C., BARNETT, G. H., VOGELBAUM, M. A., SLOAN, A. E., CHAVEZ, L., BAO, S., SCACHERI, P. C., SIQUEIRA-NETO, J. L., LIN, C. Y. & RICH, J. N. 2019. Chromatin landscapes reveal developmentally encoded transcriptional states that define human glioblastoma. *J Exp Med,* 216**,** 1071-1090.

MADABHUSHI, R., GAO, F., PFENNING, A. R., PAN, L., YAMAKAWA, S., SEO, J., RUEDA, R., PHAN, T. X., YAMAKAWA, H., PAO, P. C., STOTT, R. T., GJONESKA, E., NOTT, A., CHO, S., KELLIS, M. & TSAI, L. H. 2015. Activity-Induced DNA Breaks Govern the Expression of Neuronal Early-Response Genes. *Cell,* 161**,** 1592-605.

MAFFIA, A., RANISE, C. & SABBIONEDA, S. 2020. From R-Loops to G-Quadruplexes: Emerging New Threats for the Replication Fork. *International journal of molecular sciences,* 21**,** 1506.

MAINTAINER, B. P. 2023. liftOver: Changing genomic coordinate systems with rtracklayer::liftOver.

MANDAL, P. K., BLANPAIN, C. & ROSSI, D. J. 2011. DNA damage response in adult stem cells: pathways and consequences. *Nat Rev Mol Cell Biol*. England.

MARAMPON, F., MEGIORNI, F., CAMERO, S., CRESCIOLI, C., MCDOWELL, H. P., SFERRA, R., VETUSCHI, A., POMPILI, S., VENTURA, L., DE FELICE, F., TOMBOLINI, V., DOMINICI, C., MAGGIO, R., FESTUCCIA, C. & GRAVINA, G. L. 2017. HDAC4 and HDAC6 sustain DNA double strand break repair and stem-like phenotype by promoting radioresistance in glioblastoma cells. *Cancer Lett,* 397**,** 1-11.

MARKOVA, E., VASILYEV, S. & BELYAEV, I. 2015. 53BP1 foci as a marker of tumor cell radiosensitivity. *Neoplasma,* 62**,** 770-6.

MARTÍ, J. M., GARCIA-DIAZ, A., DELGADO-BELLIDO, D., O'VALLE, F., GONZÁLEZ-FLORES, A., CARLEVARIS, O., RODRÍGUEZ-VARGAS, J. M., AMÉ, J. C., DANTZER, F., KING, G. L., DZIEDZIC, K., BERRA, E., DE ÁLAVA, E., AMARAL, A. T., HAMMOND, E. M. & OLIVER, F. J. 2021. Selective modulation by PARP-1 of HIF-1α-recruitment to chromatin during hypoxia is required for tumor adaptation to hypoxic conditions. *Redox Biol,* 41**,** 101885.

MARTI, T. M., HEFNER, E., FEENEY, L., NATALE, V. & CLEAVER, J. E. 2006. H2AX phosphorylation within the G1 phase after UV irradiation depends on nucleotide excision repair and not DNA double-strand breaks. *Proc Natl Acad Sci U S A,* 103**,** 9891-6.

MARTIN, F. J., AMODE, M. R., ANEJA, A., AUSTINE-ORIMOLOYE, O., AZOV, ANDREY G., BARNES, I., BECKER, A., BENNETT, R., BERRY, A., BHAI, J., BHURJI, SIMARPREET K., BIGNELL, A., BODDU, S., BRANCO LINS, P. R.,

BROOKS, L., RAMARAJU, S. B., CHARKHCHI, M., COCKBURN, A., DA RIN FIORRETTO, L., DAVIDSON, C., DODIYA, K., DONALDSON, S., EL HOUDAIGUI, B., EL NABOULSI, T., FATIMA, R., GIRON, C. G., GENEZ, T., GHATTAORAYA, G. S., MARTINEZ, J. G., GUIJARRO, C., HARDY, M., HOLLIS, Z., HOURLIER, T., HUNT, T., KAY, M., KAYKALA, V., LE, T., LEMOS, D., MARQUES-COELHO, D., MARUGÁN, J. C., MERINO, GABRIELA A., MIRABUENO, LOUISSE P., MUSHTAQ, A., HOSSAIN, SYED N., OGEH, D. N., SAKTHIVEL, M. P., PARKER, A., PERRY, M., PILIŽOTA, I., PROSOVETSKAIA, I., PÉREZ-SILVA, J. G., SALAM, AHAMED IMRAN A., SARAIVA-AGOSTINHO, N., SCHUILENBURG, H., SHEPPARD, D., SINHA, S., SIPOS, B., STARK, W., STEED, E., SUKUMARAN, R., SUMATHIPALA, D., SUNER, M.-M., SURAPANENI, L., SUTINEN, K., SZPAK, M., TRICOMI, FRANCESCA F., URBINA-GÓMEZ, D., VEIDENBERG, A., WALSH, THOMAS A., WALTS, B., WASS, E., WILLHOFT, N., ALLEN, J., ALVAREZ-JARRETA, J., CHAKIACHVILI, M., FLINT, B., GIORGETTI, S., HAGGERTY, L., ILSLEY, GARTH R., LOVELAND, JANE E., MOORE, B., MUDGE, JONATHAN M., TATE, J., THYBERT, D., TREVANION, STEPHEN J., WINTERBOTTOM, A., FRANKISH, A., HUNT, S. E., RUFFIER, M., CUNNINGHAM, F., DYER, S., FINN, ROBERT D., HOWE, KEVIN L., HARRISON, P. W., YATES, A. D. & FLICEK, P. 2023. Ensembl 2023. *Nucleic Acids Research,* 51**,** D933-D941.

