Investigating the Causes and Consequences of Genome Damage in Human Pluripotent Stem Cells

By:

Owen Sam Laing

Submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

University of Sheffield
Faculty of Science
School of Bioscience

September 2023

Acknowledgements

I could (and maybe should) acknowledge a long list of family, friends, colleagues etc. here, but in the interests of conciseness I will mention just a handful by name.

I would like to thank my supervisors, Ivana and Sherif, who've guided me through my PhD. In particular, Ivana has been extraordinarily supportive of me. She has always made time for me and has kept a level head throughout the last 4 years, remaining optimistic when I've had issues in the lab and being compassionate and accommodating when I've had issues outside of the lab.

I give special thanks to my girlfriend, Emmie, who has weathered this PhD with me, for what has now been half of our relationship. Emmie doesn't really benefit from the successes of my PhD but has suffered and cared for a deflated partner during the tougher times, which seems unfair. She deserves much more acknowledgement than I can fit in a paragraph. It's frowned upon to list your partner as a co-author for morale support, so this is the best written thanks she will get. She has been patient, caring and supportive throughout, and can always make me smile. Thank you, Emmie.

Abstract

Human pluripotent stem cells (hPSC) can be cultured indefinitely *in vitro* and differentiated into any somatic cell type. These unique features make hPSC valuable tools for the study of disease and development, as well as promising candidates as sources of cellular material for regenerative medicine. Such applications typically require large numbers of cells, necessitating prolonged *in vitro* expansion. Over the course of *in vitro* culture hPSC are prone to acquiring diverse, but typically recurrent, genetic changes, which can negatively affect their utility in research applications and pose significant safety concerns for their therapeutic application. Many of the genetic changes observed in hPSC are underpinned by genome damage, specifically double-strand DNA breaks (DSBs), followed by erroneous repair. While previous work has identified constitutively elevated levels of DSBs in pluripotent cells relative to their differentiated counterparts, the mechanisms driving such damage remain uncertain. Previous studies have been limited by their use of purely quantitative assays to investigate DNA damage in pluripotent cells. Recent advances in sequencing technologies have enabled high-resolution mapping of individual DSBs within a cell population, revealing site-specific mechanisms of DSB biogenesis in diverse cell types. Here, I use one such technology, INDUCE-seq, to generate the first genome-wide map of endogenous DSBs in isogenic pluripotent and differentiated cell types. Annotation of DSB maps with publicly available sequencing datasets revealed enrichment of DSBs in transcriptionally active regions of the genome, suggesting prevalent transcription-induced DNA damage in hPSC. Further, unbiased identification of sites of DSB accumulation identified marked DSB enrichment over a broad genomic region on the 1q21 cytoband, frequently implicated in recurrent translocations in hPSC cultures. Mutation rate experiments further suggested that transcription-mediated replication stress drives DSB formation at chromosome 1q21, ultimately increasing translocation frequency in hPSC. Together, these results constitute the first mechanistic evidence for biogenesis of a specific recurrent genetic change in hPSC and pave the way for strategies aimed at reducing the frequency of genetic changes in hPSC destined for clinical use.

Table of Contents

List of figures

List of tables

List of Abbreviations

APH	Aphidicolin
ATAC	Assay for transposase accessible chromatin
ATM	Ataxia telangiectasia–mutated
ATR	Ataxia telangiectasia and Rad3-related
BAM	Binary alignment map
BER	Base excision repair
CFS	Common fragile site
ChIP	Chromatin immunoprecipitation
CNV	Copy number variant
COSMIC	Catalogue of somatic mutations in cancer
CPT	Camptothecin
CRISPR	Clustered regularly interspaced short palindromic repeats
DAPI	4’,6-diamidino-2-phenylindole
DDR	DNA damage response
DE	Differentially enriched
DNA	Deoxyribonucleic acid
DRB	5,6-dichloro-1-beta-D-ribofuranosylbenzimidazole
DRIP	DNA:RNA Immunoprecipitation
DSB	Double-strand break
EC	Embryonal carcinoma
EDTA	Ethylenediaminetetraacetic acid
ERFS	Early replication fragile site
ESC	Embryonic stem cell
(m)ESC	Mouse embryonic stem cell
(h)ESC	Human embryonic stem cell

FACS	Fluorescence activated cell sorting
FBS	Foetal bovine serum
FDR	False discovery rate
FISH	Fluorescence in situ hybridization
FSC	Forward scatter
GO	Gene ontology
HLA	Human leukocyte antigen
HTGTS	High-throughput genome translocation sequencing
HR	Homologous recombination
ICM	Inner cell mass
iPSC	Induced pluripotent stem cell
ISSCR	International society for stem cell research
IVF	In vitro fertilisation
LNA	locked nucleic acid
MEF	Mouse embryonic fibroblast
MRN	MRE11-RAD50-NBS1
NER	Nucleotide excision repair
NHEJ	Non-homologous end joining
NPC	Neural progenitor cell
dNTP	Deoxynucleotide triphosphate
ORC	Origin recognition complex
PBS	Phosphate buffered saline
PFA	Paraformaldehyde
PMT	Photomultiplier tube
(h)PSC	(human) Pluripotent stem cell
QC	Quality control

RNA	Ribonucleic acid
RNAPII	RNA polymerase II
ROS	Reactive oxygen species
RPE	Retinal pigment epithelium
RPKM	Reads per kilobase per million
SAM	Sequence alignment map
SDS	Sodium dodecyl sulphate
SNP	Single nucleotide polymorphism
SNV	Single nucleotide variant
SRA	Sequence read archive
SSB	Single -strand break
SSC	Side scatter
SV	Structural variant
TE	Tris-EDTA
TF	Transcription factor
TFBS	Transcription factor binding site
TSS	Transcription start site
TTS	Transcription termination site
UV	Ultraviolet

1. Introduction

The derivation of human pluripotent stem cells (hPSC) provided opportunities for the *in vitro* expansion of large numbers of cells, followed by directed differentiation into any specialised cell type of the soma, opening new avenues in the study of development (Moris *et al.*, 2020), disease modelling (Rowe and Daley, 2019) and cell therapy (Ilic and Ogilvie, 2022). Effective use of hPSC, particularly in the context of cell therapy, has been hindered by several obstacles, including the efficiency of differentiation protocols, immunogenicity (Koga *et al.*, 2020) and genetic instability of hPSC (Halliwell et al., 2020). This thesis focuses on genome instability in hPSC, more specifically the causes of genome damage which underpin such instability. Below, I introduce pluripotent stem cells and the issue of genomic instability in this cell type, ahead of summarising known causes of DNA damage in eukaryotic cells, with a final reflection on the unique biological characteristics of hPSC, which may pre-dispose them to certain mechanisms of DNA damage.

1.1. Pluripotent Stem Cells

Pluripotency is defined as the ability to differentiate into cells of all three germ lineages and ultimately any somatic cell type. Pluripotent cells exist transiently *in vivo* during embryonic development. Pluripotent *stem* cells are the *in vitro*-cultured counterparts to the pluripotent cells of the early embryo which, in addition to trilineage differentiation capacity, can proliferate indefinitely in culture (Yilmaz and Benvenisty, 2019). The term *pluripotent stem cell* describes various cell types, with often distinct biologies. Whilst the focus of this thesis is human pluripotent stem cells, I begin by summarising historical and non-human models of pluripotent stem cells which have contributed to our understanding of the pluripotent state, highlighting key differences between classes of pluripotent cells.

1.1.1. Embryonal carcinoma cells

The first *in vitro*-cultured pluripotent stem cells were isolated from mice germ cell tumours, known as teratocarcinomas. Teratocarcinomas comprise differentiated cell types of all three germ lineages, as well as a core of pluripotent cells, termed embryonal carcinoma (EC) cells. Early experiments demonstrated that transplantation of a single EC cell was sufficient to give rise to an entire new teratocarcinoma in mouse (Kleinsmith and Plerce, 1964), hence EC cells were able to self-renew and give rise to cells of all three germ lineages. Extraction and *in vitro*

culture of this pluripotent cell type yielded the first pluripotent stem cells, which retained proliferative capacity over prolonged *in vitro* culture, and, upon injection into mice, gave rise to tumours comprising cells of all three germ layers (Finch and Ephrussi, 1967). The ultimate functional demonstration of pluripotency is to introduce cells into a blastocyst and determine their contribution to all tissues of the resulting organism in a so-called chimera test (Mascetti and Pedersen, 2016). Later studies demonstrated that mouse EC cells could contribute to a chimeric organism upon injection into pre-implantation blastocysts (Brinster, 1974; Papaioannou *et al.*, 1975). A decade later, the first pluripotent human EC cells were isolated, able to proliferate over prolonged culture, and differentiate both *in vivo* and *in vitro* (Andrews *et al.*, 1985). Research on human and mouse EC cells demonstrated that pluripotent stem cells could be maintained in an undifferentiated state *in vitro*. EC cells could undergo spontaneous differentiation, but more importantly, via stimulation with certain compounds, could be manipulated to differentiate into specific cell types (Jones-Villeneuve *et al.*, 1982). EC cell research was fundamental in identifying culture techniques necessary for maintenance of cells in an undifferentiated state as well as the expression of key surface antigens associated with pluripotency, laying the foundation for the subsequent derivation of embryonic stem cells (Andrews *et al.*, 2005).

1.1.2. Embryonic stem cells

The first non-malignant pluripotent stem cells were isolated in 1981 in a seminal study by Evans & Kaufman (1981). In mammalian embryonic development, following fertilization, the zygote gives rise to an entire new organism, and is defined as totipotent (Baker & Pera, 2018). In mammalian embryos, three population doublings, known as cleavage divisions, follow fertilization, with no accompanied increase in embryo mass. This is followed by formation of a morula, wherein proliferating embryonic cells, known as blastomeres compact in 3D space. Subsequently a blastocyst forms, comprised of extraembryonic tissue, known as trophectoderm, with a core of pluripotent cells, known as the inner cell mass (ICM) (Figure 1.1) (Niakan *et al.*, 2012). Evans and Kaufman isolated cells of the ICM, following brief *ex vivo* culture of mouse blastocysts, and found these cells to proliferate over long term *in vitro* culture, differentiate upon subcutaneous injection in mice, and yield chimeric organisms following injection into pre-implantation mouse blastocysts (Evans and Kaufman, 1981). These mouse embryonic stem cells (mESC) remain ICM-like over *in vitro* culture in terms of both transcription factor expression, and differentiation capacity and have been termed “naïve” pluripotent cells (Nichols and Smith, 2009).

Many years after the first successful derivation of mESCs, Thompson and colleagues were successful in isolating and expanding ICM cells from equivalent stage human blastocysts (Thomson *et al.*, 1998). These human embryonic stem cells (hESCs) could proliferate over long-term culture and differentiate *in vitro.* Although chimera tests are not feasible in humans, hESCs were also able to form teratomas, tumours comprised of cells of all three germ layers, when injected into mice, demonstrating *in vivo* pluripotency (Thomson *et al.*, 1998).

1.1.3. Induced Pluripotent Stem Cells

Although hESC are derived from surplus embryos generated for in vitro fertilisation (IVF), with donor consent, the field of hESC research has been the subject of controversy owing to the necessary destruction of potentially viable human embryos during their derivation. Indeed, three years after their first derivation, the US banned federal-funded research on hESC lines derived after 2001, a policy which remained effective until 2009 (Owen-Smith *et al.*, 2012). Similarly, since the derivation of the first hESC lines in 1998, Germany's strict laws on embryo research effectively precluded the use or derivation of hESC lines, which was only relaxed in 2008 to allow import of previously derived hESC lines for research purposes (Gottweis, 2002; Stafford, 2008).

In 2007, Takahashi and colleagues generated the first human induced pluripotent stem cell (iPSC) lines, reprogrammed from human fibroblasts, circumnavigating many of the ethical issues associated with hESC research (Takahashi *et al.*, 2007). Building on their work deriving mouse iPSC the previous year (Takahashi and Yamanaka, 2006), the group demonstrated that over-expression of four genes encoding key transcription factors, namely OCT3/4, SOX2, KLF4 and cMYC, was sufficient to revert human dermal fibroblast to a pluripotent *hESC-like* state. These iPSC could be maintained in an undifferentiated state over long-term culture, expressed pluripotency-associated transcription factors and were able to differentiate into cells of all three germ layers (Takahashi *et al.*, 2007). However, this work relied on retroviral transduction of cells with transgenes, which integrate at random in the genome and with variable copy number, precluding such cells from regenerative medicine applications (Takahashi *et al.*, 2007). To overcome this limitation, subsequent studies have developed methods of reprogramming somatic cells to iPSC using non-integrating Adenovirus or Sendai virus (Stadtfeld *et al.*, 2008; Fusaki *et al.*, 2009), episomal DNA (Okita *et al.*, 2008; Jia *et al.*, 2010), mRNA (Warren *et al.*, 2010), protein (Kim *et al.*, 2009), and most recently, small molecule inhibitors in the absence of any exogenous nucleic acid (Guan *et al.*, 2022).

1.1.4. Primed versus naïve pluripotency

With the exception of occasional reports of epigenetic memory from their parent cell types (Kim et al., 2010), hiPSC appear to be functionally largely indistinguishable from hESC and collectively both cell types come under the blanket term of human pluripotent stem cells (hPSC). mESCs and hPSC are both, by definition, pluripotent cell types. However, their pluripotent states are non-equivalent, with cells exhibiting differences in morphology and cloning efficiency (Wu and Belmonte, 2015). Most notably, the signalling pathways stimulated for maintenance of the two cell types are different, with hPSC requiring FGF signalling for maintenance in an undifferentiated state, whereas modern mESC media specifically inhibit FGF signalling to prevent differentiation (Vallier *et al.*, 2005; Ying *et al.*, 2008). Brons and colleagues successfully isolated mESCs from post-implantation epiblast cells, termed epiSC, which were morphologically more similar to hPSC, required FGF signalling for maintenance of an undifferentiated population and yielded cell types of each germ layer upon differentiation with *hESC-optimised* protocols (Brons *et al.*, 2007). Later transcriptional profiling studies demonstrated that, whilst mESC are most similar, in terms of gene expression, to blastomeres of the ICM in early blastocysts (Boroviak *et al.*, 2014), epiSC gene expression was more similar to that of the late gastrulation epiblast, i.e. a later developmental stage than that from which the cells were derived (Kojima *et al.*, 2014). The current dogma is that mESCs are an *in vitro* equivalent of the cells of the ICM in the pre-implantation blastocyst termed *naive*. By contrast hPSC, like epiSC, represent a later developmental stage equivalent to the post-implantation epiblast, termed *primed* (Figure 1.1) (Nichols and Smith, 2009). Whilst several studies have devised methods for the derivation and maintenance of naïve hPSC (Taei *et al.*, 2020), the work in this thesis focuses on cells in the primed state and any reference to hPSC from herein alludes to primed hPSC.

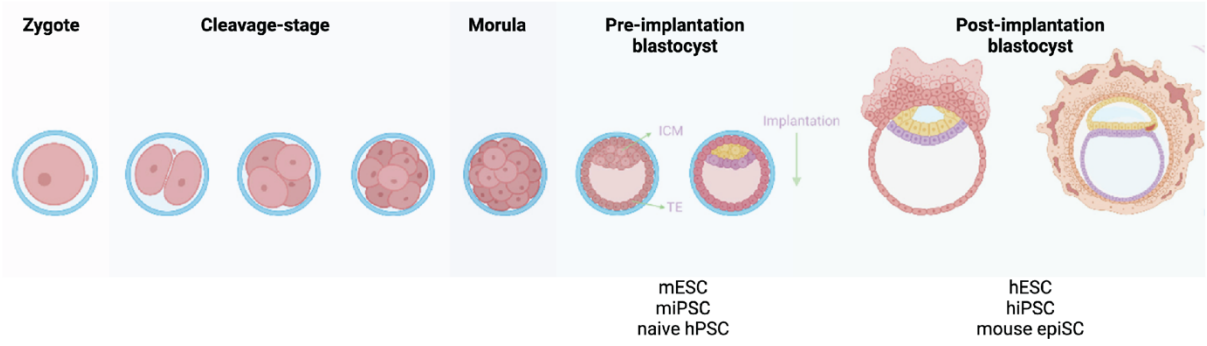


Figure 1.1 Mammalian embryogenesis and pluripotent stem cell equivalents

Following fertilization, the zygote undergoes cleavage divisions, where cells divide without an increase in total embryo volume. Cells continue to proliferate and compact together to form the morula which goes on to form the

pre-implantation blastocyst, containing an inner cell mass of pluripotent cells. The mature blastocyst implants into the uterine epithelium ahead of gastrulation and further embryonic development. Both mouse and human ESC are derived from the inner cell mass of the pre-implantation blastocyst. mESC, miPSC and naïve hPSC are most similar to the inner cell mass of the pre-implantation blastocyst. hESC, hiPSC and mouse epiSC are most similar to the epiblast of the post implantation embryo (yellow cells). Figure adapted from (Wang & Wu, 2022).

1.1.5. Characterising pluripotency

Pluripotency can be considered an unstable developmental substate, in which stimuli perturbing the signalling equilibrium, can induce gene expression changes leading to reversible lineage biases or irreversible lineage commitments (Enver *et al.*, 2009). As such, pluripotency cannot be taken for granted for any given cell line at any given time. Pluripotency is defined functionally, by the ability to give rise to all somatic cell types, the ultimate test being contribution to fertile chimeric offspring upon introduction of cells into the pre-implantation blastocyst (Brinster, 1974). For obvious ethical reasons, such a test is not possible for human cells and differentiation capacity is determined by surrogate tests. Historically, teratoma assays have been the method of choice, whereby putative pluripotent cells are subcutaneously injected into immunodeficient mice and left to form tumours. *Teratomas* are tumours comprising cells of all three germ layers and can only arise from pluripotent cells, hence the teratoma assay serves as a functional readout of cell's pluripotency (Montilla-Rojo *et al.*, 2023). More recently, amid efforts to minimise usage of animals, *in vitro* assays have become more popular. For characterisation of newly derived hPSC lines, the *International Society for Stem Cell Research* recommends directed *in vitro* differentiation into progenitor cell types of all three lineages, demonstrating differentiation by induction of lineage-specific marker expression accompanied by loss of expression of markers of the undifferentiated state (Table 1.1). Further, terminally differentiated cell types of each lineage should be generated and functionally characterised using multiparameter assays (ISSCR, 2023).

Table 1.1 Markers of undifferentiated and lineage-committed cells

Non-exhaustive list of commonly used markers of undifferentiated cells and progenitors of each germ layer.

Antigen	Localisation	Lineage
SSEA3	Plasma membrane	Undifferentiated
SSEA4	Plasma membrane	Undifferentiated
TRA-160	Plasma membrane	Undifferentiated
TRA-181	Plasma membrane	Undifferentiated

OCT4	Intracellular	Undifferentiated
NANOG	Intracellular	Undifferentiated
SOX1	Intracellular	Ectoderm
PAX6	Intracellular	Ectoderm
Nestin	Intracellular	Ectoderm
Brachyury	Intracellular	Mesoderm
NCAM	Plasma membrane	Mesoderm
HAND1	Intracellular	Mesoderm
SOX17	Intracellular	Endoderm
FOXA2	Intracellular	Endoderm

1.2. Applications of hPSC

In their 1998 paper describing the derivation of the first hESC lines, Thomson and colleagues discussed the potential use of these cells in the study of human embryo development, regulation of differentiation and regenerative medicine applications (Thomson *et al.*, 1998). All applications postulated in this manuscript have been realised in some form, in addition to further uses in drug and toxicology screening.

1.2.1. Modelling embryonic development

Legislation on human embryo research prohibits the *in vitro* culture of human embryos beyond 14 days post-fertilization (Warnock, 1985). As a result, much of our understanding of human embryogenesis is inferred from studies in mouse and non-human primate systems. As hPSC represent an epiblast-like cell type, they can be used to model embryonic development. Two-dimensional culture systems have been used to study cell patterning upon the onset of differentiation in hPSC, for instance BMP4-induced differentiation of hESCs grown on 2D circular micropatterned slides leads to formation of concentric rings of each germ layer, reminiscent of gastrulation in embryogenesis (Warmflash *et al.*, 2014). Subsequent studies have focused on three-dimensional models to more accurately recapitulate the organisation of cells *in vivo*. Moris and colleagues describe a “gastruloid” differentiation in 3D cell aggregates, in which cells form spatially separated germ layers and undergo axial patterning,

mirroring day 25-27 post-fertilisation human embryos and therefore modelling an embryonic stage impossible to study in cultured human embryos (Moris *et al.*, 2020).

Three dimensional models can also be used to determine the effect of drugs or environmental contaminants on human embryogenesis. Three dimensional aggregates of hPSC, called embryoid bodies, can spontaneously differentiate to give rise to cells of all three germ layers. The scalability of embryoid bodies makes them ideal for toxicology studies and have been employed for teratogenicity screens for drugs and other compounds, increasing throughput and reducing the use of animals in such tests (Mayshar *et al.*, 2011; Flamier *et al.*, 2017).

1.2.2. Modelling disease phenotypes

hPSC constitute an invaluable tool for the study of human disease. The basic principle of disease modelling with hPSC, is that a cell line harbouring a disease-specific mutation is differentiated into the affected cell type, and these cells can be used to determine physiological effects on cell function as well as for drug screens to determine the preliminary efficacy and safety of candidate drugs on the disease phenotype. Such a system has several advantages over animal models, most notably the cells used are human and therefore don't suffer translational issues stemming from idiosyncratic differences between species. Moreover, the scale and ease of culture facilitates higher throughput and lower cost testing than the use of animal models.

The earliest and simplest form of hPSC-based disease modelling entails taking somatic cells, typically fibroblasts, from a disease patient and reprogramming to iPSC before differentiating into the clinically relevant cell type. Such an approach has been used to effectively elucidate various disease aetiologies such as neuropathies in autistic neural progenitors cells (Wang *et al.,* 2020*)* and Charcot Marie Tooth Disease (Rizzo *et al.*, 2016). Following the advent of CRISPR-Cas9 genome editing technologies, disease models no longer require hiPSC reprogramming from affected patients, and existing hPSC lines, including hESC, can be genetically edited to introduce disease relevant mutations (Sen and Thummer, 2022). Such an approach is more readily accessible for many laboratories and limits the extensive characterisation and functional testing associated with deriving new hiPSC lines (ISSCR, 2023).