MCCOWN, P. J., WANG, M. C., JAEGER, L. & BROWN, J. A. 2019. Secondary Structural Model of Human MALAT1 Reveals Multiple Structure-Function Relationships. *Int J Mol Sci,* 20.

MCKINNON, P. J. 2016. Topoisomerases and the regulation of neural function. *Nature Reviews Neuroscience,* 17**,** 673-679.

MICHEL, N., YOUNG, H. M. R., ATKIN, N. D., ARSHAD, U., AL-HUMADI, R., SINGH, S., MANUKYAN, A., GORE, L., BURBULIS, I. E., WANG, Y.-H. & MCCONNELL, M. J. 2022. Transcription-associated DNA DSBs activate p53 during hiPSC-based neurogenesis. *Scientific Reports,* 12**,** 12156.

MIMORI, T., HARDIN, J. A. & STEITZ, J. A. 1986. Characterization of the DNA-binding protein antigen Ku recognized by autoantibodies from patients with rheumatic disorders. *J Biol Chem,* 261**,** 2274-8.

MOU, X., LIEW, S. W. & KWOK, C. K. 2022. Identification and targeting of G-quadruplex structures in MALAT1 long non-coding RNA. *Nucleic Acids Res,* 50**,** 397-410.

NATHANSON, D. A., GINI, B., MOTTAHEDEH, J., VISNYEI, K., KOGA, T., GOMEZ, G., ESKIN, A., HWANG, K., WANG, J., MASUI, K., PAUCAR, A., YANG, H., OHASHI, M., ZHU, S., WYKOSKY, J., REED, R., NELSON, S. F., CLOUGHESY, T. F., JAMES, C. D., RAO, P. N., KORNBLUM, H. I., HEATH, J. R., CAVENEE, W. K., FURNARI, F. B. & MISCHEL, P. S. 2014. Targeted therapy resistance mediated by dynamic regulation of extrachromosomal mutant EGFR DNA. *Science,* 343**,** 72-6.

NEALE, M. J. & KEENEY, S. 2006. Clarifying the mechanics of DNA strand exchange in meiotic recombination. *Nature,* 442**,** 153-8.

NEGLIA, J. P., ROBISON, L. L., STOVALL, M., LIU, Y., PACKER, R. J., HAMMOND, S., YASUI, Y., KASPER, C. E., MERTENS, A. C., DONALDSON, S. S., MEADOWS, A. T. & INSKIP, P. D. 2006. New primary neoplasms of the central nervous system in survivors of childhood cancer: A report from the childhood cancer survivor study. *Journal of the National Cancer Institute,* 98**,** 1528-1537.

NEGRINI, S., GORGOULIS, V. G. & HALAZONETIS, T. D. 2010. Genomic instability — an evolving hallmark of cancer. *Nature Reviews Molecular Cell Biology,* 11, 220-228.

NICKOLOFF, J. A. 2022. Targeting Replication Stress Response Pathways to Enhance Genotoxic Chemo- and Radiotherapy. *Molecules,* 27.

NICKOLOFF, J. A., SHARMA, N. & TAYLOR, L. 2020. Clustered DNA Double-Strand Breaks: Biological Effects and Relevance to Cancer Radiotherapy. *Genes (Basel),* 11.

NICKOLOFF, J. A., SHARMA, N., TAYLOR, L., ALLEN, S. J. & HROMAS, R. 2021. The Safe Path at the Fork: Ensuring Replication-Associated DNA Double-Strand Breaks are Repaired by Homologous Recombination. *Front Genet,* 12, 748033.

NODA, A., HIRAI, Y., HAMASAKI, K., MITANI, H., NAKAMURA, N. & KODAMA, Y. 2012. Unrepairable DNA double-strand breaks that are generated by ionising radiation determine the fate of normal human cells. *J Cell Sci,* 125, 5280-7.

NOUBISSI, F. K., MCBRIDE, A. A., LEPPERT, H. G., MILLET, L. J., WANG, X. & DAVERN, S. M. 2021a. Detection and quantification of γ-H2AX using a dissociation enhanced lanthanide fluorescence immunoassay. *Sci Rep,* 11, 8945.

NOUBISSI, F. K., MCBRIDE, A. A., LEPPERT, H. G., MILLET, L. J., WANG, X. & DAVERN, S. M. 2021b. Detection and quantification of γ-H2AX using a dissociation enhanced lanthanide fluorescence immunoassay. *Scientific Reports,* 11, 8945.

OHKURA, H. 2015. Meiosis: an overview of key differences from mitosis. *Cold Spring Harb Perspect Biol,* 7.

OLIVE, P. L. & BANÁTH, J. P. 2006. The comet assay: a method to measure DNA damage in individual cells. *Nature Protocols,* 1, 23-29.

OSTER, S. & AQEILAN, R. I. 2020. Mapping the breakome reveals tight regulation on oncogenic super-enhancers. *Mol Cell Oncol,* 7, 1698933.

OSTROM, Q. T., GITTLEMAN, H., LIAO, P., VECCHIONE-KOVAL, T., WOLINSKY, Y., KRUCHKO, C. & BARNHOLTZ-SLOAN, J. S. 2017. CBTRUS Statistical Report: Primary brain and other central nervous system tumors diagnosed in the United States in 2010-2014. *Neuro Oncol,* 19, v1-v88.