1.2.3. Regenerative medicine

The most exciting application of hPSC is undoubtedly their potential use in regenerative medicine. Many diseases, such as Parkinson's or macular degeneration, result from the loss

of a single cell type. Such diseases could theoretically be remedied by *in vitro* expansion and differentiation of hPSC into the affected cell type, followed by transplantation. The first hPSC-derived cell therapy to enter clinical trials was the treatment of spinal cord injury using injection of hESC-derived oligodendrocytes, which, whilst proven safe, failed to ameliorate patient symptoms (Kaiser, 2011). A subsequent trial using a higher cell titre demonstrated increased patient motor and sensory function in 95% of cases (Ilic and Ogilvie, 2022). Upon their initial derivation, one of the touted advantages of hiPSC was that they would enable patient-isogenic cell therapies minimising the risk of immune rejection (Takahashi *et al.*, 2007). The first of such trials used patient-derived iPSC to generate retinal pigment epithelial cells for the treatment of macular degeneration. Whilst safe, the process took 10 months from patient biopsy and cost ~$1m USD rendering it prohibitively expensive and time consuming for wide-spread adoption (Mandai *et al.*, 2017). More recent approaches to address the issue of immune rejection have focussed on modulating the expression of human leukocyte antigens (HLA) which mediate recognition of “self” or “non-self” by the immune system (Trowsdale and Knight, 2013). Generating monoallelic HLA-expressing iPSC compromises between the high cost associated with isogenic iPSC derivation and rare compatibility of heterozygous HLA-expressing cells (Koga *et al.*, 2020).

There are too many clinical trials involving hPSC-derived cell therapies to list comprehensively. The most common applications currently, are in treatment of eye disease, cardiovascular disease, neurodegeneration, and malignancies (Ilic and Ogilvie, 2022). Commercial interest and promise in such therapies is ever growing, illustrated by the exponential increase in trials, with over 120 clinical trials using hPSC-based therapies currently registered with the human pluripotent stem cell registry (hPSCreg, 2023). In spite of this increased interest in hPSC for therapeutic applications, significant hurdles to the progression of such therapies to clinic exist, including the emergence of genetically variant hPSC in *in vitro* cultures. Indeed, the first iPSC-based trial for RPE-reconstitution in macular degeneration was abandoned due the presence of single nucleotide variants (SNVs) and copy number variants (CNVs) of unknown consequence in the iPSCs line derived (Garber, 2015).

1.3. Genetic variants in hPSC cultures: a barrier for applications of hPSCs

1.3.1. Observed genetic variants in hPSC cultures

Genetic variants found in hPSC cultures can be broadly grouped into two categories: *cell of origin* i.e., a variant fixed in the population prior to *in vitro* culture, or *culture acquired*. Cell of origin variants may arise in hESC, where mutation in the zygote, or indeed parental germ cells, yields a non-mosaic genetically variant embryo. Much more common however, are cell of origin mutations in hiPSC, owing to a typically high mutational burden in somatic cells such as dermal fibroblasts, prior to reprogramming (Rouhani *et al.*, 2016, 2022). Cell of origin variants pose the same problems to hPSC applications as culture acquired variants and thus choice of somatic cell used in reprogramming iPSC is an important factor for consideration, with low-mutational burden cell types such as long-term haematopoietic stem cells being preferable (Wang *et al.*, 2019). However, the focus of this thesis is genome damage and ultimately genetic variation acquired over *in vitro* culture of hPSC, hence cell of origin variants are not discussed in great detail.

As hPSC research became more prevalent following their first derivation, it quickly became apparent that these cells were prone to acquiring genetic changes. These changes were first reported in the form of chromosomal abnormalities, readily detectable via karyology, most commonly gains of chromosomes 12 and 17 (Draper *et al.*, 2004). Soon after, translocations at certain regions as well as shorter sub-karyotypic gains and losses (CNVs) were reported (Figure 1.2) (Baker *et al.*, 2007; Spits *et al.*, 2008; Närvä *et al.*, 2010; Amps *et al.*, 2011; Taapken *et al.*, 2011). Over time, resolution of the methodologies used for detection of hPSC increased and in 2016, Merkle and colleagues, screened over 200 hPSC lines via whole exome sequencing or published RNA-seq, revealing recurrent loss of function mutations in the *TP53* gene at specific amino acid residues frequently mutated in cancer (Merkle *et al.*, 2017). Subsequent sequencing studies in hPSC revealed further recurrent point mutations in cancer-related genes *CDK12, EGFR, PATZ and BCOR* (Avior *et al.*, 2021; Rouhani *et al.*, 2022).

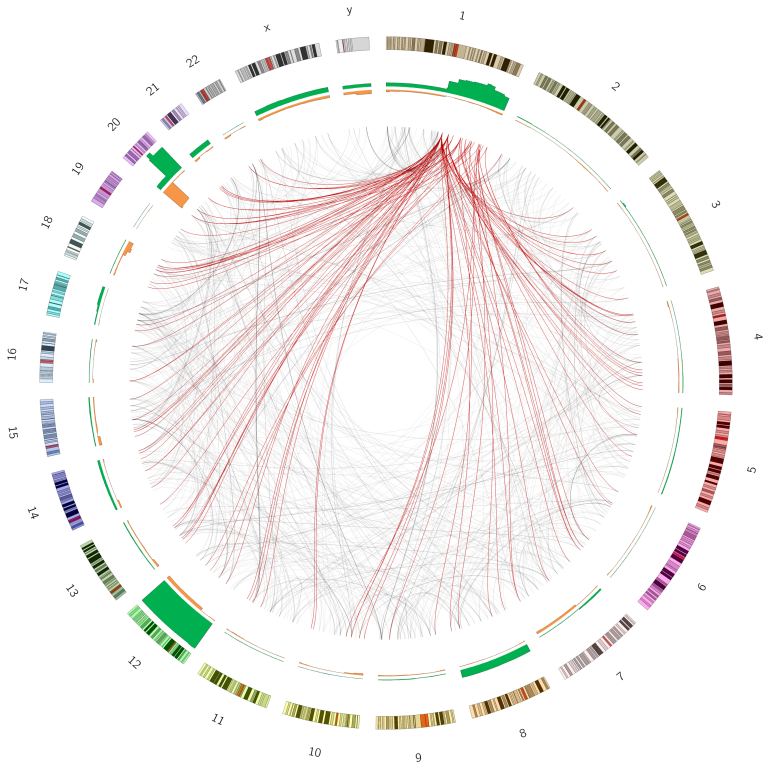


Figure 1.2 Recurrent translocations in hPSC lines

Circos plot generated from 4000 abnormal hPSC karyotypes. Chromosomes 1:XY labelled, lines represent translocation of material from one region to another. Green bars represent gain of genetic material, orange bars represent loss. Translocations involving chromosome 1q highlighted in red. Raw data obtained from WiCell (www.wicell.org). Plot extracted from Stavish et al., (manuscript in preparation)

The detectable presence of genetic variants in culture is the result of a two-step process, firstly mutation must occur, followed by a selection process which allows specific variants to rise to prevalence in culture (Figure 1.3). A commonality amongst these diverse genetic changes is their recurrent nature, i.e. the same specific changes are observed across different cell lines. Most methods for the detection of genetic variants are sensitive to around a 10% mosaic population, hence the variant population has expanded relative to the wildtype population at the time of detection, inferring that such variants have a fitness advantage allowing them to outcompete their wildtype counterparts (Baker *et al.*, 2016). Indeed, this fitness advantage is demonstrable, via co-culture of isogenic wildtype and variant hPSC clones (Olariu *et al.*, 2010; Avery *et al.*, 2013; Price *et al.*, 2021). Whilst there is ample evidence for the selective advantage of recurrent genetic variants driving recurrency in the changes observed, it remains unknown whether the genetic changes in hPSC cultures are also biased towards at the point of mutation.

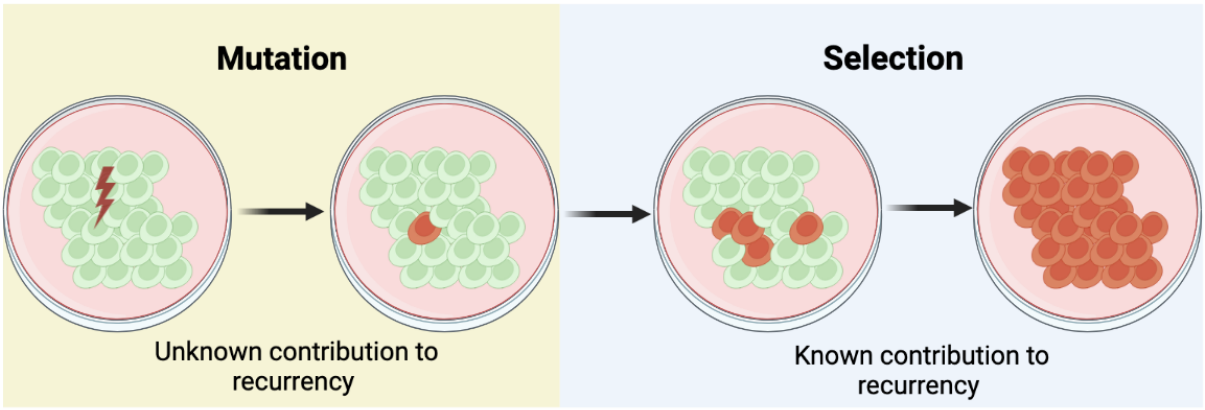


Figure 1.3 Acquisition of genetically variant hPSC: mutation and selection

Normal hPSC (green) suffer a mutation event, yielding a variant cell (red). Whether mutations occur at random is currently unknown. If the genetic variant has a fitness advantage over wildtype cells, it undergoes clonal expansion and is selected for over prolonged culture resulting in progressive overtake and elimination of wildtype cells.

1.3.2. Possible causes of mutation in hPSC cultures

Mutation arises from diverse mechanisms in cells, with often distinct causes for different classes of mutation. SNVs, recurrently found in cancer-related genes in hPSC (Merkle *et al.*, 2017; Avior *et al.*, 2021), can arise via several mechanisms including hydrolytic deamination of nucleotides (Lindahl, 1993), oxidation of nucleotides by reactive oxygen species (ROS) (Hussain *et al.*, 2003) and unfaithful DNA replication (McCulloch and Kunkel, 2008).

Indels are short insertions or deletions of DNA, typically less than 1kb in length. Whilst specific indels do not apparently recur in hPSC, indels still constitute a significant source of genetic variation in hPSC (Thompson *et al.*, 2020). Mechanistically, indels can arise via erroneous repair of DNA double-strand breaks (DSBs) (Cisneros-Aguirre *et al.*, 2022; Min *et al.*, 2023), or alternatively, slippage of DNA polymerase during DNA replication, particularly at repetitive sequences (Montgomery *et al.*, 2013).

Larger sub-chromosomal gains or losses of DNA, known as CNVs yield duplications or deletions of genes in hPSC, such as the frequently observed 20q11.21 duplication (Amps et al., 2011). Mechanistically, CNV formation is believed to be largely distinct from smaller indels, and is proposed to arise from either microhomology mediated end joining or non-allelic homologous recombination, occurring at either DSBs or stalled replication forks (Halliwell et al., 2021; Hastings et al., 2009). Similarly, chromosomal translocations, observed particularly frequently at chromosome 1q in hPSC (Figure 1.2), require DSB formation at two genomic sites, followed by unfaithful repair to generate juxtaposition of chromosomal material from one region of the genome to another (Roukos and Misteli, 2014).

Whole chromosome aneuploidies, also frequently observed in hPSC cultures (Draper *et al.*, 2004; Baker, Adam J. Hirst, *et al.*, 2016), can arise from a number of well documented mitotic errors, including incorrect or absent assembly of mitotic spindles on kinetochores (Gordon *et al.*, 2012). Chromosomal mis-segregation events are more frequent under conditions of DNA replication stress (see section 1.4.3.2) (Burrell *et al.*, 2013; Halliwell *et al.*, 2020; Böhly *et al.*, 2022) and recent CRISPR studies have demonstrated an increased mitotic error rate on chromosomes harbouring DSBs (Turocy *et al.*, 2022). The contribution of replication stress and DSBs to whole chromosome aneuploidies is likely due to incompletely replicated or indeed unresolved HR intermediates linking sister chromatids (Mankouri *et al.*, 2013).

In summary, whilst mutation origins are mechanistically diverse, DNA damage and replication errors or stress are unifying features in their biogenesis.

1.3.3. Consequences of genetic variants in hPSC cultures

The presence of genetically variant cells in hPSC cultures can affect every previously discussed application of hPSC.

In drug screening applications genetically variant hPSCs or their differentiated derivatives may confound results. The driver gene(s) in many genetic variants remains unknown, however several recurrent genetic variants endow cells with a reduced propensity to undergo apoptosis. For example, the *BCL2L1* gene located on the 20q11.21 amplicon, commonly duplicated in hPSC cultures, encodes the BCL-XL protein which antagonises BAX-mediated apoptosis (Avery et al., 2013; Amps et al., 2011). Similarly, loss of function *TP53* mutations, recurrently observed in hPSC interferes with regulation of apoptosis, likely conferring apoptotic resistance to cells (Merkle *et al.*, 2017; Avior *et al.*, 2021). A heightened apoptotic threshold in such cells is likely to affect sensitivity to cytotoxic compounds.

Certain genetically variant hPSC display altered or hindered differentiation capacities. Cells harbouring an additional copy of chromosome 17q, exhibit reduced propensity to form mesoderm upon undirected differentiation (Fazeli *et al.*, 2011). More recently, hPSC harbouring the highly recurrent 20q11.21 duplication were shown to have reduced differentiation efficiencies when compared to wild-type cells in neuroectodermal differentiations (Markouli *et al.*, 2019). Moreover, hPSC with isochromosome 20q (i.e. gain of 20q copy, loss of 20p), fail to induce expression of germ layer-associated genes following five days undirected differentiation (Vitillo *et al.*, 2023). Improper differentiation renders such cells unsuitable for developmental models.

For regenerative medicine applications, a more pressing concern than the efficacy of the cell product, is safety of a product derived from genetically variant hPSC. Many of the genetic changes observed in hPSC cultures are also found in various cancers (Table 1.2), moreover certain variants have demonstrated failure to differentiate completely in teratoma assays (Herszfeld *et al.*, 2006; Werbowetski-Ogilvie *et al.*, 2009). Teratomas derived from these cells retain a core of pluripotent cells which can be cultured *ex vivo* as seen in embryo carcinomas, indicating possible neoplastic potential (Herszfeld *et al.*, 2006; Werbowetski-Ogilvie *et al.*, 2009; Ben-David *et al.*, 2014). This uncertain oncogenic potential of genetically variant hPSCs precludes many of them from cell therapy applications, particularly where the genetic change in question has documented effect on oncogenic transformation in cancer. For instance, a recent clinical trial of the use of hPSC-derived retinal pigment epithelium for treatment of macular degeneration was terminated prematurely on account of a mutation of unknown consequence in the chosen cell line (Garber, 2015).

Table 1.2 Genetic changes in hPSC and their counterparts in cancer

The most commonly reported genetic variants in hPSC and some notable incidences in cancer. NB references listed are non-exhaustive.

hPSC Recurrent genetic change	**Example in Cancer**
Chr1(q) Gain	**Burkitt lymphoma** (Salaverria *et al.*, 2008) **Multiple myeloma** (Sawyer *et al.*, 1995)
Chr12(p) gain	**Testicular germ cell** (Castedo, 1993) **Breast cancer** (Natrajan *et al.*, 2009)
Chr17(q) gain	**Testicular germ cell cancer** (Skotheim *et al.*, 2002) **Gastric cancer** (Kokkola *et al.*, 1997) **Cervical cancer** (Harris *et al.*, 2003) **Neuroblastoma** (Lastowska *et al.*, 1997)
Chr18(q) deletion	**Cervical cancer** (Harris *et al.*, 2003) **Prostate cancer** (Yin *et al.*, 2001) **Pancreatic cancer** (Hahn *et al.*, 1996) **Breast cancer** (Cropp *et al.*, 1990) **Lung cancer** (Sameshima *et al.*, 1994)

Chr20(q11.21) Gain	**Cervical** (Harris *et al.*, 2003) **Olfactory neuroblastoma** (Guled *et al.*, 2008)
TP53 mutation	**Many (most frequently mutated gene in cancer)** (Chen *et al.*, 2022)
CDK12 mutation	**Breast cancer** (Naidoo *et al.*, 2018) **Prostate cancer** (Reimers *et al.*, 2020) **Gastric cancer** (Ji *et al.*, 2019) **Ovarian cancer** (Bell *et al.*, 2011)
EGFR mutation	**Lung cancer** (Bethune *et al.*, 2010) **Glioblastoma** (Xu *et al.*, 2017) **Breast cancer** (Teng *et al.*, 2011)
PATZ mutation	**Lung cancer** (Lucà *et al.*, 2023) **Thyroid cancer** (Chiappetta *et al.*, 2015)
BCOR mutation	**Acute myeloid leukaemia** (Grossmann *et al.*, 2011) **Endometrial carcinoma** (García-Sanz *et al.*, 2017)

In summary, hPSC acquire and expand recurrent genetic changes over the course of culture. These variants may affect the differentiation capacity and ultimate utility of hPSC but more worryingly may have oncogenic capacity, precluding them from regenerative medicine applications. Whilst the mechanisms of initial mutation are diverse, they are often preceded by DNA damage, the varieties, and sources of which are discussed in the following section.

1.4. DNA double-strand breaks and their origins

DNA damage takes diverse forms, including nucleotide base damage, inter-strand cross links, single-strand breaks (SSBs) and double-strand breaks (DSBs) (Chatterjee and Walker, 2017). The most deleterious of these lesions and the primary focus of this thesis are DSBs. Upon formation of a DSB the cell activates the DNA damage response (DDR) which broadly results in one of three outcomes. In a best-case scenario, activation of the DDR recruits components of DNA damage repair pathways facilitating faithful repair of the DSB (Ceccaldi *et al.*, 2016).

Alternatively, the level of DDR signalling, mediated by Ataxia-telangiectasia mutated protein (ATM), passes the threshold for apoptosis, resulting in an accumulation of nuclear P53 protein irreversibly activating apoptosis (Chen, 2016). A third possibility is erroneous repair of the DSB which can directly lead to the formation of indels (Cisneros-Aguirre *et al.*, 2022; Min *et al.*, 2023), CNVs (Hastings *et al.*, 2009), translocations (Roukos and Misteli, 2014) and whole chromosome copy number changes (Mankouri *et al.*, 2013). DSBs can also indirectly contribute to SNV formation, whereby ssDNA, generated as an intermediate in homologous repair (HR), exhibits increased vulnerability to hydrolytic base deamination (Roberts *et al.*, 2012). Thus, DSBs pose a threat to both the viability and genetic stability of cells.

Factors intrinsic to processes within the cell, as well as exogenous factors can contribute to DSB formation in hPSC and indeed other cell types. Ahead of discussing sources of DSBs in cells, it is important to note that mechanisms of DSB formation cannot be entirely disentangled from mechanisms giving rise to other types of DNA damage. Notably, single strand breaks (SSBs) in close proximity to one another on opposing strands, yield a DSB (Cannan and Pederson, 2016) (Figure 1.4 A). In addition, replication of DNA harbouring an unrepaired SSB or any number of base lesions or DNA adducts may hinder DNA replication, potentially leading to DSB formation (Zeman and Cimprich, 2014), discussed in more detail in section 0 (Figure 1.4 B).

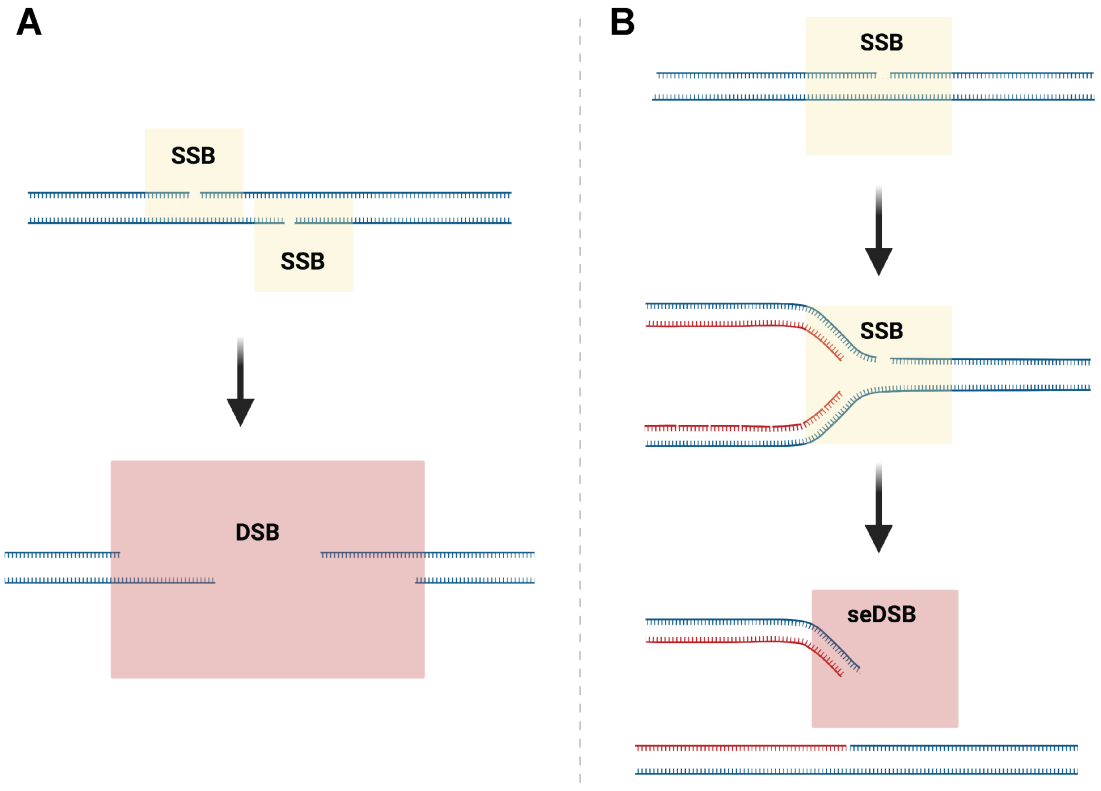


Figure 1.4 DSB formation from SSBs

A) Two SSBs on opposing DNA strands may yield a DSB if the melting temperature of the intervening stretch of DNA is sufficiently low for dissociation of complementary strands. B) An SSB may be processed to a single-ended DSB (seDSB) on encounter with a replication fork. SSB on the leading strand shown here, lagging strand SSBs also yield DSBs.

1.4.1. Environmental sources of DNA damage

DSBs can be induced by a huge range of genotoxic chemicals as well as radiation sources. However, here, I omit sources unlikely to significantly impact on the hPSC genome over routine *in vitro* culture, such as UV or ionising radiation, and chemotherapeutic agents.