OZERI-GALAI, E., LEBOFSKY, R., RAHAT, A., BESTER, A. C., BENSIMON, A. & KEREM, B. 2011. Failure of origin activation in response to fork stalling leads to chromosomal instability at fragile sites. *Mol Cell,* 43, 122-31.

PAN, M., WRIGHT, W. C., CHAPPLE, R. H., ZUBAIR, A., SANDHU, M., BATCHELDER, J. E., HUDDLE, B. C., LOW, J., BLANKENSHIP, K. B., WANG, Y., GORDON, B., ARCHER, P., BRADY, S. W., NATARAJAN, S., POSGAI, M. J., SCHUETZ, J., MILLER, D., KALATHUR, R., CHEN, S., CONNELLY, J. P., BABU, M. M., DYER, M. A., PRUETT-MILLER, S. M., FREEMAN, B. B., CHEN, T., GODLEY, L. A., BLANCHARD, S. C., STEWART, E., EASTON, J. & GEELEHER, P. 2021. The chemotherapeutic CX-5461 primarily targets TOP2B and exhibits selective activity in high-risk neuroblastoma. *Nature Communications,* 12, 6468.

PANIER, S. & BOULTON, S. J. 2014. Double-strand break repair: 53BP1 comes into focus. *Nature Reviews Molecular Cell Biology,* 15, 7-18.

PANINA, Y., GERMOND, A. & WATANABE, T. M. 2020. Analysis of the stability of 70 housekeeping genes during iPS reprogramming. *Sci Rep,* 10, 21711.

PAPADOPOULOU, C., GUILBAUD, G., SCHIAVONE, D. & SALE, J. E. 2015. Nucleotide Pool Depletion Induces G-Quadruplex-Dependent Perturbation of Gene Expression. *Cell Rep,* 13**,** 2491-2503.

PAPAEMMANUIL, E., RAPADO, I., LI, Y., POTTER, N. E., WEDGE, D. C., TUBIO, J., ALEXANDROV, L. B., VAN LOO, P., COOKE, S. L., MARSHALL, J., MARTINCORENA, I., HINTON, J., GUNDEM, G., VAN DELFT, F. W., NIK-ZAINAL, S., JONES, D. R., RAMAKRISHNA, M., TITLEY, I., STEBBINGS, L., LEROY, C., MENZIES, A., GAMBLE, J., ROBINSON, B., MUDIE, L., RAINE, K., O'MEARA, S., TEAGUE, J. W., BUTLER, A. P., CAZZANIGA, G., BIONDI, A., ZUNA, J., KEMPSKI, H., MUSCHEN, M., FORD, A. M., STRATTON, M. R., GREAVES, M. & CAMPBELL, P. J. 2014. RAG-mediated recombination is the predominant driver of oncogenic rearrangement in ETV6-RUNX1 acute lymphoblastic leukemia. *Nat Genet,* 46**,** 116-25.

PATEL, H., ESPINOSA-CARRASCO, J., LANGER, B., EWELS, P., BOT, N.-C., GARCIA, M. U., SYME, R., PELTZER, A., TALBOT, A., BEHRENS, D., GABERNET, G., JIN, M., HÖRTENHUBER, M., RODRIGUEZ, J. G., MENDEN, K. & AN, Ö. 2023. nf-core/atacseq: [2.1.2] - 2022-08-07. Zenodo.

PENNINCKX, S., CEKANAVICIUTE, E., DEGORRE, C., GUIET, E., VIGER, L., LUCAS, S. & COSTES, S. V. 2019. Dose, LET and Strain Dependence of Radiation-Induced 53BP1 Foci in 15 Mouse Strains Ex Vivo Introducing Novel DNA Damage Metrics. *Radiat Res,* 192**,** 1-12.

PERCHARDE, M., BULUT-KARSLIOGLU, A. & RAMALHO-SANTOS, M. 2017. Hypertranscription in Development, Stem Cells, and Regeneration. *Dev Cell,* 40**,** 9-21.

PETERMANN, E., ORTA, M. L., ISSAEVA, N., SCHULTZ, N. & HELLEDAY, T. 2010. Hydroxyurea-stalled replication forks become progressively inactivated and require two different RAD51-mediated pathways for restart and repair. *Mol Cell,* 37**,** 492-502.

PETKAU, A. 1987. Role of superoxide dismutase in modification of radiation injury. *Br J Cancer Suppl,* 8**,** 87-95.

PFEIFFER, P., GOEDECKE, W. & OBE, G. 2000. Mechanisms of DNA double-strand break repair and their potential to induce chromosomal aberrations. *Mutagenesis,* 15**,** 289-302.

PIETRAS, A., KATZ, A. M., EKSTRÖM, E. J., WEE, B., HALLIDAY, J. J., PITTER, K. L., WERBECK, J. L., AMANKULOR, N. M., HUSE, J. T. & HOLLAND, E. C. 2014. Osteopontin-CD44 signaling in the glioma perivascular niche enhances cancer stem cell phenotypes and promotes aggressive tumor growth. *Cell Stem Cell,* 14**,** 357-69.

POLAK, P., KARLIĆ, R., KOREN, A., THURMAN, R., SANDSTROM, R., LAWRENCE, M., REYNOLDS, A., RYNES, E., VLAHOVIČEK, K., STAMATOYANNOPOULOS, J. A. & SUNYAEV, S. R. 2015. Cell-of-origin chromatin organization shapes the mutational landscape of cancer. *Nature,* 518**,** 360-364.

POMMIER, Y., NUSSENZWEIG, A., TAKEDA, S. & AUSTIN, C. 2022. Human topoisomerases and their roles in genome stability and organization. *Nat Rev Mol Cell Biol,* 23**,** 407-427.