1.4.1.1. Low-pH environment as a cause of DNA damage

Low pH environments can cause DNA damage in human cells. Acidification of cell culture medium can occur due to improper culture medium formulation, excessive CO_2 concentrations, or inadequate frequency or volume of media replenishment (Jacobs et al., 2016a; Liu et al., 2018). This latter mechanism is due to accumulation of lactate produced by anaerobic respiration, and thus cannot be considered a purely exogenous source of DNA damage. Culture medium acidification has been demonstrated to yield increased levels of DNA damage, including DSBs in hPSC, however the mechanistic basis of this damage remains unclear (Jacobs *et al.*, 2016; Liu *et al.*, 2018). Low-pH has previously been associated with increased DNA damage burden in the tumour microenvironment, Yuan and colleagues assayed the ability of cells cultured under low pH and hypoxia to repair plasmid DNA and found them to exhibit a significant reduction in repair capacity, proposed to result from improper protein folding under low pH (Yuan *et al.*, 2000). Consistent with a model of acid-interference with protein function, studies into the mechanism of acid-induced DSBs in breast cancer cells find such damage to be dependent on TOP2 enzymes (see section 1.4.2.2), proposing a model whereby, at low pH, TOP2 cleaves but fails to re-ligate DNA (Xiao *et al.*, 2003). Low pH exposure, from acid reflux is a risk factor for oesophageal adenocarcinoma development. In contrast with a TOP2-mediated model, studies into the effect of acid exposure on epithelial cells demonstrate that DSB induction is dependent on intracellular reactive oxygen species (ROS) production following acidification, implicating oxidative damage as the downstream effector of low pH (Zhang *et al.*, 2009; Zhao *et al.*, 2021). Finally, spontaneous de-purination of DNA is enhanced at low pH (An *et al.*, 2014). The resulting abasic sites are typically repaired by the base excision repair (BER) pathway which generates intermediate single strand breaks (Kitsera *et al.*, 2019), which in turn may be processed to DSBs (Caldecott, 2008).

1.4.1.2. Oxidative stress as a source of DNA damage

Reactive oxygen species (ROS) comprise several oxidising agents including superoxide anions, hydrogen peroxide, and nitric oxide (Slimen *et al.*, 2014). Mechanistically, ROS can cause DNA damage via several mechanisms. DNA bases, most commonly guanine, are readily oxidised by ROS, which can yield SNVs if unrepaired, alternatively repair via the BER pathway generates SSB intermediates (Lindahl *et al.*, 1997). Oxidised DNA bases can also pose a block to DNA replication forks (Sedletska *et al.*, 2013), potentially yielding replication-associated damage (see section 1.4.3).

Largely generated as a by-product of cellular aerobic metabolism, ROS could be considered an endogenous source of DNA damage, however ROS equilibrium is, to a significant extent, affected by external factors, most notably media formulation and external oxygen concentration. Human PSC, like most mammalian cell lines, are routinely cultured at ~20% oxygen, representing a 10-20 fold increase on the oxygen tension likely experienced in the stem cell niche *in vivo* (Jauniaux *et al.*, 1999). At atmospheric oxygen concentrations, photo reactivity of riboflavin in cell culture media can generate ROS, in the absence of cells (Grzelak *et al.*, 2001), moreover atmospheric oxygen increases cellular ROS production in various cell types (Boregowda *et al.*, 2012; Nasto *et al.*, 2013).

ROS levels in cells, and indeed culture medium, are governed by the rate at which ROS are produced versus the rate at which they are inactivated. Media formulation can affect both the rate of ROS formation and quenching. Whilst hPSC rely predominantly on anaerobic glycolysis for ATP production (Varum *et al.*, 2011), modern media formulations have been show to increase levels of oxidative phosphorylation in hPSC accompanied by increased ROS production and DNA damage (Bangalore *et al.*, 2017). Culture medium can also buffer ROS via inclusion of ROS scavengers such as albumin, which is able to bind and inactivate certain ROS via a free thiol group (Taverna *et al.*, 2013). Indeed light-mediated ROS generation in culture media at atmospheric oxygen is largely mitigated by inclusion of serum in media (Grzelak *et al.*, 2001). In spite of this, modern hPSC media formulations have been simplified dramatically in recent years, with preferences for defined culture media with limited and characterised components for batch to batch consistency and ease of translation to clinic (Amit *et al.*, 2004; Ludwig *et al.*, 2006; Akopian *et al.*, 2010; Bergström *et al.*, 2011). The popular *E8* medium for example, has been reduced to the “essential 8” components required to support cells in an undifferentiated state. E8 has a protein content of less than 30 mg l^{-1} and, notably, lacks albumin (Chen *et al.*, 2011).

Culture of hPSC under atmospheric oxygen is accompanied by an apparent increase in DNA damage when compared to low oxygen culture (Forsyth *et al.*, 2006; Guo *et al.*, 2013). A

recent whole-genome sequencing study from our lab revealed that hPSC exhibit a mutational signature indicative of oxidative DNA damage, moreover, low oxygen culture significantly reduces mutation rate (Thompson *et al.*, 2020). Thus, it is highly likely that oxidative stress contributes to the genome damage burden of hPSC.

1.4.2. Endogenous sources of DNA damage: replication independent

Routine DNA transactions such as transcription, and regulation thereof can pose a threat to genome stability and require fastidious regulation to prevent breakage of DNA.

1.4.2.1. Active DNA demethylation

As discussed above, oxidation of DNA bases is readily repaired via the BER pathway, generating a transient SSB in the process. Similarly, oxidised methyl-cytosine derivatives can act as a substrate of the BER pathway generating DNA damage in the process. As a means of epigenetic regulation, DNA, most commonly cytosine at CpG dinucleotides, is methylated by de-novo methyl transferase enzymes, yielding 5-methyl-cytosine, conferring a transcriptionally repressive state (Moore *et al.*, 2012). In the reverse process, 5-methyl-cytosine can be actively demethylated, during which 5-methyl-cytosine is iteratively oxidized by the ten-eleven translocase enzyme to 5-hyroxy-methyl-cytosine, 5-formyl-cytosine and 5-carboxyl-cytosine. 5-carboxyl-cytosine and 5-formyl-cytosine are actively excised from DNA by thymine DNA glycosylase, yielding an abasic site which is repaired by via the BER pathway (He *et al.*, 2011; Maiti & Drohat, 2011). As such, active demethylation can induce DNA damage, indeed a recent study in post-mitotic neurons identified that cytosine residues in active enhancers are hotspots of SSB formation and accompanied repair via the BER pathway, moreover these sites overlapped significantly with regions of 5-hyrdoxy-methyl-cytosine and 5-formyl-cytosine accumulation, suggesting active demethylation induces breakage (Wu et al., 2021). A follow-up study by the same group demonstrated that genetic ablation of thymine DNA glycosylase eliminates SSB formation and associated DNA repair at enhancers, and concurrently interferes with cell specification (Wang et al., 2022). These studies demonstrate that active demethylation at enhancer regions is required for lineage-specific transcriptional activation but is capable of concurrent SSB induction.

1.4.2.2. Topoisomerase activity

The action of endogenous topoisomerase enzymes, predominantly TOP1 and TOP2 in mammalian cells also pose a threat to genome integrity. Topoisomerase enzymes resolve

diverse DNA structures, typically generated by replication, transcription, or chromatin remodelling (Pommier *et al.*, 2016). Mechanistically, TOP1 and TOP2 induce transient breaks on one or both strands respectively of the DNA backbone, allowing rotation of DNA strands relative to one another and/or decatenation of linked DNA duplexes. Cleavage of the phosphodiester backbone is typically followed by rapid re-ligation (Pommier *et al.*, 2016), however this re-ligation is not always effective, with TOP1 and TOP2 frequently carrying out abortive catalysis, yielding more permanent SSBs and DSBs respectively (Morimoto *et al.*, 2019).

Topoisomerase-mediated DNA breakage has previously been shown to facilitate transcription. During transcription, RNAPII processes along the DNA template generating positive supercoiling ahead of the transcription complex and negative supercoiling behind (Liu & Wang, 1987). Excessive supercoiling can prevent RNAPII processivity and induce formation of non-b-form DNA (Ma and Wang, 2016). Both TOP1 and TOP2 act to alleviate supercoiling arising from active transcription (Kouzine *et al.*, 2013) (Figure 1.5 A). Topoisomerase-mediated DSB induction is also required for expression of certain genes. Pioneering work in MCF7 cells found that TOP2 generation of persistent DSBs is required for active transcription of estrogen-responsive genes (Ju *et al.*, 2006), and subsequent studies have identified that promoter-proximal TOP2 cleavage of DNA activates gene transcription in diverse cell types (Pommier *et al.*, 2022). Similarly, TOP1 has been shown to associate with RNAPII, activating upon canonical pause release, to alleviate positive supercoiling and facilitate transcriptional elongation (Baranello *et al.*, 2016). Moreover TOP1 cleavage at enhancer regions is necessary for enhancer RNA (eRNA) expression, and high-resolution DSB mapping studies have found DSBs recur at active enhancers upon treatment of cells with the selective TOP1 poison, CPT (Hazan *et al.*, 2019).

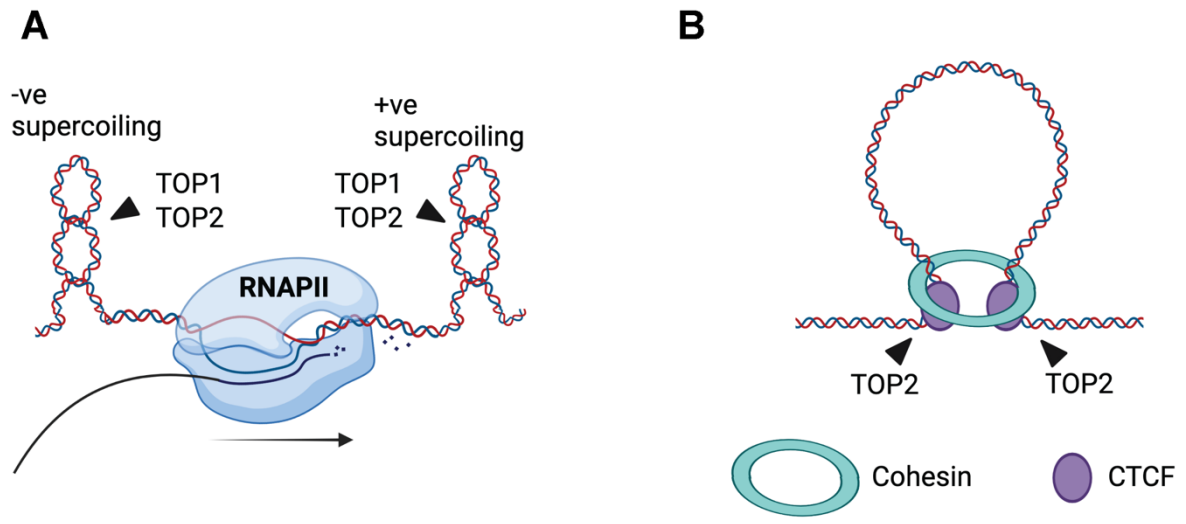


Figure 1.5 Topoisomerase mediated breaks

A) Active transcription generates positive and negative supercoiling ahead of and behind RNA polymerase, supercoiling is alleviated via enzymatic activity of TOP1 and TOP2 enzymes. B) DNA is extruded through cohesin rings to generate loop structures, the base of which is tethered by CTCF. TOP2 enzymes cleave non-B-form DNA arising at loop boundaries.

Topoisomerase enzymes also act in extrusion of chromatin loops through cohesin rings. To facilitate interactions of distal regulatory elements, such as promoters and enhancers, DNA is organised into domains. One mechanism of organising the genome in 3D, involves extrusion of a DNA loop through a ring-like protein complex, cohesin. The boundaries of such loops are dictated by binding of the CCCT-binding factor (CTCF) (Rao *et al.*, 2014). TOP2 enzymes associate with cohesin and are proposed to resolve DNA structures refractory to loop extrusion (Uusküla-Reimand *et al.*, 2016) (Figure 1.5 B). Several DSB-mapping studies have noted an enrichment of DSBs at CTCF binding sites (Tchurikov *et al.*, 2015; Canela *et al.*, 2017, 2019; Gothe *et al.*, 2019; Szlachta *et al.*, 2020), with two computational studies independently finding CTCF motifs to be amongst the best predictors of DSB formation (Mourad *et al.*, 2018; Ballinger *et al.*, 2019). In 2017, Canela and colleagues demonstrated that TOP2 induces DSBs at loop boundaries marked by cohesin and CTCF, and interestingly found these sites to be hotspots of chromosomal rearrangements in leukaemia, implicating TOP2 activity in oncogenic mutagenesis (Canela *et al.*, 2017, 2019).

In summary, topoisomerase enzymes act at various sites genome wide, depending on function. Whilst typically catalysing faithful re-ligation of the DNA template, abortive catalysis can yield genotoxic DSB formation, the presence of which coincides with sites of recurrent structural variation in cancer.

1.4.2.3. R-loops

Transcription can induce genome damage via the action of topoisomerase enzymes (discussed above) or alternatively though interference with DNA replication machinery (discussed in detail in section 1.4.3). However, transcription can also induce DNA damage independent of DNA replication, mediated by R-loops (Sollier and Cimprich, 2015). R-loops are three-stranded nucleic acid structures, which occur when nascently transcribed RNA re-anneals to its template DNA strand, displacing the non-template strand as ssDNA (Figure 1.6 A). R-loops range in length from a few base pairs to several kilobases, and are highly prevalent in the mammalian genome, forming at 5-10% of regions genome-wide (Brickner *et al.*, 2022). R-loops have diverse physiological roles including regulation of transcriptional initiation and termination as well as acting as intermediates in the DNA damage response (Niehrs and Luke, 2020). R-loop metabolism is believed to be tightly regulated, however disruption to this via interference with chromatin structure, transcriptional activity, or expression of key factors involved in R-loop resolution, can lead to formation of genotoxic, unscheduled R-loops (Brickner *et al.*, 2022). The ssDNA component of R-loops is vulnerable to deamination, indeed during somatic hypermutation, physiological R-loops facilitate programmed cytidine deamination on the displaced ssDNA strand in B cells (Basu *et al.*, 2011). As discussed in the previous section, ssDNA is also vulnerable to unprogrammed DNA damage, and a recent study in melanoma cells found dysregulation of R-loops leads to SSB formation, proposed to arise via enzymatic or spontaneous deamination of the displaced ssDNA moiety of R-loops (Figure 1.6 B) (Safari *et al.*, 2021).

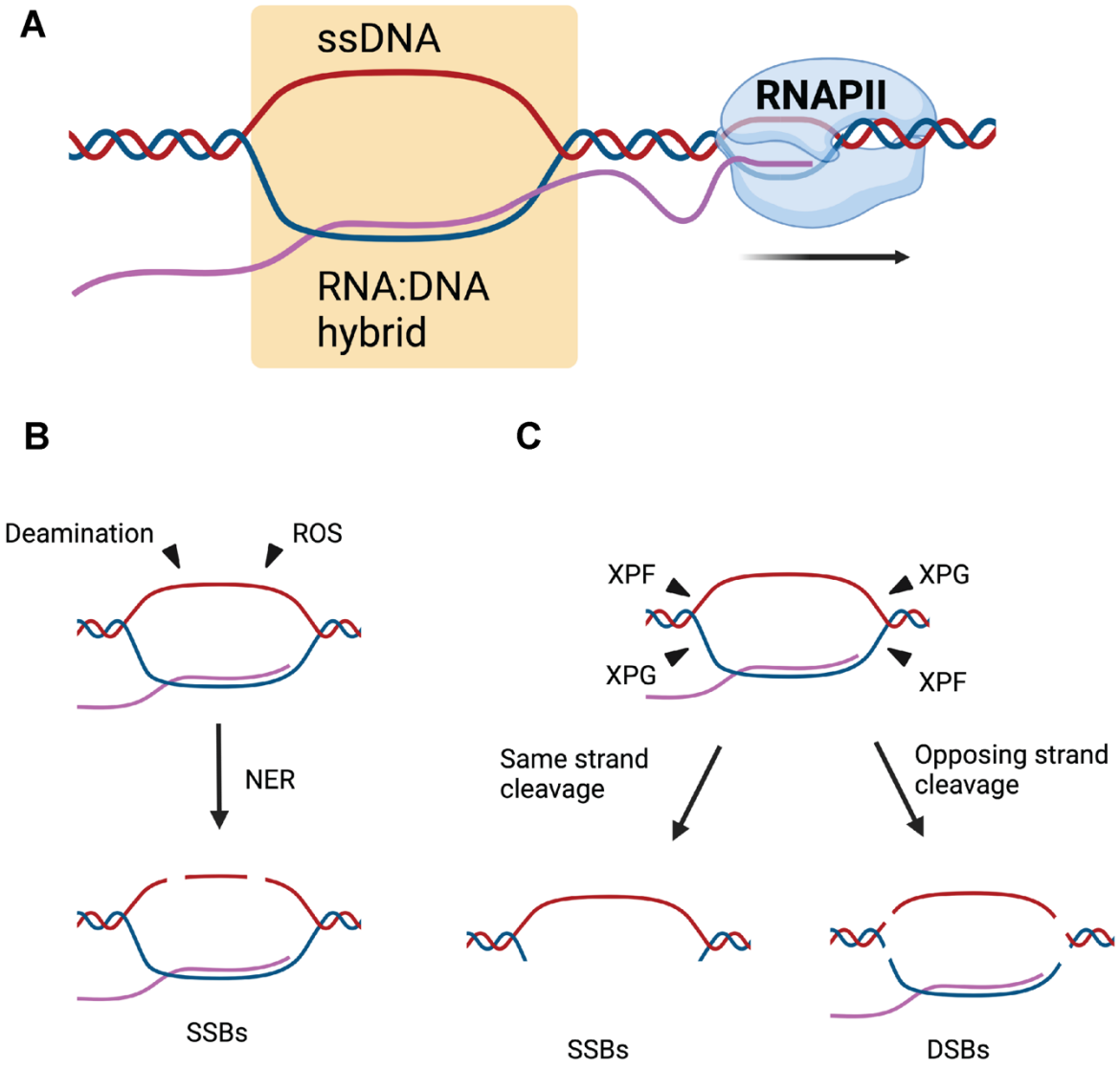


Figure 1.6 Replication-independent R-loop-mediated DNA damage

A) R-loops form when nascent RNA transcripts (purple) re-anneal to the template DNA strand (blue) outside of the transcription bubble, displacing the non-template strand as ssDNA. The R-loop structure (highlighted in the yellow box) comprises both DNA:RNA hybrid and ssDNA. B) The ssDNA moiety of the R-loop is vulnerable to oxidative damage and enzymatic or spontaneous base deamination. Repair of the resulting DNA structures by the NER pathway yields intermediate SSBs. C) Flap endonuclease enzymes, XPF and XPG, induce SSBs in the DNA flap structures flanking the R-loop. SSBs generated on the same strand of DNA generate SSBs and may have a physiological role in R-loop resolution. SSBs generated on opposing strands yield DSBs.

R-loops can also act as substrates for endogenous nucleases. Based on observations that the flap endonuclease enzymes XPF and XPG could cleave R-loop structures in vitro (Tian and Alt, 2000), Sollier and colleagues investigated whether the same was true *in vivo* and found that unscheduled R-loops, arising from diverse mechanisms, induced DSB formation dependent on the expression of XPF/XPG (Figure 1.6 C) (Sollier *et al.*, 2014). Interestingly, subsequent studies have found R-loops, cleaved by XPF and XPG occurring at common fragile sites (Jurga *et al.*, 2021), and sites of recurrent chromosomal rearrangements in breast

cancer (Stork *et al.*, 2016), suggesting processing of pathological R-loops may drive chromosomal instability.

1.4.3. Endogenous sources of DNA damage: replication-dependent

DNA replication in proliferative cells requires faithful duplication of the entire genome exactly once per cell cycle. Whilst largely successful, errors occurring during DNA replication can yield mutagenic DNA damage and recombination events. Such errors can be more frequent under conditions known as replication stress, as discussed below.

1.4.3.1. Eukaryotic DNA replication

Much of the understanding of DNA replication in humans is drawn from studies in yeast. During unperturbed eukaryotic DNA replication, pre-replication complexes are loaded onto the DNA exclusively during G1 phase of the cell cycle in a process known as origin licensing. Pre-replication complexes comprise two core replicative helicase hexamers, MCM2-7 bound to chromatin in opposing orientations which facilitate bidirectional DNA replication from each replication origin (Yuan and Li, 2020). In humans, a surplus of replication origins are licensed (Woodward *et al.*, 2006), the positions of which are not strictly defined by DNA sequence (Schaarschmidt *et al.*, 2004; Akerman *et al.*, 2020), rather by epigenetic status of chromatin, thus origin position and usage is not universal across cell types (Cayrou *et al.*, 2011). During origin licensing, proteins of a six-subunit complex known as the origin recognition complex (ORC) are recruited to DNA in a step-wise manner (Li and Stillman, 2012). Whilst certain subunits of ORC may remain bound to chromatin throughout the cell cycle, complete ORC binding is restricted to G1 phase, regulated by proteasomal degradation of specific subunits during S-phase (Méndez *et al.*, 2002). Chromatin-bound ORC is bound and activated by cdc6 during G1, which complexes with ORC to encircle the DNA (Speck *et al.*, 2005). The resulting ORC-cdc6 complex recruits the MCM2-7 core hexamer of the replicative helicase via the cdt1 chaperone protein, following which cdt1 dissociates and a second MCM2-7 hexamer is loaded, forming an MCM2-7 double hexamer in a head-to-head orientation (Fernández-Cid *et al.*, 2013; Miller *et al.*, 2019). By stringent restriction of origin licensing to the G1 phase, cells limit the opportunity for re-duplication of DNA.

During S-phase, functional replicative helicases are formed from pre-replication complexes and are activated in a process known as origin firing. In this process, MCM2-7 associates with GINS and cdc45 to form the intact replicative helicase CMG, and, through interaction with several proteins known as firing factors, each CMG helicase extrudes the lagging DNA strand,

allowing helicases to pass one another, unwinding the DNA template in a bidirectional fashion (Botchan and Berger, 2010). Replication of unwound DNA on both leading and lagging strands is initiated by the pol α-primase complex which synthesises a short RNA primer followed by ~20nt of DNA, this primer is then bound by proliferating cell nuclear antigen (PCNA), and either pol ε on the leading strand or pol δ on the lagging strand, which extend DNA synthesis in 5'-3' orientation (Johansson and MacNeill, 2010). During unperturbed replication, DNA polymerase enzymes associate with the CMG assembly, collectively forming the "replisome" and leading strand synthesis occurs in a continuous fashion, whereas in lagging strand synthesis is discontinuous, achieved by iterative cycles of priming and pol δ synthesis of ~100nt Okazaki fragments (Garg and Burgers, 2005).

Provided the replication fork does not encounter a block, it will proceed until it converges with a second oncoming replication fork, to terminate replication. In this process CMG helicases pass one another on their respective leading strands and leading strands of opposing forks complete synthesis of their counterpart fork's lagging strand, thus completing DNA replication between two origins (Dewar *et al.*, 2015).

1.4.3.2. Replication stress

Replication stress is the term given to perturbed DNA replication, manifesting in low fork speed, fork stalling, accumulation of ssDNA and DSB formation (Gaillard *et al.*, 2015). Replication stress can arise through various mechanisms, for instance scarcity of metabolites, the presence of unrepaired DNA lesions or interference with transcription (Zeman and Cimprich, 2014).

Lack of deoxynucleotide triphosphates (dNTPs) can cause replication stress, which is readily rescued by supplementation with exogenous nucleosides (Bester *et al.*, 2011). Mechanistically, dNTPs can be depleted by rapid proliferation and premature entry into S-phase upon overexpression of certain oncogenes (Bester *et al.*, 2011). Alternatively, a lack of precursor metabolites can induce dNTP shortage and associated replication stress, as demonstrated in a study by Lamm and colleagues, where chronic deprivation of folate, a dNTP precursor, lead to replication stress in fibroblasts (Lamm *et al.*, 2015). Interestingly dNTP shortage has been shown to increase the frequency chromosome copy number changes and structural variants in colorectal cancer cells (Burrell *et al.*, 2013). Moreover, we recently identified decreased replication stress in hPSC cultures supplemented with nucleosides, raising the possibility that replication stress, driven by a paucity of dNTPs, directly contributes to genetic variation in hPSC (Halliwell *et al.,* 2020).

Replication stress is also readily induced by the presence of DNA lesions. As discussed above, SSBs can yield DSBs upon encounter with DNA replication forks (Figure 1.4 B). SSBs on the leading strand, lead to passive run-off of the CMG helicase and therefore collapse of the replication fork (Strumberg *et al.*, 2000). Recently however, it was also found that lagging strand SSBs generate fork collapse, whereby CMG helicase bound to the leading strand encircles the lagging strand downstream of the SSB, and processes along dsDNA, akin to transcription termination, where it is actively unloaded via proteasomal degradation (Vrtis *et al.*, 2021).

Whilst SSBs lead to rapid replisome disassembly, other lesions can cause a physical block to replication forks, resulting in stalling. Bulky DNA adducts, occurring on exposure to UV radiation or chemotherapeutic drugs are potent replication fork blocks (Wang *et al.*, 2001). However due to low UV exposure and the absence of chemotherapeutic agents, these are unlikely to be prevalent in the cultured hPSC genome. Ribonucleotides, are erroneously incorporated into nascent DNA at a rate of 1 in 1500 and 1 in 5000 for pol ε and pol δ respectively (McElhinny et al., 2010). In the following S-phase, mis incorporated ribonucleotides induce replication fork stalling (Lazzaro *et al.*, 2012). Additionally, secondary DNA structures, such as hairpins (Anand *et al.*, 2012), z-form DNA (Brinton *et al.*, 1991), H-form DNA (Krasilnikova and Mirkin, 2004) and G-quadruplexes (Paeschke *et al.*, 2011), often arising at low-complexity, repetitive sequences, have all been shown to impede fork progression. Importantly, many of these non-B-form DNA structures involve intra-strand base pairing and are more energetically favourable in the context of ssDNA. Replication fork stalling can cause uncoupling of the replicative helicase from the replication fork, leading to excess ssDNA generation (Byun *et al.*, 2005). This favours secondary DNA structure formation, increasing the likelihood of further fork stalling at secondary structures, further exacerbating replication stress.

Transcription is frequently implicated as a cause of replication stress through physical collision with replication forks. Transcription and replication use the same DNA template making interference between them near-inevitable (Helmrich *et al.*, 2011). However, both prokaryotic and eukaryotic cells regulate transcription and replication timing to spatiotemporally separate these two processes, mitigating replication stress arising from collisions (Meryet-Figuiere *et al.*, 2014). In particularly long genes, which take more than one iteration of the cell cycle to complete transcription, concurrent transcription and replication may be inevitable. A study in B-lymphoblasts found very long genes to exhibit chromosomal fragility in a manner dependent on their expression (Helmrich *et al.*, 2011). At a broader scale, global upregulation of transcription via oncogene overexpression is accompanied by increased replication fork stalling and DNA damage (Kotsantis *et al.*, 2016).

The orientation of collision between replication and transcription complexes affects genotoxicity. Using episomal models of co-directional and head-on transcription and replication in NHEK cells, Hamperl and colleagues were able to demonstrate that collisions between replication and transcription in a head-on orientation were more genotoxic than co-directional collisions (Hamperl *et al.*, 2017). Accordingly, at a genomic level, unperturbed origin firing occurs preferentially in a co-directional orientation with active transcription (Petryk *et al.*, 2016). Induction of replication stress however, increases interactions between RNAPII and the replisome and associated damage, suggesting dormant origin activation induced to complete DNA replication, under stressed conditions, leads to an increased frequency of head-on collisions (Hamperl and Cimprich, 2016). Recent work identified a role in chromatin loop extrusion in maintaining spatiotemporal separation of transcription and replication. Using an inducible cohesin degron system, the study found that loop extrusion repositions pre-replication complexes out of transcriptionally active chromatin loops, dysregulation of which leads to increased fragility proposed to stem from an increase in transcription-replication conflicts (Wu et al., 2023).

R-loops are often implicated in replication stress mediated by transcription-replication conflicts, as demonstrated by phenotype rescue using overexpression of the RNA:DNA hybrid-specific RNase H enzyme (Helmrich *et al.*, 2011; Kotsantis *et al.*, 2016; Hamperl *et al.*, 2017). Whilst R-loop presence is not strictly necessary for induction of transcription-mediated replication stress, the presence of the hybrid is believed to stabilise RNAPII on the DNA template, forming a more potent block to DNA replication (Hamperl *et al.*, 2017).

1.4.3.3. Replication stress DSB mechanisms

One of the unifying features of replication stress is the frequent stalling of replication forks. Stalled forks are often cleaved or remodelled to generate intermediates for fork restart. The leading strand template of a stalled fork can act as a substrate for the structure-specific 3' endonuclease MUS81-EME2 (Hanada *et al.*, 2007; Pepe and West, 2014) (Figure 1.7 B). Equally, a stalled fork's lagging strand template can be cleaved by the 5' endonuclease EEPD1 (Kim et al., 2017) (Figure 1.7 B). Each of these scenarios yields a potentially mutagenetic single-ended double strand break (seDSB), the purpose of which is to prompt fork restart via homologous recombination. In this process seDSBs undergo 5' resection, mediated by EXO1 (Wu *et al.*, 2015) to generate extended stretches of ssDNA. RAD51 then mediates invasion of the ssDNA into the newly synthesized homologous sister chromatid, prompting fork restart (Figure 1.7 B) (Nickoloff *et al.*, 2021) . In the case of R-loop-induced replication stress, a recent study identified a non-canonical function of MUS81-cleavage,

whereby supercoiling arising between head-on transcription and replication induces fork stalling. MUS81 DNA cleavage relieves this supercoiling, following which seDSBs are re-ligated via non-homologous end joining (NHEJ), allowing passage of the transcription complex followed by fork restart (Chappidi *et al.*, 2020).

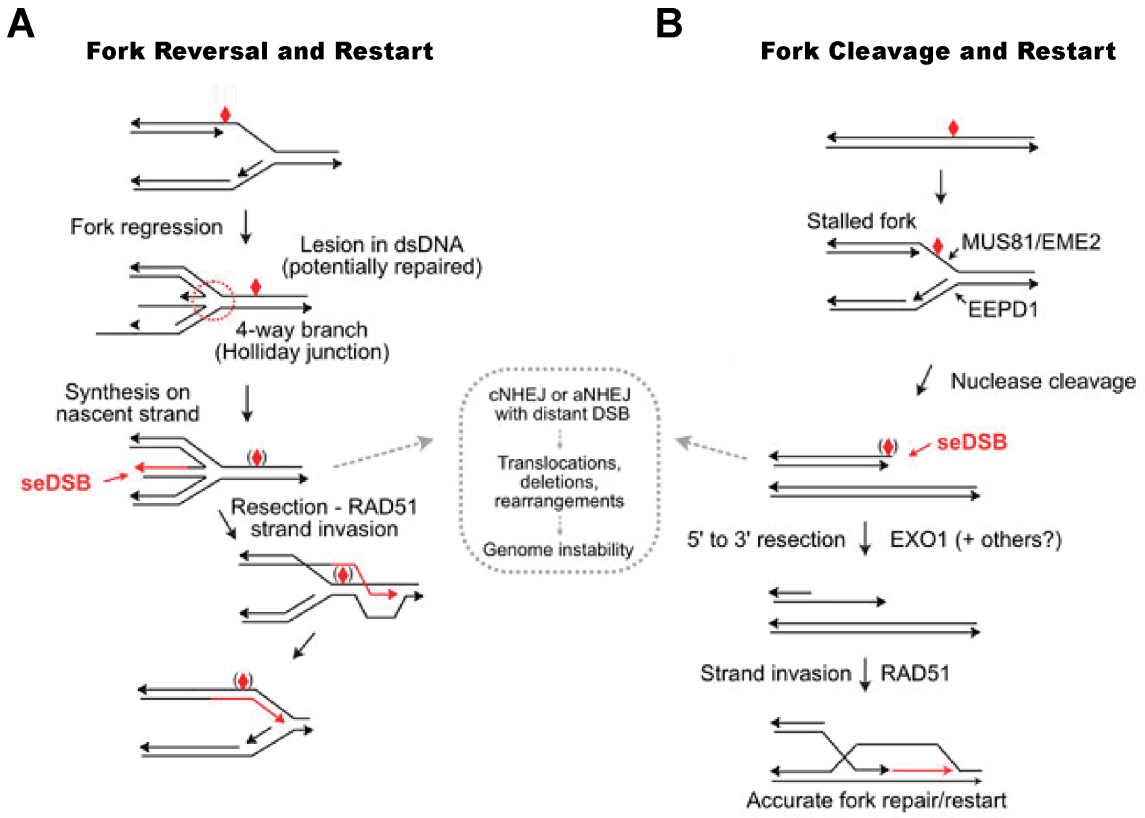


Figure 1.7 seDSB generation for replication fork restart

A) Replication fork stalls upon encounter with a barrier (red diamond) on the leading strand template. Fork reversal is catalysed by fork re-modellers, newly synthesized strands anneal and regress from the site of stall. Leading strand synthesis occurs using the lagging strand as a template, following which the lagging strand is resected, facilitating RAD51-mediated strand invasion of the leading strand and fork restart beyond the site of stalling. B) Replication fork stalls, and cleavage occurs on either the leading or lagging strands, catalysed by MUS81-EME2 or EEPD1 respectively. The resulting seDSB is resected, then undergoes RAD51-mediated strand invasion on the newly synthesised sister chromatid, facilitating fork restart. If resection and subsequent HR fails, seDSBs generated via either mechanism can act as substrates for mutagenic NHEJ, leading to potential genomic instability. Figure adapted from (Nickoloff *et al.*, 2021).

An alternative mechanism for DSB generation at stalled forks is via a process known as fork reversal. When forks stall upon encountering a barrier on the leading strand, the replicative helicase uncouples from the fork and extensive ssDNA is generated. This ssDNA is bound by RPA which, in turn, can recruit remodelling enzymes such as SMARCAL1, to catalyse fork reversal (Bétous *et al.*, 2012, 2013). In fork reversal, nascent strands anneal and regress from the site of the stall, forming a 'chicken foot' structure (Qiu *et al.*, 2021). The end of a reversed fork constitutes a seDSB, which can be resected and act as a substrate for homologous repair and replication fork restart (Figure 1.7 B) (Qiu *et al.*, 2021).

In instances where stalled forks are not rescued by either restart, or convergence of a neighbouring replication fork, incompletely replicated DNA may persist into mitosis, resulting in covalently linked sister chromatids (Mankouri *et al.*, 2013). MUS81-EME1 cleaves persistent repair intermediates to allow proper chromosome segregation, indeed in the absence of MUS81, cells subjected to replication stress show an increased rate of mis-segregation events (Ying et al., 2013).

Thus, whilst DSBs generated at sites of replication stress have physiological roles in restarting replication forks or facilitating sister chromatid separation, such transient lesions still pose a threat to genome stability.

1.4.3.4. Fragile sites

Replication stress-induced DNA damage occurs in a non-random fashion across the genome. Regions of the genome prone to breakage upon replication stress are known as fragile sites. Fragile sites can be broadly grouped into two categories based on their replication timings, *early replicating fragile sites* and *common* (late replicating) *fragile sites*. Each category differs in terms of sequence content and genetic context with distinct mechanisms proposed for their respective fragilities.

Common fragile sites (CFS), have been observed for decades as broad genomic regions which manifest as chromosome breaks in metaphase spreads, following exposure of cells to low doses of the DNA polymerase inhibitor, aphidicolin (Glover *et al.*, 1984). Whilst classically defined via Giemsa staining, modern mapping techniques involving FISH or next generation sequencing have resolved many CFS at a higher resolution (Ji et al., 2022). Such mapping reveals that common fragile sites are characterised by late replication timing, low density of replication origins, AT-rich sequences and frequent overlap with long genes (Li & Wu, 2020) . Fragility at common fragile sites may occur via one of two mechanisms, firstly, through collisions between replication and transcription machinery, inevitable in very long genes, as proposed by Helmrich and Ballarino (Helmrich *et al.*, 2011). Some studies however, present evidence against such a model, demonstrating no correlation between transcriptional activity and fragility over CFS, incomplete overlap between genes and CFS as well as comparable fragility in head-on versus head to tail transcription-replication orientations, suggesting collisions are not causative (LeTallec *et al.*, 2013; Brison *et al.*, 2019). The second model for fragility at CFS, which reconciles these observations, is via modulation of replication origin density. Transcription has been shown capable of repositioning (Chen et al., 2019; Gros et al., 2015) or even displacing pre-replication complexes in human cells prior to origin firing (Macheret and Halazonetis, 2018). Thus, over long transcriptional units, origin-poor regions

are generated and, as a consequence, fork stalling upon replication stress is not readily recovered by adjacent forks (Blin *et al.*, 2019; Brison *et al.*, 2019).

Early replicating fragile sites (ERFS) were identified much more recently, as regions accumulating HR components and fork re-modellers following acute treatments with ribonucleotide reductase inhibitor, hydroxyurea, at the onset of S-phase (Barlow *et al.*, 2013). By contrast with CFS, ERFS replicate early and occur in origin-dense, GC-rich regions which overlap with clusters of highly expressed genes (Barlow *et al.*, 2013). As genes found in ERFS are often oriented in a divergent manner, it is proposed that head-on transcription-replication collisions are likely following disruption of the replication timing programme (Barlow *et al.*, 2013).

Both CFS and ERFS map to sites of recurrent genomic rearrangements in cancer suggesting fragility through replication stress is causative of such aberrations (Barlow *et al.*, 2013; Li and Wu, 2020).

1.5. Methods for studying DSB formation.

Techniques for studying genome damage, including DSBs, have evolved over time and vary in their resolution, from assays like Western blotting, which broadly measure prevalence of damage at a population scale, all the way up to next generation sequencing-based techniques capable of identifying individual DSBs at nucleotide resolution. Such techniques, as applied to the study of endogenous DSBs, have enabled contextualisation of DSBs and ultimately lead to identification of novel damage mechanisms as discussed in the following section (Saayman and Esashi, 2022).

1.5.1. Quantitative assays for studying DSB formation

Amongst the earliest techniques for studying DSB formation was the single-cell electrophoresis, or comet assay. The technique lyses cells embedded in agarose, and measures the ability of DNA to migrate away from the nucleus upon electrophoresis, where the presence of DSBs relieves chromatin compaction, allowing DNA migration (Ostling and Johanson, 1984). The technique has the advantage of assaying DSBs in single cells, which gives information as to heterogeneity within the population. Moreover, it measures DSBs directly, rather than via proxy (Ostling and Johanson, 1984). However, owing to the nature of embedding cells in agarose, it is not possible to co-stain cells with antibodies to measure additional parameters such as cell cycle phase, and the assay has been found to detect only near-lethal levels of DSBs, hence lacks sensitivity (Ismail *et al.*, 2007).

Immunological assays using antibodies raised against proteins of the DNA damage response (DDR) are often used as a proxy for DSB presence. Upon DSB formation histone H2AX is phosphorylated on serine 139, yielding γH2AX which acts to recruit DNA repair factors (Rogakou *et al.*, 1998). Antibodies raised against γH2AX have been used in Western blot (Burma *et al.*, 2001), flow cytometry (Katakota et al., 2006) and image cytometry (Paull *et al.*, 2000) as a measure of DSB prevalence in cells. Similarly, upon induction of damage, 53BP1 forms nuclear foci, detectable via immunofluorescence microscopy, marking DSBs scheduled for repair via NHEJ (Schultz *et al.*, 2000; Gupta *et al.*, 2014). Whilst assaying for the presence of DNA damage markers like γH2AX or 53BP1 is typically more sensitive than the neutral comet assay (Ismail *et al.*, 2007), they only mark DSBs which have activated the DDR and therefore may miss transient sub-classes of DSBs or indeed DSBs in cells with compromised DDR. In addition, γH2AX induction has been previously reported to precede DSB formation in cells subjected to replication stress, hence is not specific to DSBs (Petermann *et al.*, 2010).

An interesting modern technique modifies the break labelling *in situ* and sequencing technique (BLISS, discussed in following section), to directly ligate adapter sequences to DNA break ends in fixed cells. DNA is then extracted, followed by digital droplet PCR using primers matching the adapter sequences to accurately count individual break-ends in the starting population of cells (Yao Wang et al., 2022). Whilst more quantitative than counting foci of DDR proteins as a proxy for DSBs, the technique (dc-BLIS) involves bulk extraction of DNA from a population of cells hence does not resolve heterogeneity in DSB prevalence between individual cells (Wang *et al.*, 2022).

1.5.2. Positional assays for DSB formation

Knowledge of the position of a given DSB can provide invaluable information as to its context and therefore possible causes (Saayman and Esashi, 2022). Moreover, identifying sites of specific DSBs, allows site-specific validation of damage, which may enhance sensitivity over global assay such as γH2AX immunofluorescence.

Cytogenetic and molecular biology techniques can detect DSB formation at specific sites. Break-apart FISH is a cytogenetic technique often employed in diagnosis of cancers, which employs fluorescent DNA probes flanking a putative genomic breakpoint. The distance between probes, as determined by fluorescence microscopy, can be used to score whether a given locus is in-tact or broken (Kim et al., 2011). Recently, break-apart FISH has been employed to identify rare (~1%) chromosome breaks in populations of cells via high-content microscopy (Burman *et al.*, 2015; Gothe *et al.*, 2019).

Quantitative PCR can also be used to determine DSB presence at a known genomic locus. If the DSB occurs at a very high frequency, qPCR of gDNA using primers flanking the break site may be sufficient to detect breakage (Hazan *et al.*, 2019). Alternatively chromatin in the vicinity of DSBs may be immunoprecipitated using an antibody raised against a DDR marker e.g. γH2AX, followed by qPCR for a locus of interest (Iacovoni *et al.*, 2010). Such methods can provide quantitative information as to DSB prevalence at specific sites, however, they require advance knowledge as to the genomic position of DSBs. Next generation sequencing techniques, by contrast, have enabled identification of DSB distribution without existing knowledge of break localisation.

DSB-sequencing technology has rapidly evolved in recent years. This has been catalysed, in part, by the explosion in genomic engineering following the advent of CRISPR/CAS9 genome editing technologies, and the need for a means of detecting off-target cut sites (Jinek *et al.*, 2012; Mali *et al.*, 2013). DSB-mapping techniques fall into three categories: i) ChIP of proxy proteins for DSB formation, ii) Bait DSB induction followed by translocation mapping, and iii) *in situ* break labelling.

Early DSB-mapping methodologies relied on immunoprecipitation of proteins as proxies for DSB formation, such as γh2AX, followed by sequencing or array hybridization (Iacovoni *et al.*, 2010). However, γh2AX ChIP is limited in that γh2AX can extend over megabase scales from the site of DSB, and it only captures breaks which have activated the DDR and hence induced γh2AX (Georgoulis *et al.*, 2017).

Higher resolution can be achieved by techniques mapping repair outcomes at an induced 'bait' DSB site, such as translocation capture sequencing (Klein *et al.*, 2011), high-throughput genome translocation sequencing (HTGTS) (Chiarle *et al.*, 2011), and primer-extension-mediated sequencing (Yin *et al.*, 2019). These techniques revolve around "bait DSB" induction, followed by translocation mapping. For example, HTGTS entails introducing a rare restriction site and inducible endonuclease into the genome, to generate a "bait DSB" at a known location in live cells. Translocations then occur stochastically, with regions of endogenous damage ("prey DSBs") via NHEJ and bait:prey junctions can be mapped by sequencing (Chiarle *et al.*, 2011). Such methods improve resolution of DSB mapping compared with γh2AX ChIP, however they are subject to positional bias, whereby translocations are observed more frequently to bait DSBs on the same chromosome (Wei *et al.*, 2018). Moreover, these methods do not map damage *per se,* rather erroneous repair outcomes. Identification of DSB sites is contingent on breaks being repaired by NHEJ and thus break ends subject to end resection, as part of the HR pathway, pass undetected (Bouwman and Crosetto, 2018).

The highest resolution approach entails directly ligating sequencing adapters to broken ends of DNA. The first of such approaches specific to DSBs was termed break labelling enrichment on streptavidin and sequencing (BLESS) and entails extracting genomic DNA, ligating streptavidin-conjugated adapters onto DNA ends, enriching using biotin followed by PCR-based library preparation and sequencing (Crosetto *et al.*, 2013). Since the publication of BLESS, an abundance of techniques based on the direct break labelling approach have been developed including DSB-capture (Lensing *et al.*, 2016), END-seq (Canela *et al.*, 2016), i-BLESS (Biernacka *et al.*, 2018) and Break-seq (Hoffman *et al.*, 2015), which differ subtly in approaches to library preparation, required cell numbers and sensitivity. A new generation of BLESS-like techniques aimed to quantitatively map DSBs across the genome, which has been achieved by use of DNA spike-ins in qDSB-seq (Zhu *et al.*, 2019), inclusion of unique molecular identifier sequences on adapters in BLISS (Yan *et al.*, 2017), and, most recently, via elimination of PCR amplification during library preparation in INDUCE-seq (Dobbs *et al.*, 2022).