POND, K. W. & ELLIS, N. A. 2019. Quantification of Double-Strand Breaks in Mammalian Cells Using Pulsed-Field Gel Electrophoresis. *Methods Mol Biol,* 1999**,** 75-85.

PROMONET, A., PADIOLEAU, I., LIU, Y., SANZ, L., BIERNACKA, A., SCHMITZ, A. L., SKRZYPCZAK, M., SARRAZIN, A., METTLING, C., ROWICKA, M., GINALSKI, K., CHEDIN, F., CHEN, C. L., LIN, Y. L. & PASERO, P. 2020.

Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites. *Nat Commun,* 11, 3940.
PUGET, N., MILLER, K. M. & LEGUBE, G. 2019. Non-canonical DNA/RNA structures during Transcription-Coupled Double-Strand Break Repair: Roadblocks or Bona fide repair intermediates? *DNA Repair,* 81, 102661.
QIAN, H., MARGARETHA PLAT, A., JONKER, A., HOEBE, R. A. & KRAWCZYK, P. 2024. Super-resolution GSDIM microscopy unveils distinct nanoscale characteristics of DNA repair foci under diverse genotoxic stress. *DNA Repair (Amst),* 134, 103626.
QIU, G.-H. 2015. Protection of the genome and central protein-coding sequences by non-coding DNA against DNA damage from radiation. *Mutation Research/Reviews in Mutation Research,* 764, 108-117.
QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26, 841-842.
QUINTANA, E., SHACKLETON, M., SABEL, M. S., FULLEN, D. R., JOHNSON, T. M. & MORRISON, S. J. 2008. Efficient tumour formation by single human melanoma cells. *Nature,* 456, 593-598.
RAMÍREZ, F., RYAN, D. P., GRÜNING, B., BHARDWAJ, V., KILPERT, F., RICHTER, A. S., HEYNE, S., DÜNDAR, F. & MANKE, T. 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res,* 44, W160-5.
RAMSDEN, D. A. & GELLERT, M. 1995. Formation and resolution of double-strand break intermediates in V(D)J rearrangement. *Genes Dev,* 9, 2409-20.
REINA-SAN-MARTIN, B., NUSSENZWEIG, M. C., NUSSENZWEIG, A. & DIFILIPPANTONIO, S. 2005. Genomic instability, endoreduplication, and diminished Ig class-switch recombination in B cells lacking Nbs1. *Proc Natl Acad Sci U S A,* 102, 1590-5.
REYES, A. & HUBER, W. 2018. Alternative start and termination sites of transcription drive most transcript isoform differences across human tissues. *Nucleic Acids Res,* 46, 582-592.
ROBERTS, R. W. & CROTHERS, D. M. 1992. Stability and properties of double and triple helices: dramatic effects of RNA or DNA backbone composition. *Science,* 258, 1463-6.
ROBINSON, M. D., MCCARTHY, D. J. & SMYTH, G. K. 2010. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics,* 26, 139-40.
ROGAKOU, E. P., BOON, C., REDON, C. & BONNER, W. M. 1999. Megabase Chromatin Domains Involved in DNA Double-Strand Breaks in Vivo. *Journal of Cell Biology,* 146, 905-916.
ROGAKOU, E. P., PILCH, D. R., ORR, A. H., IVANOVA, V. S. & BONNER, W. M. 1998. DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139. *J Biol Chem,* 273, 5858-68.
ROTHKAMM, K., BARNARD, S., MOQUET, J., ELLENDER, M., RANA, Z. & BURDAK-ROTHKAMM, S. 2015. DNA damage foci: Meaning and significance. *Environ Mol Mutagen,* 56, 491-504.
ROTHKAMM, K. & LÖBRICH, M. 2003. Evidence for a lack of DNA double-strand break repair in human cells exposed to very low x-ray doses. *Proc Natl Acad Sci U S A,* 100, 5057-62.
ROY, D. & LIEBER, M. R. 2009. G clustering is important for the initiation of transcription-induced R-loops in vitro, whereas high G density without clustering is sufficient thereafter. *Mol Cell Biol,* 29, 3124-33.

ROY, I. M., NADAR, P. S. & KHURANA, S. 2021. Neutral Comet Assay to Detect and Quantitate DNA Double-Strand Breaksin Hematopoietic Stem Cells. *Bio Protoc,* 11, e4130.

SAAYMAN, X., GRAHAM, E., NATHAN, W. J., NUSSENZWEIG, A. & ESASHI, F. 2023. Centromeres as universal hotspots of DNA breakage, driving RAD51-mediated recombination during quiescence. *Mol Cell,* 83, 523-538.e7.

SAFA, A. R., SAADATZADEH, M. R., COHEN-GADOL, A. A., POLLOK, K. E. & BIJANGI-VISHEHSARAEI, K. 2015. Glioblastoma stem cells (GSCs) epigenetic plasticity and interconversion between differentiated non-GSCs and GSCs. *Genes Dis,* 2, 152-163.

SANO, K., MIYAJI-YAMAGUCHI, M., TSUTSUI, K. M. & TSUTSUI, K. 2008. Topoisomerase IIbeta activates a subset of neuronal genes that are repressed in AT-rich genomic environment. *PLoS One,* 3, e4103.

SANZ, L. A. & CHÉDIN, F. 2019. High-resolution, strand-specific R-loop mapping via S9.6-based DNA-RNA immunoprecipitation and high-throughput sequencing. *Nat Protoc,* 14, 1734-1755.