DSB-mapping studies have elucidated novel mechanisms of DNA damage (Saayman and Esashi, 2022), and identified pathologically relevant regions of frequent breakage, pertaining to mutations in developmental disorders or indeed structural rearrangements in cancer (Canela *et al.*, 2016; Gothe *et al.*, 2019; Szlachta *et al.*, 2020; Wang *et al.*, 2020).

1.6. Genome damage in hPSC

Genome damage, in particular DSB formation, limits *in vitro* expansion rates of hPSC through induction of apoptosis. More concerningly, erroneous repair of DSBs can yield genetically variant cells, which both limit utility of hPSC as experimental tools, but also constitute a regulatory hurdle, hindering the progress of hPSC-derived cell therapies (Halliwell *et al.*, 2020). Arising during early embryogenesis, such variants may also contribute to the poor survival rate of human embryos (Hassold and Hunt, 2001). Understanding the mechanistic basis of DSB formation in hPSC may inform improved culture practices to limit genome damage and therefore the occurrence of variants. Moreover, the study of genome damage in hPSC may give insights into the origins of mosaic aneuploidy in embryogenesis. We and others have previously noted an increased prevalence of DSBs in hPSC when compared with their differentiated counterparts (Halliwell et al., 2020; Simara et al., 2017; Vallabhaneni et al., 2018). Below, I discuss the unique biological features of hPSC which may predispose them to DSBs, as well as our current knowledge of damage origins in hPSC.

1.6.1. hPSC biology and a predisposition to DSBs

Genome instability and replication stress are hallmarks of cancer (Negrini *et al.*, 2010) and notably, cancer cells share many of the characteristics of hPSC including abnormal cell cycle regulation and rapid proliferation (Kent and Leone, 2019), high levels of transcription (Lin *et al.*, 2012), and often, maintenance of an undifferentiated state (Batlle and Clevers, 2017). It is likely that hPSC share many DNA damage mechanisms with cancer, however, unlike cancer, wild-type hPSC readily commit to apoptosis upon genome damage (Desmarais *et al.*, 2012, 2016), resulting in relatively low mutation rates (Rouhani *et al.*, 2016; Thompson *et al.*, 2020).

Cultured hPSC are biologically similar to the pluripotent cells of the early embryo, which are required to undergo rapid population doublings to generate sufficient cell numbers for development of the entire organism. Accordingly, hPSC have an abbreviated cell cycle, typically taking less than 24 hours per iteration (Barbaric et al., 2014; Becker et al., 2006; Halliwell et al., 2020). This is enabled, in part, by short gap phases, most notably G1 (Becker *et al.*, 2006; J. A. Halliwell *et al.*, 2020). A fast G1-S transition in hPSC is facilitated by high expression of cyclin E, CDK1 and CDK2, yielding highly phosphorylated Retinoblastoma protein (Rb), derepressing the E2F transcription factor which mediates S-phase progression (Liu et al., 2019). In addition to facilitating rapid proliferation, an abbreviated G1 phase is reportedly vital to maintenance of an undifferentiated state, with hPSC only responding to differentiation cues during G1 phase (Mummery *et al.*, 1987; Sela *et al.*, 2012). Rapid progression through G1 however, has consequences for genome stability in the subsequent S-phase. Mouse ESCs have short G1 durations and exhibit markers of replication stress and associated DNA damage during S-phase (Ahuja *et al.*, 2016). Interestingly, G1 extension in mESCs alleviates the replication stress phenotype (Ahuja *et al.*, 2016). The authors propose that replication-associated lesions from the previous S-phase are processed to collapsed replication forks and DSBs, however elongating G1 provides sufficient time for repair of lesions, prior to S-phase re-entry (Ahuja *et al.*, 2016). Rapid proliferation and a short G1 may also deplete nucleotide pools. Bester and colleagues found dysregulation of S-phase entry in fibroblasts, via cyclin E overexpression, was sufficient to deplete dNTPs to an undetectable level, and induce associated replication stress (Bester *et al.*, 2011). Elsewhere, cyclin E overexpression was found to increase damage in a transcription- and replication-dependent manner (Jones *et al.*, 2013). A short G1, mediated by cyclin E overexpression was later shown to induce firing from dormant replication origins *within* gene bodies, normally removed as pre-replication complexes by transcription during G1 (Macheret and Halazonetis, 2018). Firing from these new origins lead to increased transcription-replication collisions accompanied by DSB formation (Macheret and Halazonetis, 2018).

Pluripotent stem cells are also highly transcriptionally active. mESC exhibit open chromatin, permissive of transcription (Gaspar-Maia *et al.*, 2009), moreover radiolabelling of nascent RNA reveals mESC are globally more transcriptionally active than neural progenitor cells (Efroni *et al.*, 2008). High levels of genome-wide transcriptional activity are proposed to poise cells for rapid activation of specific transcriptional programmes in response to differentiation cues (Singh and Dalton, 2009). Whilst no equivalent study has demonstrated globally elevated transcriptional activity in human, hPSC genomes are enriched for euchromatin, as measured by electron microscopy (Golob *et al.*, 2008; Courtot *et al.*, 2014), and open chromatin as determined by ATAC-seq (Meléndez-Ramírez *et al.*, 2021). This chromatin profile is indicative of high transcriptional activity relative to somatic cell types, and thus it stands to reason that hPSC may suffer increased levels of transcription-associated damage, meditated by R-loops and topoisomerase activity. In line with this hypothesis, we recently identified that gene mutation rate correlates with expression level in hPSC (Thompson *et al.*, 2020). Given the concurrent high rates of replication and transcription in hPSC, it is possible that they suffer an increased frequency of transcription-replication collisions over their differentiated counterparts.

Human PSCs predominantly rely on glycolysis to generate cellular energy (Varum *et al.*, 2011), this is accompanied by the generation of high levels of lactate, and a drop in culture medium pH. This medium acidification has been identified as a major cause of genome damage in hPSC culture (Jacobs et al., 2016; Liu et al., 2018), where increased buffering capacity of the culture medium is sufficient to rescue the resulting DNA damage (Liu et al., 2018). Despite relatively low oxidative phosphorylation activity, the mutational signature of hPSCs suggests that they suffer oxidative DNA damage (Thompson et al., 2020). This counterintuitive observation may reflect increased ROS production in hPSC cultured in a low pH environment, as is seen epithelial cells (Zhang *et al.*, 2009; Zhao *et al.*, 2021). Interestingly, both culture medium formulation and external oxygen tension have been shown to alter levels of oxidative phosphorylation in hPSC, thus culture conditions directly affect DNA damage in hPSC (Bangalore *et al.*, 2017; Lees *et al.*, 2019; Xu *et al.*, 2021).

hPSC also have an unusual DDR, particularly in response to replication stress. Early work claimed to find faster repair kinetics for both oxidative lesions and DSBs in hPSCs, relative to differentiated cell types, and noted increased expression of certain genes of DNA repair pathways. Results however were not consistent across differentiated comparisons and lacked controls to monitor levels of apoptosis between time points (Maynard *et al.*, 2008).

Upon sensing stress, proliferative cells activate checkpoints to restrict progression through the cell cycle. Under replication stress, most proliferative cell types form nuclear foci of RPA, due to coating of ssDNA arising at stalled replication forks (Byun *et al.*, 2005). RPA recruits ataxia-

telangiectasia and RAD3-related (ATR), which in turn phosphorylates checkpoint kinase 1 (CHK1), suppressing origin firing and arresting the cell cycle in S-phase to allow time for restart of stalled forks (Cimprich and Cortez, 2008). Interestingly, upon induction of replication stress, hPSC generated less ssDNA than somatic cells, did not to form significant RPA foci and consequently failed to activate the S-phase ATR/CHK1 checkpoint (Desmarais *et al.*, 2012, 2016). The authors propose that this favours DNA damage sensing and P53-mediated apoptosis, reflected by high proportions of apoptotic cells following replication stress (Desmarais *et al.*, 2012, 2016). As cells of the early embryo yield all differentiated cell types, favouring apoptosis over DNA repair may protect the embryo from acquisition of deleterious genetic variants. Consistent with this idea, hPSC reportedly have lower rates of mutation compared to somatic cells (Rouhani *et al.*, 2016; Thompson *et al.*, 2020). Alternatively, lack of S-phase checkpoint activation may serve to facilitate rapid proliferation in hPSC. However, hPSC notably have intact checkpoints at both the G1/S transition (Bárta *et al.*, 2010) and mitotic spindle assembly (Lyu *et al.*, 2022). In either scenario, failure to activate the S-phase checkpoint likely exacerbates genome damage in hPSC, akin to ATR inhibition in cancer (Toledo *et al.*, 2013). Moreover, in instances where apoptosis is not triggered, the resulting increase in DSBs may lead to erroneous repair and genetic change.

1.6.2. Studies on genome damage in pluripotent stem cells

Several studies have noted a quantitative increase in DNA damage burden in pluripotent cells when compared with their differentiated counterparts. The process of reprogramming fibroblasts to iPSC has been associated with genomic instability (Gore *et al.*, 2011; Ji *et al.*, 2012). Shortly after the first derivation of iPSC, it was noted that, during reprogramming, activation of the DDR leads to apoptosis and low reprogramming efficiencies, suggestive of increased DNA damage (Marión *et al.*, 2009). Reprogramming-associated DSBs were found to be partially rescued by antioxidant supplementation in human systems, suggesting oxidative damage is causative (Ji *et al.*, 2014). Ruiz and colleagues were the first to quantitatively demonstrate increased levels of DNA damage and replication stress during reprogramming, where reprogramming of mouse embryonic fibroblasts (MEFS) to iPSC was accompanied by decreased replication fork speed and increased levels of yH2AX, indicative of replication stress-induced DSBs (Ruiz *et al.*, 2015). The authors went on to show that the phenotype could be alleviated with supplementation of nucleosides during reprogramming, or genetically by addition of a further copy of the CHK1 gene, suggesting that, at least during reprogramming, mouse cells are able to activate the S-phase checkpoint to mitigate replication stress (Ruiz *et al.*, 2015).

In steady-state culture of pluripotent stem cells, constitutive replication stress and accompanied DSB formation was first identified in mESC (Ahuja *et al.*, 2016). An increase genome damage, proportional to the increase in replication was subsequently noted in hPSC when compared with differentiated counterparts (Simara et al., 2017; Vallabhaneni et al., 2018). Most recently, we identified that hPSC suffer constitutive replication stress with decreased replication fork speeds, increased stalling and high levels of yH2AX when compared with isogenic differentiated cell types (Halliwell *et al.*, 2020). Notably, this phenotype could be partially rescued by supplementation of culture medium with nucleosides (Halliwell *et al.*, 2020).

Interestingly, whilst culture conditions have been shown to affect replication stress and DNA damage in hPSC, these phenomena are believed to be highly prevalent in the early embryo. A recent investigation into the origins of aneuploidy in embryogenesis found DSBs and markers of replication stress as early as the first S-phase following fertilization in IVF embryos (Palmerola *et al.*, 2022). Fragility occurred in late-replicating regions and was independent of active transcription (Palmerola *et al.*, 2022). Owing to restrictions on embryo culture, it is unknown whether the same is true of hPSC's equivalent in the peri-implantation blastocyst.

1.7. Hypothesis and aims

The phenotype of elevated genome damage in pluripotent cells relative to their differentiated counterparts is well documented (Ahuja *et al.*, 2016; Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018; J. A. Halliwell *et al.*, 2020). Previous work, however, falls short of identifying mechanistic causes of genome damage in hPSC, and a common limitation of these studies is their reliance on purely quantitative assays for genome damage. I reason that genomic mapping of endogenous DSBs in pluripotent cells would enable annotation of genetic and epigenetic features, to identify commonality amongst DSB sites and infer putative causes. My hypothesis is that DSB formation occurs in a non-random fashion and that, owing to their unique biology, hPSC are predisposed to DSBs arising from specific biological processes and exogenous culture conditions. In this thesis, I aim to map the genomic location of DSBs in hPSC and their differentiated counterparts. Using DSB maps I will identify and experimentally validate putative causes of damage and determine whether sites of recurrent DSB formation correspond to recurrent genetic variants observed in hPSC cultures.

2. Materials and Methods

2.1. Cell culture

All cell culture procedures were carried out in a class II laminar flow hood, using sterilised consumables and reagents.

2.1.1. hPSC culture

Experiments were conducted using four different ESC lines: H7 (Thomson *et al.*, 1998), H9 (Thomson *et al.*, 1998), MShef4 (Thompson *et al.*, 2020), MShef11 (Thompson *et al.*, 2020); and one iPSC line: MIFF-1(Desmarais *et al.*, 2016). Human PSC lines were routinely cultured at 37°C under humidified air supplemented with 5% CO_2. For low-oxygen culture, lines were cultured continuously at 5% O_2/5% CO_2, in a hypoxic workstation (Baker Ruskinn), with media and reagents equilibrated for at least 2 hours before use in a high surface area vessel. Cells were cultured in 2D, adhered to tissue culture-treated flasks coated with vitronectin (Gibco, A31804; STEMCELL Technologies, 100-0763), in either S8 medium (prepared in-house, based on Kuo et al. 2020) or mTeSR1 (STEMCELL Technologies, 85850). Media was replenished daily. Cells were passaged every 3-5 days.

2.1.2. Fibroblast culture

The CCD-1112Sk human foreskin fibroblast line (ATCC CRL2429) was cultured in 5% CO_2 ~100% relative humidity incubators at 37°C. Cells were cultured on tissue culture-treated flasks in Iscove's Modified Dulbecco's Medium (Gibco, 12440-046), supplemented with 10% foetal bovine serum (FBS) (HyClone, SV30160.03). Cells were passaged every 3-5 days.

2.1.3. Cell culture vessel preparation

Recombinant human vitronectin, (Gibco, A31804; STEMCELL Technologies, 100-0763) was diluted to 5µg/ml or 7.5µg/ml, respectively, in sterile PBS. This working solution was added to tissue culture treated culture vessels at 0.15ml cm^{-2}. Vessels were shaken briefly to ensure coverage of the entire growth surface, then incubated at room temperature for 1 hour or more.

Coated vessels were stored at room temperature for up to 5 days before use. Vitronectin solution was removed immediately prior to cell seeding.

2.1.4. S8/E6 Culture medium preparation

A 50× supplement for each medium was prepared in house as per Table 2.1. Working solutions of sodium selenite, holo-transferrin, insulin, FGF2-G3 and TGFβ1 were made up in distilled water, fresh on the day of preparation. Insulin working solution was titrated to pH 3 using 100mM HCl. 50× supplements were stored at -20°C in 10ml aliquots.

To prepare complete medium, 50× aliquots were thawed at room temperature then added to 500ml DMEM/F12 (Sigma D6421), mixed by inversion, and filter sterilised using 0.2µm polyethersulfone vacuum filters. Complete medium was stored at 4°C for up to 14 days or -20°C for up to 1 year.

Table 2.1 components and quantities for 1l batch of 50X supplement of S8 or E6 medium

Component	Manufacturer	Catalogue #	Working solution concentration	Quantity: E6	Quantity: S8
DMEM/F12	Sigma	D6434	N/A	346ml	346ml
L-ascorbic acid	Sigma	A8960	N/A	3.2g	3.2g
Sodium selenite	Sigma	S5761	14µg/l	50µl	50µl
Insulin	Sigma	91077C	10mg/ml	97ml	97ml
NaHCO3	Sigma	S5761	N/A	27.15g	69.15g
Holo-transferrin	Sigma	T0665	10mg/ml	53.5ml	53.5ml
Glutamax	Gibco	35050-038	N/A	500ml	500ml
FGF2-G3	QKine	QK053-0500	500µg/ml	N/A	4ml
TGFβ1	Peprotech	100-21C	100µg/ml	N/A	1ml

2.1.5. hPSC passaging

At 60-80% confluence, culture media was removed from cells and replaced with sufficient RelesR (STEMCELL Technologies, 100-0484) to cover the entire growth surface (typically 50µl cm^{-2}). RelesR was removed immediately, and cells incubated at room temperature for 4-7 minutes, until colonies begin to detach from the vessel surface. Fresh medium was then added and rocked over the vessel's growth surface to lift colonies into the medium. Large colonies in suspension were broken apart further by gentle trituration using a 10ml serological pipette. The resulting suspension of colonies was seeded into new vitronectin-coated culture vessels, containing adequate pre-warmed culture medium to achieve a split ratio of 1:3-1:12 depending on the cell line and culture density.

2.1.6. hPSC single cell dissociation and seeding

When cells were 60-80% confluent, culture media was removed from cells, followed by one wash with PBS. 50µl cm^{-2} TryplE (Gibco, 1260401) was then added, rocked to cover the entire growth surface of the vessel, and incubated at 37°C for 3-4 minutes. Vessels were tapped lightly to lift cells, followed by addition and trituration in an additional 9 volumes of DMEM/F12. The resulting cell suspension was transferred to a 15ml centrifuge tube, 10µl was taken for counting with a haemocytometer, and the remaining cell suspension was spun at 200×G for 4 minutes. Supernatant was aspirated and cell pellets resuspended to a volume of 10^6 cells/ml. Cells were then seeded at the specified density in hPSC medium supplemented with 10µM Y-27632 Rho-kinase inhibitor (Generon, HY-10583) for the first 24 hours. At 24 hours, media was replenished without Y-27632.

2.1.7. Cryopreservation of hPSC

At 60-80% confluence, cells were incubated with RelesR (as per 2.1.5). Following RelesR incubation, an appropriate volume of STEMCELL-BANKER (AMSBIO, 11924) freezing medium was added, culture vessels were rocked to lift cells into the freezing medium, followed by trituration with a 10ml serological pipette. The resulting suspension of small colonies was then aliquoted into cryovials at 0.5ml per vial. Vials were stored at -80°C in a CoolCell (Corninig 432002) for 24 hours, before transfer to liquid nitrogen for long term storage.

2.1.8. Thawing of Cryopreserved hPSC

Vials of hPSC were removed from liquid nitrogen and transferred directly to a 37°C water bath for ~2 minutes, with regular shaking, until the contents had completely thawed. The vial exterior was then thoroughly washed in 70% ethanol and ~1ml of room temperature DMEM/F12 was added dropwise to the contents before transferring to a 15ml centrifuge in an additional 3.5ml of DMEM/F12. Cells were spun at 200×g for 1 minute, the supernatant aspirated, then pellets were gently resuspended in ~3ml of medium supplemented with 10μM Y-27632 and transferred to a vitronectin-coated T12.5 culture flask. Media was replenished at 24 hours, without Y-27632.

2.1.9. Human Fibroblast Passaging

At 70-80% confluence, culture media was removed from cells, followed by one wash with PBS. 50μl cm^{-2} TrypLE was then added, rocked to cover the entire growth surface of the vessel, and incubated at 37°C for 8-10 minutes. Culture vessels were tapped gently to detach cells, at which point 9 volumes of fibroblast culture medium was added and triturated at full speed with a 10 ml serological pipette to obtain a single-cell suspension. This suspension was transferred to a 15 ml centrifuge tube and spun at 200×g for 4 minutes. Supernatant was aspirated and pellets were resuspended in 4 ml fibroblast culture medium which was then seeded into vessels containing adequate pre-warmed fibroblast culture medium, to achieve a split ratio of 1:4-1:8.

2.2. Genotyping of hPSC

2.2.1. Karyology

Karyotypes of cultures were determined by analysis of 20 metaphase spreads. All karyotypes were scored by trained clinical geneticists at Sheffield Children's NHS Foundation Trust (Sheffield Diagnostic Genetics Service).

2.2.2. SNP Arrays

DNA was extracted from hPSC pellets in house using the DNeasy blood and tissue kit (Qiagen, 69506). Extracted DNA was then analysed via OGT 60K v3 ISCA Oligo array by trained clinical geneticists at Sheffield Diagnostic Genetics Service.

2.2.3. qPCR CNV assay

Cell cultures were routinely screened for copy number changes of commonly duplicated regions, via qPCR of genomic DNA, as previously described (Baker, Adam J Hirst, *et al.*, 2016; Laing *et al.*, 2019). Briefly, Genomic DNA was extracted from hPSC pellets, using the DNeasy blood and tissue kit as per manufacturer's instructions. DNA was quantified using a nanodrop spectrophotometer and 1µg DNA was digested with EcoRI (Thermo Fisher Scientific, FD0274) for 30 minutes at 37°C in a 50µl reaction. Digested DNA was diluted to 10µg/ml in nuclease free water.

Meanwhile, master-mixes of Taq polymerase (IDT, 1055770), hydrolysis probes and primers (Table 2.2) were prepared for each locus to be tested as well as an internal reference locus. 8µl of master-mix was then aliquoted into wells of a 384 well PCR plate, using an electronic multi-dispense pipette. 2µl of digested DNA solution was then aliquoted into wells accordingly. Plates were spun briefly, sealed with adherent film then loaded into a QuantStudio 12K Flex Real-time PCR machine (Thermo Fisher Scientific) and run using the following profile:

$$\begin{array}{ll} Hold & 50°C\ 2min; \\ Denature & 95°C\ 10min \\ 40 \times Cycles & \begin{cases} 95°C\ 15sec \\ 60°C\ 1min \end{cases} \end{array}$$

Copy number of loci was then determined using the $2^{-\Delta\Delta CT}$ method (Livak and Schmittgen, 2001), calibrated against a known diploid sample, with a rarely duplicated locus (*NWD2*) as internal reference control across samples.

Table 2.2 qPCR primer and probe sequences

5'-3' primer and probe sequences. Nucleotides in square parentheses, prefaced with '+' are locked nucleic acids (LNA). All hydrolysis probes are 5' labelled with FAM1 fluorophore and 3' labelled with BHQ1 quencher. Primers custom ordered from IDT, probes custom ordered from Merck.

Locus	**Primer sequence (F and R)**	**Hydrolysis probe sequence**
NWD2 (4p14)	GCAAGAGCCCAACACCTTTG AGCAGAGGGGAAGGATGGAT	A[+G][+G][+C][+A][+G][+A][+G]
MDM4 (1q32.1)	GCCCCCAGACCTAAATCAAT TCGGTATGACAGCAATGTCTCT	A[+G][+G][+C][+A][+G][+A][+G]

BCL2L1 (20q11.21)	TCTGCAGAAGGCTACCCCTA TGCTGTGTCTAAGACCTCTTTCA	G[+C][+T][+G][+C][+C][+C][+A]

2.2.4. High-throughput DNA extraction and quantification

Confluent hPSC cultures were harvested directly from 96 well plates by addition of 100µl Lysis buffer (10mM Tris, 10mM EDTA, 10mM NaCl, 0.5% SDS, proteinase K 1mg/ml), sealed with an adhesive film and incubated at 37°C overnight. Lysates were then transferred to 96 well PCR plates, and 79µl of 5M NaCl was added. Plates were sealed and shaken vigorously for 20 seconds, then centrifuged at 1000×g for 30 minutes to precipitate protein. 120µl of supernatants were transferred to 96 well v-bottom plates, and 72µl isopropanol added. Plates were then sealed and rocked on an orbital shaker for 30 minutes at room temperature to precipitate DNA. Plates were spun at 1000×g for 5 minutes to pellet DNA. Pellets were then washed twice with 70% ethanol, air dried for one minute and resuspended in 30µl TE buffer (IDT, 11-05-01-09). Following incubation at room temperature overnight, DNA was quantified using a miniaturised Qubit assay. Qubit dsDNA broad range (Thermo Fisher Scientific, Q32850) working solution was prepared as per manufacturer's instructions. The 100 ng/µl dsDNA standard was serially diluted 1:2 four-fold for generation of a standard curve. 1µl of standards or test DNA samples were then pipetted into the bottom of a 384 well black wall plate (Corning, 3573), 9µl dsDNA BR working solution added, plates spun briefly and fluorescence quantified using a plate reader (BMG LABTECH, FLUOstar OPTIMA; Excitation 485, Emmission 520).

2.3. Flow cytometry

2.3.1. Cell surface marker staining: Flow cytometry

Cells were dissociated using TrypLE as per section 2.1.6 and resuspended to 5×10^6 cells/ml in FACS buffer (Table 2.3). Antibodies were aliquoted into 5 ml polypropylene FACS tubes to achieve dilutions listed in Table 2.4-2.5 and 100µl of cell suspension was added. Cells were mixed gently by pipetting, then incubated for 20 minutes at 4°C. Cells were then washed using 3ml of FACS buffer, spun at 200×g for 3 minutes, and supernatant poured off. Pellets were then disrupted by gentle tapping and the wash step was repeated. An appropriate secondary antibody was then added to the sample to achieve dilution listed in Table 2.6 in a total volume of ~200µl. Cells were incubated at 4°C for 20 minutes in the dark followed by one wash, as

previous. Finally, pellets were resuspended in an additional 300µl of FACS buffer before assaying on the BD FACSJazz™ cytometer.

2.3.2. Intracellular marker staining: Flow cytometry

Cells were dissociated using TrypLE as per section 2.1.6 and pellets were resuspended in 1ml 4% paraformaldehyde (PFA) to fix for 10 minutes. 4ml PBS was then added to the cells, followed by centrifugation at 300×g for 4 minutes. Cells were washed once in 1ml PBS, spun at 300×g for 4 minutes, and resuspended in 1ml PBS. At this stage cells were stored at 4°C for up to 2 weeks, or immediately stained for flow cytometry.

Fixed cells were centrifuged at 300×g for 4 minutes to pellet and resuspended in 1ml permeabilization buffer (Table 2.3) followed by incubation at room temperature for 10 minutes. Cells were washed once with 3ml intracellular FACS buffer, pelleted and resuspended to 10^7 cells/ml in intracellular FACS buffer. 50µl of the cell suspension was added to 50µl of appropriate primary antibodies at 2× concentration (Table 2.4) in intracellular FACS buffer. Samples were incubated at 4°C overnight, then washed twice with 3ml intracellular FACS buffer. Samples were then incubated with appropriate secondary antibodies for 1 hour at room temperature in the dark. Cells were washed once with 3ml intracellular FACS buffer and resuspended in an additional 300µl of intracellular FACS buffer before assaying on the BD FACSJazz™ cytometer.

Table 2.3 Buffers for FACS staining. Components for 100 ml working solution.

	Permeabilization buffer	**FACS buffer**	**Intracellular FACS buffer**
PBS	89ml	90ml	90ml
Bovine Serum Albumin	1g	-	1g
FBS	10ml	10ml	10ml
Triton-X-100	0.5ml	-	-

2.3.3. Flow cytometric analysis

A BD FACSJazz™ cytometer was used for all flow cytometric analyses. Prior to running samples, 8 peak calibration beads (Biolegend, 422903) were run through the cytometer and laser positions were adjusted manually to obtain greatest resolution between fluorescence peaks on each photomultiplier tube (PMT). Negative control samples were next run through the machine (isotype control for cell surface staining or secondary only control for intracellular FACS). Forward scatter and side scatter voltages were adjusted to detect cells in the centre of scatter plot, and the bulk cell population (P1) was taken forward to doublet discrimination (Figure 2.1 A). Events from the P1 gate were then gated based on trigger pulse width, to eliminate cell doublets from analysis (P2) (Figure 2.1 B). Only events passing both P1 and P2 gates were assayed for fluorescence. PMT voltages were adjusted, so that negative fluorescence distributions were contained in the first decade (Figure 2.1 C). Gates for positive cells were then set using negative controls, gates were positioned to exclude >99% of the negative control sample (Figure 2.1 C, D). Fully stained samples were then assayed, capturing 10,000 single cells for each sample. Plots and statistics were then generated using FlowJo™ software.

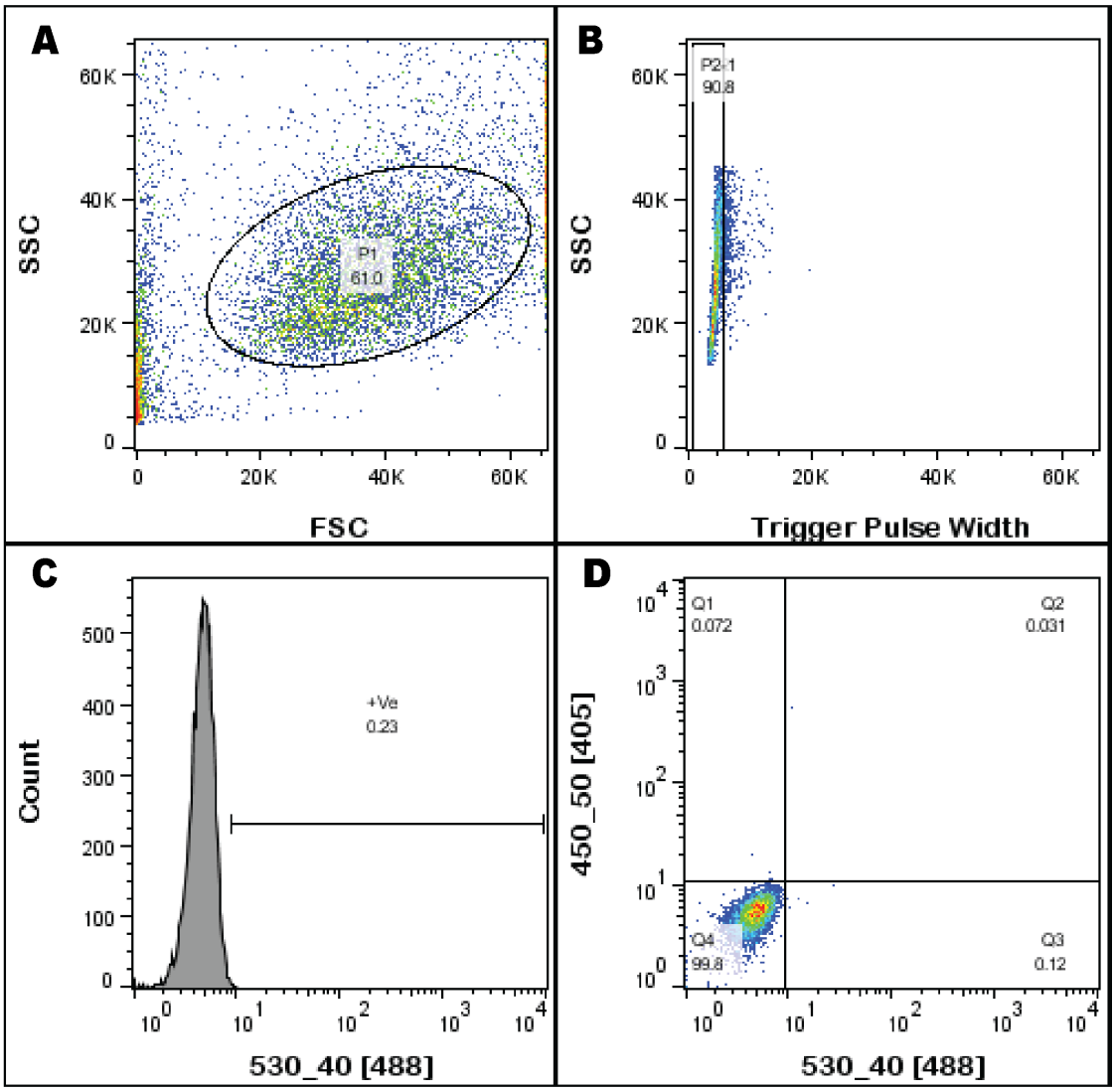


Figure 2.1 FACS gating example

A) Gate P1 is set to encompass only events with FSC/SSC values typical of cells. B) Events in P1 are further filtered based on trigger pulse width, to eliminate doublet cells. Events passing gates P1 and P2 are then analysed based on fluorescence at specific wavelengths. Gates for positive populations are set against a negative control sample stained with secondary antibody only. C) Example single channel fluorescence gating (-ve control in grey). D) Example dual fluorescence gating (-ve control pseudo colour dot plot).

2.3.4. FACS single cell cloning

All single-cell sorting was carried out using a FACSJazz™ cytometer. Laser alignment was carried out as per section 2.3.3, following which drop delay for sorting was optimised using Accudrop fluorescent beads (BD, 345249). Next, rainbow calibration beads were individually sorted into wells of an empty 96 well plate. The plate was screened using a fluorescent microscope to verify that only one bead per well was present and that the bead was deposited

in the centre of the well. Cells were then dissociated as per section 2.1.6, strained through a 70µm cell strainer to remove clumps (Miltenyi, 130-095-823) and resuspended in mTeSR Plus medium (STEMCELL Technologies, 00-0276). Events passing gate P2 (section 2.3.3) were sorted into wells of a 96 well plate, pre-coated with Geltrex (Thermo Fisher Scientific, A1413201), containing 60µl mTeSR Plus supplemented with CloneR2 (STEMCELL Technologies, 100-0691) and Normicin (InvivoGen ant-nr-2). Post-sorting, plates were centrifuged briefly, to aid cell adherence to matrix. Plates were then incubated at 37°C, with media half replenished every two-three days.

2.4. Immunocytochemistry

2.4.1. Intracellular staining

Adherent cells in culture were washed once with adequate PBS to cover the growth surface. Cells were then fixed for 15 minutes at room temperature by addition of adequate 4% PFA to cover the growth surface. Cells were washed 3 times with excess of PBS and stored in excess PBS at 4°C for up to 6 months prior to staining and imaging.

Fixed cells were next permeabilised and blocked for 1 hour by addition of permeabilization buffer. Permeabilization buffer was then removed and replaced with appropriate primary antibodies diluted in intracellular FACS buffer. Samples were incubated for 1 hour at room temperature, or 24 hours at 4°C. Primary antibodies were removed, followed by three 5-minute washes with intracellular FACS buffer. Next, appropriate secondary antibodies were added, diluted, as per Table 2.6, in intracellular FACS buffer with DAPI (Thermo Fisher Scientific, 62248) at 1µg/ml to counterstain nuclei. Secondary antibodies were left to incubate for 1 hour at room temperature in the dark. Samples were then washed 3 times for 5 minutes with intracellular FACS buffer. Finally, wells were filled with PBS, prior to immediate imaging or short-term storage at 4°C (up to 1 week).

Table 2.4 Commercial Primary Antibodies and dilutions

Antibody	Species	Manufacturer	Catalogue #	Application	Dilution
OCT4	Mouse	AbCam	ab184665	FACS	1:1600
OCT4	Rabbit	Cell signalling	2890	ICF	1:1600
TBXT	Rabbit	AbCam	ab209665	FACS	1:400

Nanog	Rabbit	Cell Signaling	4903	ICF	1:400
γH2AX	Mouse	AbCam	ab26350	ICF	1:1000
γH2AX	Rabbit	Cell Signalling	9718	FACS	1:400
Cleaved Caspase-3	Rabbit	Cell Signalling	9661	ICF/FACS	1:400

Table 2.5 In-house primary antibodies and dilutions

Antibody	**Species**	**Manufacturer**	**Catalogue #**	**Application**	**Dilution**
P3X	Mouse	In-house	NA	FACS	1:10
SSEA-3	Rat	In-house	NA	FACS	1:10
SSEA-4	Mouse	In-house	NA	FACS	1:50
TRA-181	Mouse	In-house	NA	FACS	1:50
TRA-160	Mouse	In-house	NA	FACS	1:10
TRA-185	Mouse	In-house	NA	FACS	1:50

Table 2.6 Commercial secondary antibodies and dilutions

Antibody	**Species**	**Manufacturer**	**Catalogue #**	**Application**	**Dilution**
Anti mouse-AF647	Goat	Jackson	115-605-044	ICF/FACS	1:200
Anti mouse-AF488	Goat	Jackson	115-545-044	ICF/FACS	1:200
Anti mouse-AF594	Goat	Jackson	115-585-044	ICF	1:200
Anti mouse-FITC	Goat	Jackson	115-095-044	ICF	1:200
Anti-Rabit-AF594	Goat	Jackson	111-585-045	ICF	1:200
Anti-Rabit-AF647	Goat	Jackson	111-605-045	ICF/FACS	1:200

Anti-Rabbit-BV421	Goat	Becton-Dickinson	565014	FACS	1:500

2.5. Image analysis

Fluorescent images were acquired using a high content fluorescent microscope (GE, InCell 2200). Per cell or per condition parameters were measured using CellProfiler (Stirling *et al.*, 2021), as detailed below.

2.5.1. Cell count analysis

Plates of PFA-fixed cells, counterstained for DNA with either DAPI or Hoechst 33342 (Thermo Fisher Scientific, H3570) were imaged on the DAPI channel using a 10X objective lens, with sufficient exposure to obtain high signal:noise ratio with minimum saturation. Images were saved in Tiff format. For images with high background fluorescence, pre-processing was carried out in FIJI (Schindelin *et al.*, 2012), using the *subtract background* function, with a rolling window diameter of 50 pixels. Pre-processed images were imported into *CellProfiler4* (Stirling *et al.*, 2021) and nuclei were segmented based on intensity and shape. Per-field nuclear counts were written to a csv file.

2.5.2. γH2AX foci quantification

Plates of PFA-fixed cells stained for γH2AX and counterstained with DAPI were imaged using a 40X objective lens, on appropriate channels. Images were then pre-processed in FIJI (Schindelin *et al.*, 2012) using the *subtract background* function, with a rolling window diameter of 100 pixels.

Pre-processed images were imported into *CellProfiler4* and γH2AX foci were enumerated using a custom pipeline. Briefly: non-overlapping nuclei were segmented based on fluorescence intensity and used to generate a mask. γH2AX foci were next enhanced using the *Enhance Features* module to enhance speckles of diameter 8 pixels. Enhanced foci were then masked using the nuclear image and individual foci were thresholded based on intensity. Per nucleus metrics, including γH2AX foci count and DAPI integrated intensity were exported as a csv file.

Downstream, in instances where cells with differing cell cycle phase occupancies were being compared e.g., pluripotent vs differentiated, S/G2 phase cells were subset based on DAPI integrated intensity.

2.6. INDUCE-seq

2.6.1. INDUCE-seq plate preparation

96 well tissue culture-treated plates were coated using poly-D-lysine (Thermo Fisher Scientific, A3890401) diluted to 50µg/ml in PBS. Plates were then washed twice with excess of PBS to remove surplus poly-D-lysine.

2.6.2. INDUCE-seq plate seeding

Following 5 days growth under respective conditions, cells were dissociated using TrypLE, counted and pelleted. Pellets were resuspended to a concentration of 10^6 cells/ml in DMEM/F12 medium. 100µl of the resulting suspension was then added to a well of a 96 well plate, prepared as per 2.6.1. and left to adhere for 20 minutes at room temperature. 100µl methanol-free formaldehyde (Thermo Fisher Scientific, 28908), diluted to 8% w/v in PBS, was then added to the existing media to fix cells (10 minutes at room temperature). Cells were then washed twice with excess PBS and stored in 200µl PBS at 4°C, sealed with adhesive plate seals.

2.6.3. INDUCE-seq library preparation and sequencing

All library preparation and sequencing was performed by Broken String Biosciences, as described previously (Dobbs *et al.*, 2022). Briefly, fixed cells were permeabilised, and double-stranded ends of DNA blunted using T4 polymerase (NEB, E1201S). A full length P5 sequencing adapter was then ligated *in-situ* to blunted ends (NEB, M0202M). DNA was then extracted and fragmented to ~500bp. A half functional p7 sequencing adapter was next ligated to fragments, and the resulting library purified and hybridised directly to Illumina flow cells without PCR amplification. Sequencing was performed using the Illumina NextSeq 500 platform.

2.7. Bioinformatic analysis of sequencing data

Key analysis software and packages used in the following section are listed in Table 2.7.

Table 2.7 Key bioinformatics software

Software	Version	Reference
R	4.2.0	https://cran.r-project.org/
Bedtools	2.30.0	(Quinlan and Hall, 2010)
Deeptools	3.5.1	(Ramírez *et al.*, 2016)
Spectacle	1.4	(Song and Chen, 2015)
Samtools	1.7	(Danecek *et al.*, 2021)
Python	3.9.13	https://www.python.org/
RLpipes	0.9.3	https://anaconda.org/bioconda/rlpipes
metagene2	1.16.0	https://github.com/ArnaudDroitLab/metagene2
Conda	22.9.0	https://github.com/conda
CiiiDER	0.9	(Gearing *et al.*, 2019)
BiomaRt	2.54.1	(Durinck *et al.*, 2009)
bowtie2	2.5.1	(Langmead and Salzberg, 2012)
IGV	2.8.0	(Robinson *et al.*, 2011)
clusterProfiler	4.6.0	(T. Wu *et al.*, 2021)

2.7.1. Read alignment and mapping of INDUCE-seq break end coordinates

De-multiplexing, read alignment and identification of break-end coordinates was carried out by Broken String Biosciences, using a custom, published pipeline (Dobbs *et al.*, 2022). Briefly, de-multiplexed FASTQ files were trimmed of adapter sequences using *Trim Galore!* Reads were aligned using the *Burrows-Wheeler Aligner* (BWA) (Li and Durbin, 2009), low-quality alignment score reads were removed using *Samtools* (Danecek *et al.*, 2021). The resulting BAM files were trimmed to the 5' nucleotide and converted to bed format using *bedtools bamtobed* (Quinlan and Hall, 2010).

2.7.2. Read QC: replicate correlation.

The hg19 reference genome was divided into sequential 100kb bins using *bedtools makewindows* (Quinlan and Hall, 2010). Break ends for each sample were counted in bins, using *bedtools intersect* (Quinlan and Hall, 2010). The resulting files were imported into *R* (https://cran.r-project.org/) and correlation matrices were constructed per line and plotted using *pheatmap* (https://cran.r-project.org/web/packages/pheatmap/).

2.7.3. Determining Break enrichment over genomic features

Refseq genomic hg19 annotations were downloaded as bed files from UCSC. For a given sample, 10^5 break ends were sampled at random using GNU *sort -R* (https://www.gnu.org/software/coreutils/manual/html_node/sort-invocation.html). Sampled break ends were shuffled within chromosomes at random (excluding a mask region of unmapped regions) using *bedtools shuffle* (Quinlan and Hall, 2010). The shuffling process was repeated 10^4 times, to obtain a random distribution of break ends. Each real and shuffled dataset was intersected with the genomic annotation bed file to determine the frequency of overlap, using *bedtools intersect* (Quinlan and Hall, 2010).

Z-scores were calculated as:

$$zscore = observed\ events - \frac{mean\ permuted\ events}{SD\ permuted\ events}$$

P-values were calculated as:

$$p = \frac{1 + N\ Permuted\ frequencies\ more\ extreme\ than\ observed\ frequencies}{N\ Permutaitons}$$

2.7.4. Chromatin state annotation

ChIP-seq, DRIP-seq and ATAC-seq datasets for the H9 hESC line (Yan *et al.*, 2020) were downloaded from NCBI sequence read archive (SRA) and aligned to hg19 by the University of Sheffield Bioinformatics Core. BAM files were input into Spectacle software, using the default parameters. Variable numbers of chromatin states were identified to empirically determine the most suitable output. Ten was the highest number of chromatin states that were readily distinguishable from one another and thus used for annotation. A bed file of chromatin

states was intersected with H9 pluripotent and randomised break end bed files, and enrichment was determined as in 2.6.2.

2.7.5. R-loop DSB enrichment analysis

DRIP-seq data from Yan *et al.,* (2020) was processed by the University of Sheffield Bioinformatics core using the RLpipes pipeline (https://anaconda.org/bioconda/rlpipes), to call peaks for each H9 pluripotent replicate independently (n=4). Subsequently, *bedtools intersect* (Quinlan and Hall, 2010) was used to identify a consensus peak set common to all 4 replicates. Flanking regions were generated using *bedtools flank* (Quinlan and Hall, 2010), and the above bed files were intersected with H9 pluripotent and shuffled break end bed files and enrichment determined as previously.

2.7.6. Gene expression analysis

H9 RNA-seq data (n=2 replicates) from Yan *et al.,* (2020) was aligned to hg19 and transcripts per million (TPM) for Ensembl genes were calculated by the University of Sheffield Bioinformatics core. Expressed genes were subsequently defined as any gene with a mean TPM value >1. Expressed genes were next divided into quartiles based on expression level (TPM). Transcription start sites (TSS) were defined as the start coordinate of the first Ensembl transcript, transcription termination sites (TTS) were defined as the end coordinate. Breaks per gene (TSS-TTS) and breaks per promoter (TSS+1kb/-2kb) in H9 pluripotent samples were determined using *bedtools intersect* (Quinlan and Hall, 2010)

2.7.7. Metagene plots

Metagene profiles were plotted using the metagene2 package (https://github.com/ArnaudDroitLab/metagene2), using mean coverage of H9 pluripotent INDUCE-seq datasets in the region +/-5kb of Ensembl genes' TSS. Categorisation of genes based on the presence/absence of a CpG island in the promoter was carried out as described previously (Yang *et al.*, 2015). Briefly, CpG island bed files for the hg19 reference genome were downloaded from UCSC. Using *bedtools intersect* (Quinlan and Hall, 2010), genes with at least one CpG island in the region +/-500bp of the TSS were designated as CpG+ve.