SCHINDELIN, J., ARGANDA-CARRERAS, I., FRISE, E., KAYNIG, V., LONGAIR, M., PIETZSCH, T., PREIBISCH, S., RUEDEN, C., SAALFELD, S., SCHMID, B., TINEVEZ, J.-Y., WHITE, D. J., HARTENSTEIN, V., ELICEIRI, K., TOMANCAK, P. & CARDONA, A. 2012. Fiji: an open-source platform for biological-image analysis. *Nature Methods,* 9, 676-682.

SCHIPLER, A. & ILIAKIS, G. 2013. DNA double-strand–break complexity levels and their possible contributions to the probability for error-prone processing and repair pathway choice. *Nucleic Acids Research,* 41, 7589-7605.

SCHNEIDER, C. A., RASBAND, W. S. & ELICEIRI, K. W. 2012. NIH Image to ImageJ: 25 years of image analysis. *Nature Methods,* 9, 671-675.

SCHOEFFLER, A. J. & BERGER, J. M. 2008. DNA topoisomerases: harnessing and constraining energy to govern chromosome topology. *Q Rev Biophys,* 41, 41-101.

SCULLY, R., PANDAY, A., ELANGO, R. & WILLIS, N. A. 2019. DNA double-strand break repair-pathway choice in somatic mammalian cells. *Nat Rev Mol Cell Biol,* 20, 698-714.

SHACKLETON, M., QUINTANA, E., FEARON, E. R. & MORRISON, S. J. 2009. Heterogeneity in cancer: cancer stem cells versus clonal evolution. *Cell,* 138, 822-9.

SHARMA, S. 2011. Non-B DNA Secondary Structures and Their Resolution by RecQ Helicases. *J Nucleic Acids,* 2011, 724215.

SHIBATA, A. & JEGGO, P. A. 2020. Roles for 53BP1 in the repair of radiation-induced DNA double strand breaks. *DNA Repair (Amst),* 93, 102915.

SHIOTANI, B. & ZOU, L. 2009. ATR signaling at a glance. *Journal of Cell Science,* 122, 301-304.

SHORT, S. C., MARTINDALE, C., BOURNE, S., BRAND, G., WOODCOCK, M. & JOHNSTON, P. 2007. DNA repair after irradiation in glioma cells and normal human astrocytes. *Neuro Oncol,* 9, 404-11.

SINGH, D. K., KOLLIPARA, R. K., VEMIREDDY, V., YANG, X. L., SUN, Y., REGMI, N., KLINGLER, S., HATANPAA, K. J., RAISANEN, J., CHO, S. K., SIRASANAGANDLA, S., NANNEPAGA, S., PICCIRILLO, S., MASHIMO, T., WANG, S., HUMPHRIES, C. G., MICKEY, B., MAHER, E. A., ZHENG, H., KIM, R. S., KITTLER, R. & BACHOO, R. M. 2017. Oncogenes Activate an Autonomous Transcriptional Regulatory Circuit That Drives Glioblastoma. *Cell Rep,* 18, 961-976.

SINGH, S., SZLACHTA, K., MANUKYAN, A., RAIMER, H. M., DINDA, M., BEKIRANOV, S. & WANG, Y.-H. 2020. Pausing sites of RNA polymerase II on actively transcribed genes are enriched in DNA double-stranded breaks. *Journal of Biological Chemistry,* 295**,** 3990-4000.

SKENE, P. J. & HENIKOFF, S. 2017. An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites. *Elife,* 6.

SKOURTI-STATHAKI, K., KAMIENIARZ-GDULA, K. & PROUDFOOT, N. J. 2014. R-loops induce repressive chromatin marks over mammalian gene terminators. *Nature,* 516**,** 436-9.

SMITH, C. J., PERFETTI, T. A., CHOKSHI, C., VENUGOPAL, C., ASHFORD, J. W. & SINGH, S. K. 2024. Risk factors for glioblastoma are shared by other brain tumor types. *Hum Exp Toxicol,* 43**,** 9603271241241796.

SMOLKA, J. A., SANZ, L. A., HARTONO, S. R. & CHÉDIN, F. 2020. Recognition of cellular RNAs by the S9.6 antibody creates pervasive artefacts when imaging RNA:DNA hybrids. *bioRxiv***,** 2020.01.11.902981.

SO, A., LE GUEN, T., LOPEZ, B. S. & GUIROUILH-BARBAT, J. 2017. Genomic rearrangements induced by unscheduled DNA double strand breaks in somatic mammalian cells. *Febs j,* 284**,** 2324-2344.

SOBEK, S. & BOEGE, F. 2014. DNA topoisomerases in mtDNA maintenance and ageing. *Exp Gerontol,* 56**,** 135-41.

SPIEGEL, J., ADHIKARI, S. & BALASUBRAMANIAN, S. 2020. The Structure and Function of DNA G-Quadruplexes. *Trends Chem,* 2**,** 123-136.

STAVNEZER, J., GUIKEMA, J. E. & SCHRADER, C. E. 2008. Mechanism and regulation of class switch recombination. *Annu Rev Immunol,* 26**,** 261-92.

STEWART, D. J., RICHARD, M. T., HUGENHOLTZ, H., DENNERY, J. M., BELANGER, R., GERIN-LAJOIE, J., MONTPETIT, V., NUNDY, D., PRIOR, J. & HOPKINS, H. S. 1984. Penetration of VP-16 (etoposide) into human intracerebral and extracerebral tumors. *J Neurooncol,* 2**,** 133-9.