2.7.8. Promoter heatmaps and k-means clustering

Promoter heatmaps were generated over the region +/-5kb of the TSS of expressed Ensembl genes using *deeptools plotHeatmap* function (Ramírez *et al.*, 2016) on pooled BAM files of each pluripotent sample. Promoters were clustered via *deeptools* using the kmeans algorithm (Ramírez *et al.*, 2016) for the H9 pluripotent sample only. Clustering was optimised using variable cluster counts (k) to determine the highest cluster number with discernible DSB distributions across the promoter region. The clusters identified in H9 were used for plotting heatmaps in the remaining 4 pluripotent samples to check reproducibility.

2.7.9. Gene ontology enrichment analysis

Genes of each cluster were analysed for enrichment of biological process gene ontologies using the *clusterProfiler* package in *R (*https://cran.r-project.org/).

2.7.10. Transcription factor binding site enrichment analysis.

For Gene clusters 1 and 2, HGNC Symbols were retrieved from Ensembl IDs using BiomaRt (Durinck *et al.*, 2009). The resulting list of genes were assayed for differential enrichment of transcription factor binding sites in TSS +/-500bp from the JASPAR 2020 core mammalian database (Fornes *et al.*, 2020). CiiiDER software (Gearing *et al.*, 2019) was run using the default parameters with cluster 1 genes as the test set and cluster 2 as the background.

2.7.11. RNAPII pause index calculation

RNAPII ChIP-seq FASTQ files (Lyu *et al.*, 2018) were downloaded from the SRA. FASTQ files were already trimmed of adapter sequences and low quality sequencing reads and were therefore directly aligned to hg19 using bowtie2 (Langmead and Salzberg, 2012). BAM files were generated from the resulting SAM files using Samtools (Danecek *et al.*, 2021). Pause indices for genes of each cluster were calculated as previously (Ray *et al.*, 2022). Briefly, pause sites and elongation sites were designated as TSS+300bp/-100bp and TSS+500/+2000bp respectively. RNAPII read coverage was determined for each site using *bedtools multicov* function (Quinlan and Hall, 2010) and pause indices for each gene were calculated as RNAPII coverage in pause site/ RNAPII coverage in elongation site.

2.7.12. DSB Hotspot calling method comparison.

For the purpose of comparison, hotspots were calculated on a 1kb scale using three methods with a sample dataset of the H7 cell line. In the sequential binned approach, the hg19 reference genome was divided into sequential 1kb bins using the *bedtools makewindows* function (Quinlan and Hall, 2010). Individual breaks were counted in each interval using *bedtools intersect -c* (Quinlan and Hall, 2010).

In the merge approach, Individual breaks were merged into broader intervals within 500bp upstream/downstream of one another using the *bedtools merge -n* (Quinlan and Hall, 2010) to simultaneously count the number of breaks merged in each interval.

For the sliding windows approach, the hg19 reference genome was divided into overlapping bins of 1kb, offset by 25bp, using the *bedtools makewindows -s* function (Quinlan and Hall, 2010). Sliding windows were then intersected with the break end bed files.

Of the resulting intervals from each of the three methods, the break density value of the top fifth percentile was calculated for each method in *R* (https://cran.r-project.org/), and all intervals with a density greater than the threshold were taken as “hotspots”. Overlap between identified hotspots was calculated in *R*, using the *GenomicRanges* package (Lawrence *et al.*, 2013).

2.7.13. Sliding window optimization

To determine optimal window size, the hg19 reference genome was divided into variable width windows with a 10% slide distance ranging from 50bp (5bp slide) to 10Mb (1Mb slide). A real and randomly shuffled H7 pluripotent dataset were intersected with the resulting windows. Top percentile, fifth percentile and quartile break density values were called for each dataset, and the fold increase in real/random threshold values were plotted as a function of window distance, as were threshold pass rates. For each window size, the break density threshold calculated on the real dataset was applied to both real and randomized datasets and the fold change in number of hotspots called in real/random datasets plotted as a function of window size to determine specificity.

2.7.14. Sliding window hotspot calling

The hg19 reference genome was divided into sliding windows of length 100kb, with a slide distance of 10kb, break end bed files for each replicate of each sample were intersected

independently of one another. For each sample, the threshold break value for the top percentile of intervals was calculated in *R*, then applied using a GNU *awk* script (https://www.gnu.org/software/gawk/manual/gawk.html).

2.7.15. Differential enrichment of DSBs analysis

Analysis of differential enrichment of DSBs in hotspot regions was carried out using the DiffBind package (Ross-Innes *et al.*, 2012) in R (https://cran.r-project.org/). First, a consensus peak set, i.e. hotspots overlapping in two or more samples was identified using *dba.peakset()*, specifying the minimum overlap between hotspots as 1bp. DSBs were then counted for each sample across the consensus peak set, followed by trimming of peaks to the 100kb regions centered on the maxima across all samples using *dba.count().* Break counts over hotspots were normalized as reads per kb per million (RPKM) using *dba.normalize().* The contrast group was set as "condition" i.e. pluripotent vs differentiated, with cell line specified as a confounding factor for more sensitive results, using *dba.contrast().* Differential enrichment analysis was then run using the DEseq2 parameters with *dba.analyze(),* hotspots with a false discovery rate (FDR) value <0.05 and a log2FC >0 were designated pluripotent-specific hotspots.

2.7.16. Hotspot long highly expressed gene overlap

Quartile values for gene length, and expression level (TPM values from Yan et al., 2020 H9 RNA-seq dataset) were calculated in *R* (https://cran.r-project.org/). Long, highly transcribed genes were defined as genes in the top quartile for length and the top quartile for expression. Fisher's exact test was carried out to determine the significance of overlap with pluripotent-specific hotspots, using the *bedtools Fisher* function (Quinlan and Hall, 2010).

2.7.17. Hotspot translocation frequency

A database of 4000 abnormal hPSC karyotypes (Stavish *et al., manuscript in preparation)* was used to obtain hg19 genomic intervals of cytogenetic bands implicated in translocation breakpoints. Karyotypes varied in resolution, (e.g. where one karyotype may report a translocation mapping to 1q21, others may report translocations mapping to the subsidiary band 1q21.1). Overlapping cytogenetic bands were therefore merged to the size of the largest reported region using *bedtools merge* (Quinlan and Hall, 2010). Cytogenetic bands comprising >25% unmapped genome were excluded from analysis. Normalized translocation frequency

was calculated as the number of times a cytogenetic band was implicated in a chromosomal translocation in hPSC, divided by its length. The resulting bed file was intersected with DSB hotspot bed files using *bedtools intersect* (Quinlan and Hall, 2010) and hotspot break density plotted against normalized translocation frequencies of the cytogenetic band.

2.8. Statistical analysis

All statistical analysis was performed in either *Graphpad Prism* software (www.graphpad.com) or *R* (https://cran.r-project.org/). Distributions of data were inspected to inform choice of parametric versus non-parametric test. In instances of multiple testing, P-values were adjusted accordingly. Details of specific tests, correction methods for multiple comparisons, and n numbers are detailed in figure legends.

3. Cell line characterisation and process validation for sequencing

3.1. Introduction

HPSC acquire recurrent genetic changes over prolonged *in vitro* culture (Halliwell et al., 2020). Many of these changes are preceded by DNA damage in the form of DSBs. Using quantitative techniques, previous studies have identified higher basal levels of DNA damage in pluripotent cells than their differentiated counterparts (Ahuja et al. 2016; Halliwell et al. 2020; Ruiz et al. 2015; Vallabhaneni et al. 2018a). However, the molecular basis of this damage remains poorly characterized. Culture conditions are known to affect levels of genome damage in hPSC, for instance low culture medium pH (~pH7.0-pH6.5) increases DNA damage (Jacobs *et al.*, 2016; Liu *et al.*, 2018). By contrast, low-oxygen culture has been shown to reduce the incidence of chromosomal breaks in hPSC and decrease mutation frequency (Forsyth *et al.*, 2006; Thompson *et al.*, 2020), raising the intriguing possibility that modulating culture conditions, specifically oxygen tension, may alter the position and density of DSBs across the pluripotent genome.

To date, studies investigating genome damage in hPSC lack information on the genomic location of DSBs. I reasoned that by mapping DSBs and annotating genomic, transcriptomic and epigenomic features, I would be able to infer putative causes of genome damage in hPSC and determine whether recurrent genetic changes in hPSC correspond to hotspots of genome damage. Moreover, I postulated that mapping DSBs under low and high oxygen tension would also allow me to determine how modifying culture conditions impacts the distribution and abundance of DSBs. Finally, I posited that mapping DSBs in undifferentiated hPSC and their differentiated derivatives would enable me to identify DSBs that are specific to the pluripotent state. I therefore aimed to use INDUCE-seq (Dobbs *et al.*, 2022) to map DSBs in hPSC lines cultured under the following conditions:

1) Undifferentiated hPSC grown under atmospheric oxygen (“Pluripotent 20% O_2”),

2) Differentiated cells grown under atmospheric oxygen (“Differentiated 20% O_2”),

3) Undifferentiated hPSC grown under low oxygen (“Pluripotent 5% O_2”) (Figure 3.1). Together, these conditions will facilitate the comparison of breakage between pluripotent and differentiated states, determining sites of recurrent damage unique to pluripotent cells, and determining the effect of oxygen tension on DNA damage in hPSC.

Prior to embarking on such a large-scale sequencing experiment, I first set out to characterise the hPSC lines I intended to use and optimise experimental conditions to generate samples for subsequent sequencing.

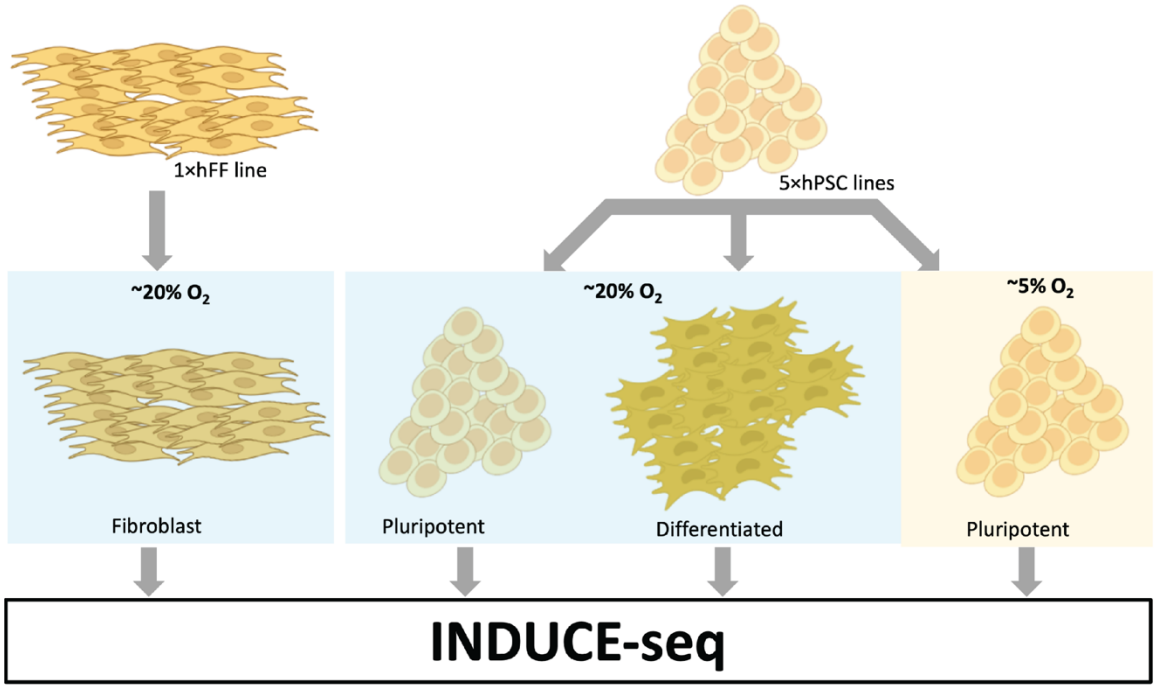


Figure 3.1 Schematic of proposed INDUCE-seq experiment

3.1.1. DNA double-strand breaks in hPSC

Several studies have noted increased levels of DNA damage in pluripotent cells, when compared with isogenic differentiated counterparts. Ruiz and colleagues first demonstrated that reprogramming of mouse embryonic fibroblasts (MEFs) to iPSC induces replication-associated DSBs (Ruiz *et al.*, 2015). Mouse ESCs were later shown to constitutively express markers of replication stress and associated DNA-damage, namely γH2AX and 53BP1 (Ahuja *et al.*, 2016).

The first of such studies in human cells compared γH2AX and 53BP1 foci accumulation in iPSCs and their parental fibroblasts and found that, whilst iPSC universally harboured higher numbers of foci per cell than fibroblasts at a population level, when subsetting only replicative cells via EdU incorporation, fibroblasts harboured comparably high numbers of foci (Simara *et al.*, 2017). Similarly, Vallabhaneni and colleagues identified increased levels of γH2AX, and

replication stress markers in pluripotent cells over differentiated counterparts, but found this to be proportional to the fraction of cells in S phase of the cell cycle (Vallabhaneni et al. 2018).

Previous work from our lab corroborated higher levels of DNA damage in pluripotent than differentiated cells, however, unlike the two aforementioned studies, γH2AX was reportedly higher in pluripotent than differentiated states, when comparing only S/G2 fractions of the cell cycle (Halliwell et al., 2020). All of the above studies point to prevalent replication-associated damage in pluripotent cells but fall short of elucidating the mechanistic basis of this damage.

3.1.2. Low oxygen in hPSC culture

HPSC are commonly cultured under atmospheric oxygen (~20%), in humidified 5% CO_2 incubators. This constitutes a super-physiological oxygen tension. Intrauterine oxygen measurements suggest cells of the epiblast, hPSC's closest developmental equivalent (Nichols and Smith, 2009), are exposed to ~2% oxygen (Jauniaux *et al.*, 1999). Several studies have investigated the effect of lower, more physiological oxygen tensions on hPSC cultures and revealed that low oxygen conditions enhance growth rate (Forristal *et al.*, 2010), increase expression of pluripotency-associated markers (Närvä *et al.*, 2013; Mathieu *et al.*, 2014) and reduce levels of apoptosis (Mathieu *et al.*, 2014). Most interestingly, low oxygen culture has also been shown to reduce 53BP1 foci, a marker of DSBs (Guo *et al.*, 2013), yield fewer chromosomal breaks (Forsyth *et al.*, 2006), and reduce mutation rate (Thompson *et al.*, 2020), relative to cells cultured under atmospheric oxygen. By contrast, culture of hPSC under atmospheric oxygen can lead to increased generation of ROS, and ultimately oxidative DNA damage (Jagannathan *et al.*, 2016).

3.1.3. Characterisation of hPSCs

Since the derivation of the first hESC lines in 1998 (Thomson *et al.*, 1998), hundreds of additional hPSC lines have been derived; these numbers exploded following the advent of hiPSC, with 895 hESC and 3950 hiPSC lines currently registered on the hPSCreg database (Seltmann *et al.*, 2016). Pluripotency is defined functionally, by the ability of cells to give rise to all 3 germ layers, which can be demonstrated via trilineage differentiation either *in vivo* or *in vitro* (Allison *et al.*, 2018). For cell lines where pluripotency has previously been functionally demonstrated (as used in this study), it is routine to use expression of markers of the undifferentiated state, including pluripotency-governing transcription factors such as POU5F1 (also known as OCT4) (Nichols *et al.*, 1998), to demonstrate the undifferentiated status of cells (ISSCR, 2023). As cells can undergo spontaneous differentiation even under conditions

supportive of pluripotency (Kurek *et al.*, 2015), it is critical to regularly validate the undifferentiated status of hPSC lines.

Individual hPSC lines exhibit widely variable characteristics such as growth kinetics and differentiation capacity. This variation can stem from natural differences between lines, or alternatively derive from the presence of genetically variant populations of cells (Kilpinen *et al.*, 2017; Strano *et al.*, 2020; Vitillo *et al.*, 2023). The presence of genetic variants in hPSC cultures could have consequences for the outcome of the planned sequencing experiment. For instance, the common duplication of *BCL2L1* or deletion of *P53,* endows cells with a higher apoptotic threshold which may, in turn, affect levels of DNA damage endured by a cell, prior to committing to apoptosis (Avery *et al.*, 2013; Amir *et al.*, 2017). Alternatively, a gross karyotypic abnormality involving gain or loss of genetic material, could give rise to artefactual over or under-representation of the implicated genomic region in sequencing read depth.

Historically, karyotyping has been widely used for the detection of genetic variants in hPSC cultures (Draper *et al.*, 2004) and remains an important tool, providing genome wide information on the structure and copy number of genomic regions. However, the resolution of karyotyping is limited to ~5-10Mb and hence smaller aberrations can pass undetected (Baker et al., 2016). In instances of known, highly recurrent CNVs, such as the commonly gained amplicon on 20q11.21 (Amps et al. 2011), a targeted qPCR-based method can be used to calculate copy number of a gene within a sample, relative to a known diploid control (Baker et al., 2016; Laing et al., 2019). SNP arrays, unlike karyology, lack structural information but do provide a relatively high-resolution (~1kb), genome-wide map of genetic variation, useful in detecting genetic duplications, or deletions causing loss of heterozygosity in tumour suppressor genes such as *TP53* (Merkle *et al.*, 2017).

The presence of genetic variants can confound experimental results and it is therefore important to interrogate the genetic status of cells prior to undertaking experiments. As no *one* method can detect all types of genetic changes observed in hPSC, it is necessary to use a range of complementary techniques to effectively characterise the genetic status of cells.

3.1.4. Aims

In this chapter I set out to validate hPSC lines and optimise experimental conditions to be used in my subsequent INDUCE-seq experiment. Specifically, I aim to:

1) Characterise the genetic status of all cell lines for use in the study

2) Optimise a protocol for the differentiation of all hPSC lines and quantitatively compare DSBs between states.

3) Optimise low-oxygen culture conditions and validate a low-oxygen phenotype

4) Generate characterised cellular material for INDUCE-seq

3.2. Results

3.2.1. Genetic Characterisation of a diverse panel of hPSC lines.

To ultimately define a pluripotent damage phenotype, I sought to characterize banks of six different cell lines: four hESC lines (H7, H9, MShef4, MShef11), one iPSC line (MIFF-1) and the parental human foreskin fibroblast (hFF) line (CRL2429) from which MIFF-1 was reprogrammed. The hPSC lines used in this study have all been published previously, and their pluripotency demonstrated functionally, either by teratoma assays or *in vitro* differentiation. For simplicity, cells determined to be undifferentiated by antigen expression are herein referred to as *pluripotent*, however it should be noted that no functional tests of pluripotency have been carried out in parallel with the experiments described in this thesis.

As a first-pass phenotypic confirmation that cell lines were genetically normal, I sought to compare cloning efficiencies of all five hPSC lines, alongside two known genetically variant subclones of the H7 line, harbouring duplications of chromosomes 17 and 1 respectively as a positive control (Figure 3.2 A, B). Wild-type hPSC lines exhibit very low cloning efficiencies, typically between 1-20%, depending on the cell line (Chen *et al.*, 2010, 2021; Barbaric *et al.*, 2014). I therefore posited that a significant deviation from this range may be due to the presence of genetic variants in culture. Lines were seeded as single cells, at a density of 500 cells cm $^{-2}$. Following five days of growth, cells were fixed, stained with Hoechst33342 and colony numbers enumerated via high-content microscopy. Whilst all lines exhibited cloning efficiencies comparable to genetically normal lines in the literature, MShef11 appeared to deviate from the other four lines used in this study, with a mean cloning efficiency of ~10%.

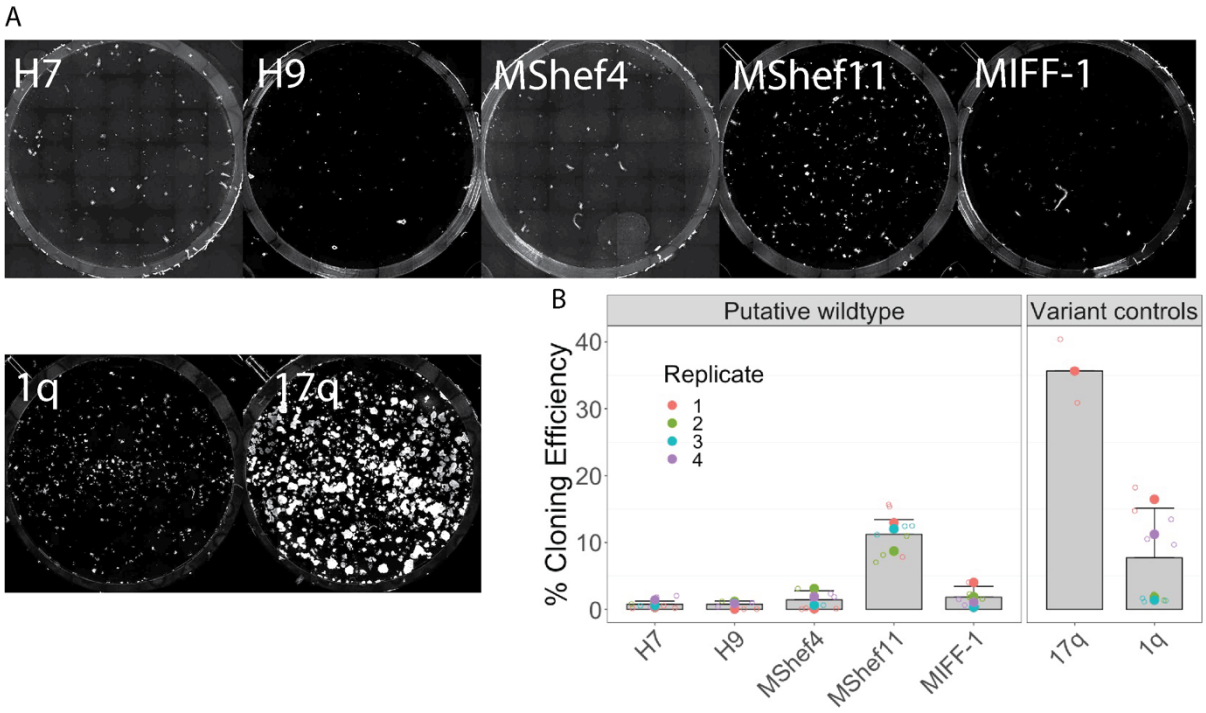


Figure 3.2 Cell line Cloning efficiency

A) Representative stitched images of entire wells of a 12 well plate wells stained for DNA (white). B) Cloning efficiencies of hPSC lines to be used for sequencing, as well as variant clones of H7, harbouring duplications of chromosomes 17q and 1q respectively. Bars represent mean cloning efficiencies of all biological replicates of a given line +/- SD. Filled points denote mean values for a given biological replicate. Empty points denote cloning efficiencies calculated for individual technical replicate wells of each biological replicate.