STRUMBERG, D., PILON, A. A., SMITH, M., HICKEY, R., MALKAS, L. & POMMIER, Y. 2000. Conversion of topoisomerase I cleavage complexes on the leading strand of ribosomal DNA into 5'-phosphorylated DNA double-strand breaks by replication runoff. *Mol Cell Biol,* 20**,** 3977-87.

STUPP, R., MASON, W. P., VAN DEN BENT, M. J., WELLER, M., FISHER, B., TAPHOORN, M. J., BELANGER, K., BRANDES, A. A., MAROSI, C., BOGDAHN, U., CURSCHMANN, J., JANZER, R. C., LUDWIN, S. K., GORLIA, T., ALLGEIER, A., LACOMBE, D., CAIRNCROSS, J. G., EISENHAUER, E. & MIRIMANOFF, R. O. 2005. Radiotherapy plus concomitant and adjuvant temozolomide for glioblastoma. *N Engl J Med,* 352**,** 987-96.

SUNTER, N. J., COWELL, I. G., WILLMORE, E., WATTERS, G. P. & AUSTIN, C. A. 2010. Role of Topoisomerase IIß in DNA Damage Response following IR and Etoposide. *J Nucleic Acids,* 2010.

TACHON, G., CORTES, U., GUICHET, P. O., RIVET, P., BALBOUS, A., MASLIANTSEV, K., BERGER, A., BOISSONNADE, O., WAGER, M. & KARAYAN-TAPON, L. 2018. Cell Cycle Changes after Glioblastoma Stem Cell Irradiation: The Major Role of RAD51. *Int J Mol Sci,* 19.

TAKAISHI, S., OKUMURA, T., TU, S., WANG, S. S., SHIBATA, W., VIGNESHWARAN, R., GORDON, S. A., SHIMADA, Y. & WANG, T. C. 2009. Identification of gastric cancer stem cells using the cell surface marker CD44. *Stem Cells,* 27**,** 1006-20.

TASCHNER-MANDL, S., SCHWARZ, M., BLAHA, J., KAUER, M., KROMP, F., FRANK, N., RIFATBEGOVIC, F., WEISS, T., LADENSTEIN, R., HOHENEGGER, M., AMBROS, I. M. & AMBROS, P. F. 2016. Metronomic topotecan impedes

tumor growth of MYCN-amplified neuroblastoma cells in vitro and in vivo by therapy induced senescence. *Oncotarget,* 7**,** 3571-86.

THOMAS, P. D., EBERT, D., MURUGANUJAN, A., MUSHAYAHAMA, T., ALBOU, L. P. & MI, H. 2022. PANTHER: Making genome-scale phylogenetics accessible to all. *Protein Sci,* 31**,** 8-22.

THONGTHIP, S., CARLSON, A., CROSSLEY, M. P. & SCHWER, B. 2022. Relationships between genome-wide R-loop distribution and classes of recurrent DNA breaks in neural stem/progenitor cells. *Scientific Reports,* 12**,** 13373.

TIRINO, V., CAMERLINGO, R., BIFULCO, K., IROLLO, E., MONTELLA, R., PAINO, F., SESSA, G., CARRIERO, M. V., NORMANNO, N., ROCCO, G. & PIROZZI, G. 2013. TGF-β1 exposure induces epithelial to mesenchymal transition both in CSCs and non-CSCs of the A549 cell line, leading to an increase of migration ability in the CD133+ A549 cell fraction. *Cell Death Dis,* 4**,** e620.

TRUONG, L. N., LI, Y., SHI, L. Z., HWANG, P. Y.-H., HE, J., WANG, H., RAZAVIAN, N., BERNS, M. W. & WU, X. 2013. Microhomology-mediated End Joining and Homologous Recombination share the initial end resection step to repair DNA double-strand breaks in mammalian cells. *Proceedings of the National Academy of Sciences of the United States of America,* 110**,** 7720-7725.

TSAI, H. Z., LIN, R. K. & HSIEH, T. S. 2016. Drosophila mitochondrial topoisomerase III alpha affects the aging process via maintenance of mitochondrial function and genome integrity. *J Biomed Sci,* 23**,** 38.

TSAI, S. Q., ZHENG, Z., NGUYEN, N. T., LIEBERS, M., TOPKAR, V. V., THAPAR, V., WYVEKENS, N., KHAYTER, C., IAFRATE, A. J., LE, L. P., ARYEE, M. J. & JOUNG, J. K. 2015. GUIDE-seq enables genome-wide profiling of off-target cleavage by CRISPR-Cas nucleases. *Nat Biotechnol,* 33**,** 187-197.

TSEGAY, P. S., LAI, Y. & LIU, Y. 2019. Replication Stress and Consequential Instability of the Genome and Epigenome. *Molecules,* 24.

TUTT, A., BERTWISTLE, D., VALENTINE, J., GABRIEL, A., SWIFT, S., ROSS, G., GRIFFIN, C., THACKER, J. & ASHWORTH, A. 2001. Mutation in Brca2 stimulates error-prone homology-directed repair of DNA double-strand breaks occurring between repeated sequences. *EMBO Journal,* 20**,** 4704-4716.

VAKKILAINEN, S., SKOOG, T., EINARSDOTTIR, E., MIDDLETON, A., PEKKINEN, M., ÖHMAN, T., KATAYAMA, S., KRJUTŠKOV, K., KOVANEN, P. E., VARJOSALO, M., LINDQVIST, A., KERE, J. & MÄKITIE, O. 2019. The human long non-coding RNA gene RMRP has pleiotropic effects and regulates cell-cycle progression at G2. *Scientific Reports,* 9**,** 13758.

VALDÉS, A., SEGURA, J., DYSON, S., MARTÍNEZ-GARCÍA, B. & ROCA, J. 2018. DNA knots occur in intracellular chromatin. *Nucleic Acids Res,* 46**,** 650-660.

VALDIGLESIAS, V., GIUNTA, S., FENECH, M., NERI, M. & BONASSI, S. 2013. γH2AX as a marker of DNA double strand breaks and genomic instability in human population studies. *Mutat Res,* 753**,** 24-40.

VAN WAARDE, M. A. W. H., VAN ASSEN, A. J., KONINGS, A. W. T. & KAMPINGA, H. H. 1996. Feasibility of measuring radiation-induced DNA double strand breaks and their repair by pulsed field gel electrophoresis in freshly isolated cells from the mouse RIF-1 tumor. *International Journal of Radiation Oncology*Biology*Physics,* 36**,** 125-134.

VENKATA NARAYANAN, I., PAULSEN, M. T., BEDI, K., BERG, N., LJUNGMAN, E. A., FRANCIA, S., VELOSO, A., MAGNUSON, B., DI FAGAGNA, F. D. A., WILSON, T. E. & LJUNGMAN, M. 2017. Transcriptional and post-transcriptional

regulation of the ionizing radiation response by ATM and p53. *Scientific Reports,* 7**,** 43598.
VILENCHIK, M. M. & KNUDSON, A. G. 2003. Endogenous DNA double-strand breaks: production, fidelity of repair, and induction of cancer. *Proc Natl Acad Sci U S A,* 100**,** 12871-6.
VILLEPONTEAU, B. 1997. The heterochromatin loss model of aging. *Experimental Gerontology,* 32**,** 383-394.
VÍTOR, A. C., HUERTAS, P., LEGUBE, G. & DE ALMEIDA, S. F. 2020. Studying DNA Double-Strand Break Repair: An Ever-Growing Toolbox. *Front Mol Biosci,* 7**,** 24.
VRAL, A., WILLEMS, P., CLAES, K., POPPE, B., PERLETTI, G. & THIERENS, H. 2011. Combined effect of polymorphisms in Rad51 and Xrcc3 on breast cancer risk and chromosomal radiosensitivity. *Mol Med Rep,* 4**,** 901-12.
WANG, G. & VASQUEZ, K. M. 2006. Non-B DNA structure-induced genetic instability. *Mutat Res,* 598**,** 103-19.
WANG, J., YU, X., CAO, X., TAN, L., JIA, B., CHEN, R. & LI, J. 2023. GAPDH: A common housekeeping gene with an oncogenic role in pan-cancer. *Comput Struct Biotechnol J,* 21**,** 4056-4069.
WANG, M., WEI, P. C., LIM, C. K., GALLINA, I. S., MARSHALL, S., MARCHETTO, M. C., ALT, F. W. & GAGE, F. H. 2020. Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress-Associated Genome Instability. *Cell Stem Cell,* 26**,** 221-233.e6.
WANG, Z., KATSAROS, D., BIGLIA, N., SHEN, Y., FU, Y., LOO, L. W. M., JIA, W., OBATA, Y. & YU, H. 2018. High expression of long non-coding RNA MALAT1 in breast cancer is associated with poor relapse-free survival. *Breast Cancer Res Treat,* 171**,** 261-271.
WARD, J. F. 2000. Complexity of damage produced by ionizing radiation. *Cold Spring Harb Symp Quant Biol,* 65**,** 377-82.
WATANABE, F., HOLLINGSWORTH, E. W., BARTLEY, J. M., WISEHART, L., DESAI, R., HARTLAUB, A. M., HESTER, M. E., SCHIAPPARELLI, P., QUIÑONES-HINOJOSA, A. & IMITOLA, J. 2024. Patient-derived organoids recapitulate glioma-intrinsic immune program and progenitor populations of glioblastoma. *PNAS Nexus,* 3**,** pgae051.
WEI, P.-C., CHANG, A. N., KAO, J., DU, Z., MEYERS, R. M., ALT, F. W. & SCHWER, B. 2016. Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells. *Cell,* 164**,** 644-655.
WICKHAM, H. 2016. *ggplot2,* Cham, Springer International Publishing.
WICKHAM, H., AVERICK, M., BRYAN, J., CHANG, W., MCGOWAN, L., FRANÇOIS, R., GROLEMUND, G., HAYES, A., HENRY, L., HESTER, J., KUHN, M., PEDERSEN, T., MILLER, E., BACHE, S., MÜLLER, K., OOMS, J., ROBINSON, D., SEIDEL, D., SPINU, V., TAKAHASHI, K., VAUGHAN, D., WILKE, C., WOO, K. & YUTANI, H. 2019. Welcome to the tidyverse. *J. Open Source Softw.,* 4**,** 1686.
WILSON, T. E., ARLT, M. F., PARK, S. H., RAJENDRAN, S., PAULSEN, M., LJUNGMAN, M. & GLOVER, T. W. 2015. Large transcription units unify copy number variants and common fragile sites arising under replication stress. *Genome Res,* 25**,** 189-200.
WIRSCHING, H.-G., GALANIS, E. & WELLER, M. 2016. Chapter 23 - Glioblastoma. *In:* BERGER, M. S. & WELLER, M. (eds.) *Handbook of Clinical Neurology*. Elsevier.
XIONG, Y., QI, Y., PAN, Z., WANG, S., LI, B., FENG, B., XUE, H., ZHAO, R. & LI, G. 2022. Pancancer landscape analysis of the thymosin family identified

TMSB10 as a potential prognostic biomarker and immunotherapy target in glioma. *Cancer Cell Int,* 22**,** 294.

XU, C., WU, Z., DUAN, H. C., FANG, X., JIA, G. & DEAN, C. 2021a. R-loop resolution promotes co-transcriptional chromatin silencing. *Nat Commun,* 12**,** 1790.

XU, W., DING, M., WANG, B., CAI, Y., GUO, C. & YUAN, C. 2021b. Molecular Mechanism of the Canonical Oncogenic lncRNA MALAT1 in Gastric Cancer. *Curr Med Chem,* 28**,** 8800-8809.

YADAV, A., BISWAS, T., PRAVEEN, A., GANGULY, P., BHATTACHARYYA, A., VERMA, A., DATTA, D. & ATEEQ, B. 2023. Targeting MALAT1 Augments Sensitivity to PARP Inhibition by Impairing Homologous Recombination in Prostate Cancer. *Cancer Res Commun,* 3**,** 2044-2061.

YAN, W. X., MIRZAZADEH, R., GARNERONE, S., SCOTT, D., SCHNEIDER, M. W., KALLAS, T., CUSTODIO, J., WERNERSSON, E., LI, Y., GAO, L., FEDEROVA, Y., ZETSCHE, B., ZHANG, F., BIENKO, M. & CROSETTO, N. 2017. BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks. *Nature Communications,* 8**,** 15058.

YANG, K. S., KOHLER, R. H., LANDON, M., GIEDT, R. & WEISSLEDER, R. 2015. Single cell resolution in vivo imaging of DNA damage following PARP inhibition. *Sci Rep,* 5**,** 10129.

YANG, Z., WANG, Z., FAN, Y. & ZHENG, Q. 2012. Expression of CD133 in SW620 colorectal cancer cells is modulated by the microenvironment. *Oncol Lett,* 4**,** 75-79.

YASUHARA, T., KATO, R., HAGIWARA, Y., SHIOTANI, B., YAMAUCHI, M., NAKADA, S., SHIBATA, A. & MIYAGAWA, K. 2018. Human Rad52 Promotes XPG-Mediated R-loop Processing to Initiate Transcription-Associated Homologous Recombination Repair. *Cell,* 175**,** 558-570.e11.

YIN, J., LIU, M., LIU, Y. & HU, J. 2019. Improved HTGTS for CRISPR/Cas9 off-target detection. *Bio Protoc,* 9**,** e3229.

YU, G., WANG, L.-G. & HE, Q.-Y. 2015. ChIPseeker: an R/Bioconductor package for ChIP peak annotation, comparison and visualization. *Bioinformatics,* 31**,** 2382-2383.

ZACK, T. I., SCHUMACHER, S. E., CARTER, S. L., CHERNIACK, A. D., SAKSENA, G., TABAK, B., LAWRENCE, M. S., ZHSNG, C. Z., WALA, J., MERMEL, C. H., SOUGNEZ, C., GABRIEL, S. B., HERNANDEZ, B., SHEN, H., LAIRD, P. W., GETZ, G., MEYERSON, M. & BEROUKHIM, R. 2013. Pan-cancer patterns of somatic copy number alteration. *Nat Genet,* 45**,** 1134-40.

ZEMAN, M. K. & CIMPRICH, K. A. 2014. Causes and consequences of replication stress. *Nat Cell Biol,* 16**,** 2-9.

ZHANG, J., LIAN, H., CHEN, K., PANG, Y., CHEN, M., HUANG, B., ZHU, L., XU, S., LIU, M. & ZHONG, C. 2021. RECQ1 Promotes Stress Resistance and DNA Replication Progression Through PARP1 Signaling Pathway in Glioblastoma. *Front Cell Dev Biol,* 9**,** 714868.

ZHANG, L., REYNOLDS, T. L., SHAN, X. & DESIDERIO, S. 2011. Coupling of V(D)J recombination to the cell cycle suppresses genomic instability and lymphoid tumorigenesis. *Immunity,* 34**,** 163-74.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology,* 9**,** R137.

ZHAO, Y., HUANG, Q., ZHANG, T., DONG, J., WANG, A., LAN, Q., GU, X. & QIN, Z. 2008. Ultrastructural studies of glioma stem cells/progenitor cells. *Ultrastruct Pathol,* 32**,** 241-5.

ZHONG, Y., NELLIMOOTTIL, T., PEACE, J. M., KNOTT, S. R., VILLWOCK, S. K., YEE, J. M., JANCUSKA, J. M., REGE, S., TECKLENBURG, M., SCLAFANI, R. A., TAVARÉ, S. & APARICIO, O. M. 2013. The level of origin firing inversely affects the rate of replication fork progression. *J Cell Biol,* 201**,** 373-83.

ZHU, Y., BIERNACKA, A., PARDO, B., DOJER, N., FOREY, R., SKRZYPCZAK, M., FONGANG, B., NDE, J., YOUSEFI, R., PASERO, P., GINALSKI, K. & ROWICKA, M. 2019. qDSB-Seq is a general method for genome-wide quantification of DNA double-strand breaks using sequencing. *Nat Commun,* 10**,** 2313.

ZLOTORYNSKI, E., RAHAT, A., SKAUG, J., BEN-PORAT, N., OZERI, E., HERSHBERG, R., LEVI, A., SCHERER, S. W., MARGALIT, H. & KEREM, B. 2003. Molecular basis for expression of common and rare fragile sites. *Mol Cell Biol,* 23**,** 7143-51.

ZYLKA, M. J., SIMON, J. M. & PHILPOT, B. D. 2015. Gene length matters in neurons. *Neuron,* 86**,** 353-5.