To further characterise lines, at the point of freezing, cells were karyotyped via analysis of 20 metaphase spreads, (95% chance of detecting 14% mosaicism (Baker, Adam J. Hirst, *et al.*, 2016)). All lines appeared karyotypically normal (Figure 3.3, Table 3.1).

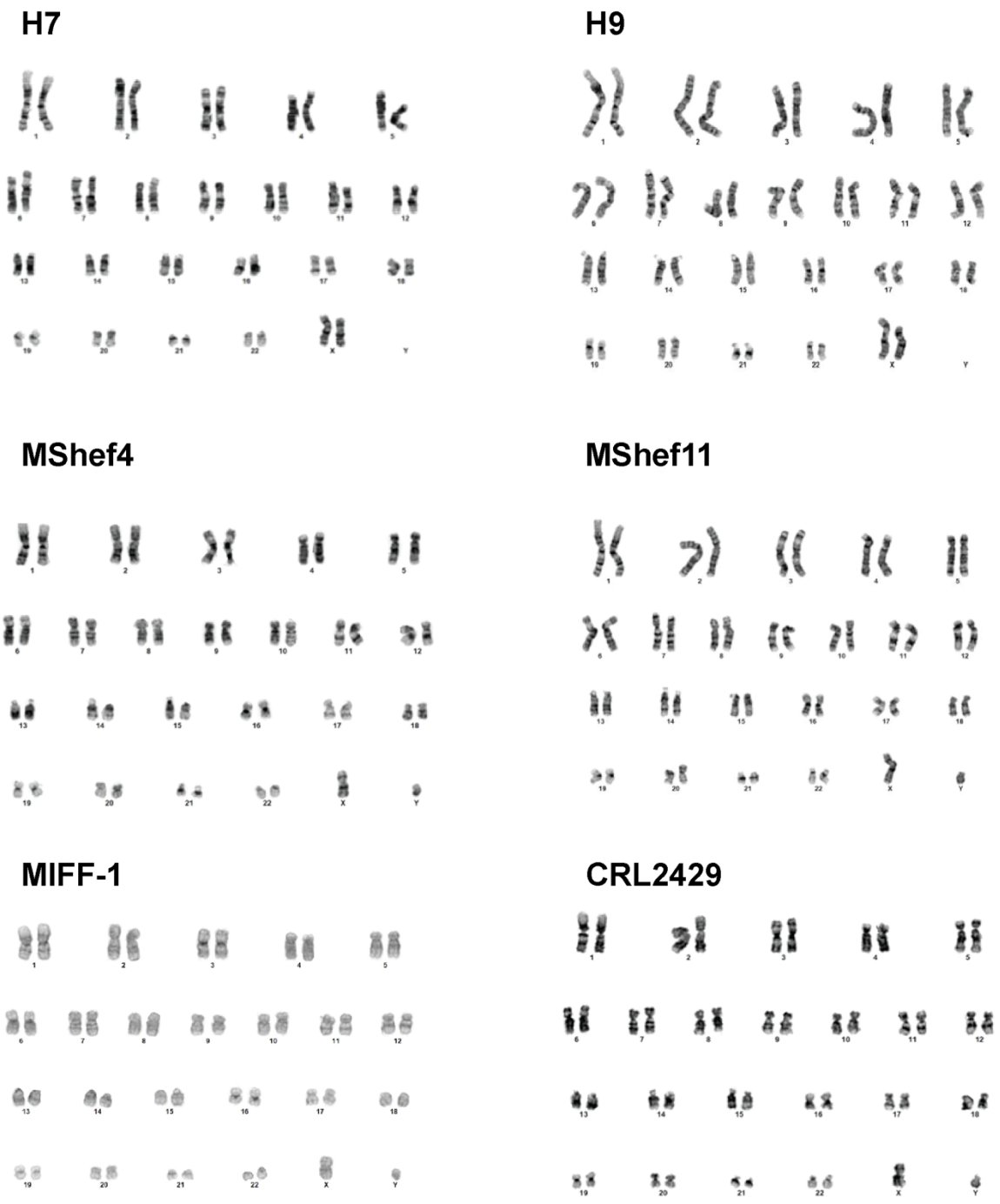


Figure 3.3 Representative karyotypes of all lines used in the study.

All karyotyping was carried out by clinical cytogeneticists at Sheffield Diagnostic Genetics Service.

Table 3.1 Cell line karyotypes at point of freezing.

Numbers of metaphases scored denoted in parentheses. Passage numbers indicated, numbers represent passages in continuous culture, '+' denotes cryopreservation, "/" denotes cloning events.

Cell Line	Cell Type	Karyotype	Passage
H7	hESC	46,XX[20]	P7+2+4+4
H9	hESC	46,XX[20]	P47/5+3
MShef4	hESC	46,XY[20]	P11+17+8+3+3
MShef11	hESC	46,XY[20]	P12+3+3+3+3+2
MIFF-1	hiPSC	46,XY[20]	P2+5+5+5
CRL2429	hFF	46,XY[20]	P4+2+1

Karyology has a relatively coarse resolution (~5-10Mb) and as such, many commonly observed genetic changes in hPSC can pass below this threshold and therefore remain undetected. To complement karyology, a targeted qPCR method was used to detect changes in copy number of genes *BCL2L1* and *MDM4* on the commonly amplified regions 20q11.21 and 1q32, respectively (Fig3.4A, B). Again, no population of variants harbouring these aberrations was detected in any of the cell lines by qPCR.

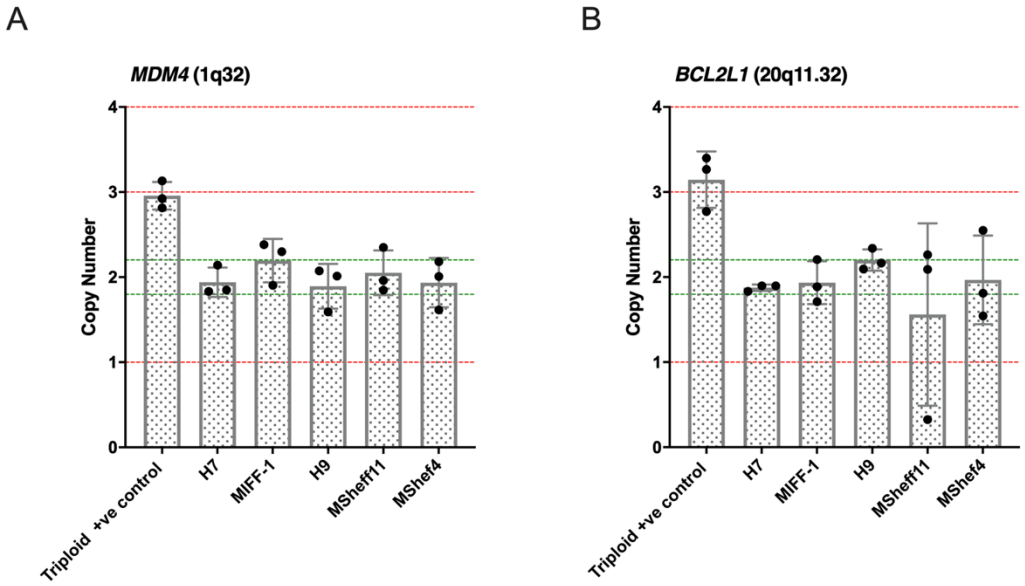


Figure 3.4 hPSC line Copy numbers at loci of frequent CNVs

Copy numbers of genes on frequently duplicated regions 1q32 (A) and 20q11.21 (B), as determined by qPCR of gDNA, relative to a known diploid control sample. Known triploid line included as a positive control. Bar heights represent mean values of 3 triplicate wells, points represent values of individual wells, error bars +/- SD.

In addition to targeted qPCR, to gain a genome-wide picture of CNVs, gDNA extracted from each of the cell lines was used in SNP arrays to detect the presence of small duplications and deletions (Table 3.2). All lines harboured some CNVs, ranging from ~4Kb to ~8Mb in length. No CNVs were common across lines. To determine the risk of identified CNVs contributing to oncogenic transformation of cells, genes in CNV regions were cross referenced with the COSMIC Cancer Gene Census (CGC) database (Sondka *et al.*, 2018). Of the 29 genes implicated in CNVs across the 5 lines, *RUNX1T1 (gained in H9)*, was the only hit in the CGC database. *RUNX1T1* is classified as a *Tier 1* CGC gene, meaning mutations therein have been reproducibly found in cancers and demonstrated to contribute to oncogenic transformation (Sondka *et al.*, 2018). *RUNX1T1* is commonly implicated in *RUNX1T1/RUNX1* gene fusions in acute myeloid leukaemia, leading to aberrant repression of genes under RUNX1 control (Al-Harbi *et al.*, 2020). In other pathologies, *RUNX1T1* is typically under-expressed, leading to increased proliferation in gastric and ovarian cancers (Hu *et al.*, 2022). It is therefore unlikely that *RUNX1T1* copy gain, endows a fitness advantage on this cell line.

MShef11 exhibited an increased cloning efficiency compared to other lines (Figure 3.2). SNP array analysis identified a 96kb gain in MShef11 on Chromosome 12p (Table 3.2), which is commonly involved in whole-chromosome or whole-arm gains in both hPSC and germ cell tumours (Castedo, 1993; Draper *et al.*, 2004). This 96Kb CNV encompasses the *SINHCAF* gene, shown to be essential for rapid proliferation and maintenance of the pluripotent state in mESC (Streubel *et al.*, 2017) and recently proposed as a possible driver gene for 12p gains in germ cell tumours (Jhuang *et al.*, 2022), raising the possibility that *SINHCAF* gain endows MShef11 with a fitness advantage, resulting in increased cloning efficiency. As there is no documented effect of *SINHCAF* on the DDR or apoptosis, MShef11 was not excluded from subsequent experiments.

Table 3.2 SNP Array summary of gained and lost regions in all 5 hPSC lines at point of freeze

Cell Line	Locus	Type	Length (Kb)	Genes
H7	2q37.3	LOSS	209.4	No genes
	3p21.31	GAIN	42.8	No genes
	3q26.1	GAIN	250.0	No genes
	12p12.1	GAIN	4.2	*SOX5*
H9	5p15.31	GAIN	49.9	*ADCY2*
	7q11.21	GAIN	655.5	No genes
	8p23.1	GAIN	51.2	*AC105233.4*
	8q21.3-8q22.1	GAIN	2009.6	*RUNX1T1, NR_125827.1, TRIQK, MIR8084, C8orf87, LINC00535* (partial)
	10q22.1	LOSS	27.0	*DNAJB12*, *MICU1*
	14q23.2	GAIN	295.6	*PPP2R5E*, *WDR89*, *SGPP1*
	14q23.3	GAIN	374.9	No genes
MShef4	7q35	LOSS	96.6	*CNTNAP2*
	19p12	LOSS	102.3	*ZNF85*
	Xp22.33	GAIN	7.2	*CRLF2Y*, *CRLF2*
MSheff11	2q37.3	LOSS	195.9	No genes
	12p11.21	GAIN	95.9	*SINHCAF*, *DENND5B*
MIFF-1	3p26.1	LOSS	101.2	*SUMF1*
	3q21.3	GAIN	130.4	*CHCHD6*
	10q21.1	LOSS	49.3	*PCDH15*
	12q24.33	GAIN	236.1	*POL5E*, *PXMP2*, *PGAM5*, *ANKLE2*, *GOLGA3*, *CHFR*
	13q13.3	LOSS	47.5	No genes
	13q31.3	LOSS	9.3	No genes
	16q21	LOSS	430.7	No genes

3.2.2. Optimising a universal protocol for the differentiation of characterized cell lines.

A plethora of differentiation protocols are available for driving hPSC to specific cell fates (Hong and Do, 2019; Mennen *et al.*, 2022). For the purposes of my INDUCE-seq experiments, I wanted a relatively quick and inexpensive protocol that would drive hPSC out of the pluripotent state in a robust and reproducible manner. To this end, I chose a five-day mesodermal differentiation protocol, whereby cells are seeded in mTeSR1 pluripotent cell medium for 24hours, then replenished with E6 medium (E6 represents E8 pluripotent cell medium without TGFβ and FGF2) supplemented with the canonical WNT agonist, CHIR99021 (Lippmann *et al.*, 2014).

Different cell lines have different levels of endogenous WNT signalling (Strano *et al.*, 2020). Activating WNT signalling with a given concentration of CHIR99021 in one cell line, may therefore be inadequate in another. I therefore tested CHIR99021 at 3 concentrations in each of my five cell lines, with the aim of identifying a universal concentration which would give robust exit from the pluripotent state across all lines (Figure 3.5).

Following 5 days culture in each condition, cells were fixed and stained for the pluripotency-governing transcription factor, POU5F1 (also known as OCT4 and from herein referred to as OCT4) and the mesendodermal-associated transcription factor Brachyury (T). Images were analysed using a custom CellProfiler pipeline to approximate the numbers of OCT4+ve and T+ve cells in each condition, gated against a secondary staining control (Figure 3.5). Treatments with 3μM CHIR99021 yielded ~100% T+ve MShef11 cells (Figure 3.5 A, B), however failed to induce T expression robustly in the remaining 4 cell lines (Figure 3.5 E). Similarly, 3μM CHIR99021 treatment lead to a reduction in OCT4 integrated intensity (Figure 3.5 A, B), but a significant OCT4+ve population remained in H7, H9, and MShef11 (Figure 3.5 D). By contrast, 10μM CHIR99021 treatment gave typically high proportions of OCT4-ve, T+ve cells but was toxic to several cell lines, resulting in low cell numbers by D5 (Figure 3.5 C). 5μM CHIR99021 was well tolerated by most cell lines (NB low cell counts in H7 and MIFF-1 can be accounted for by washing away during fixation), gave ~100% T+ve cells across all lines, and consistently low levels of OCT4+ve cells (Fig3.5). I therefore selected 5μM CHIR99021 as a suitable concentration for the differentiation protocol used for sequencing.

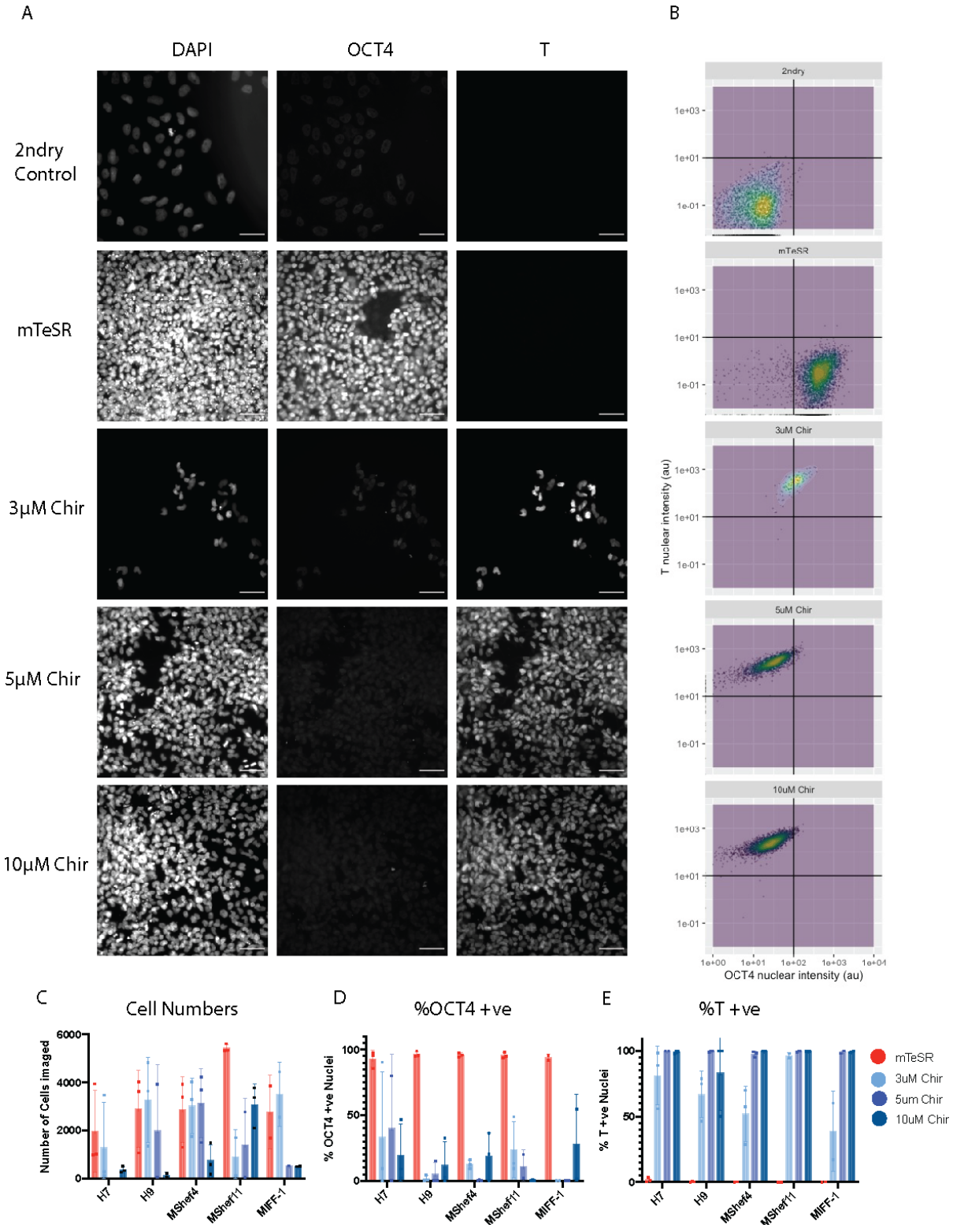


Figure 3.5 Differentiation optimisation

A) *Representative fluorescence images of MShef11 cells cultured over 4 days in either mTeSR medium (pluripotent control), or E6 medium supplemented with increasing levels of CHIR99021. Cells nuclei stained with DAPI, counterstained for OCT4 and T. Top panel is a 2ndry antibody only control well. B) Representative scatterplots of individual nuclei's OCT4 /T integrated intensity. lines denote gates, positioned using 2ndry control well, to categorise +ve/-ve nuclei. C) Cell numbers D) %OCT4 +ve cells E) %T +ve cells for all 5 hPSC lines. Points represent values of individual biological replicates, bar height represents the mean value for 3 biological replicates. +/-SD.*

The above differentiation protocol yields high density cultures by the end point. Quantifying proportions of T/OCT4+ve cells *in situ* using immunocytochemistry can be challenging due to the tendency of cells to detach during fixation and subsequent washing, and the inaccuracy of nuclear segmentation when handling high-density images (Figure 3.5 A). Culture density affects cell differentiation (Wilson *et al.*, 2015), and given the efficacy of the protocol used here, I sought to avoid lowering cell density solely for the purpose of image analysis. I therefore adapted the staining protocol for fixed cells in suspension, to be analysed via flow cytometry. To this end, H7 cells were dissociated at day 5 of the differentiation protocol, fixed in suspension, then subjected to a similar intracellular staining protocol as used in the immunocytochemistry experiment. Analysis via flow cytometry allowed accurate resolution of individual cells and revealed near-pure populations of OCT4+ve/T-ve and OCT4-ve/T+ve cells in the pluripotent and differentiated conditions, respectively (Figure 3.6 A, B). Based on this data, I decided to use flow cytometric analysis of OCT4/T expression as a robust readout of cell state in INDUCE-seq experiments.

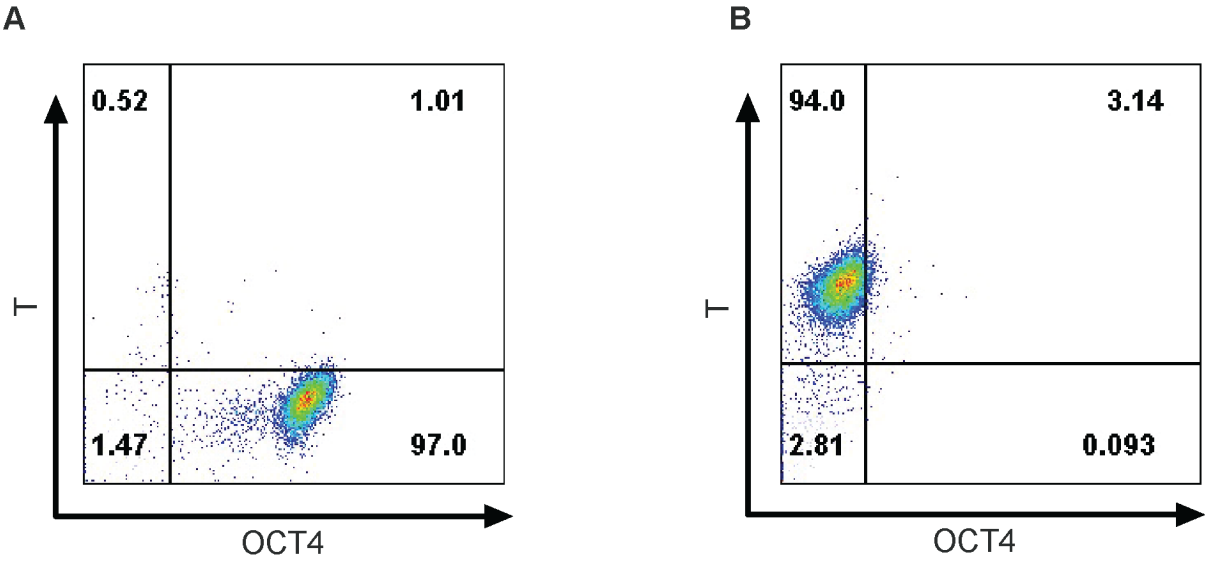


Figure 3.6 Adaptation of OCT4/T staining for flow cytometry

FACS plots of OCT4/T staining intensity in H7 cells Cultured for 5 days in A) mTeSR or B) E6 + 5μM CHIR99021. Quadrants position based on 2ndry staining control, numbers denote percentages in each quadrant. Points represent individual cells, gated on SSC/FSC/Trigger pulse width values.

3.2.3. Confirming the Pluripotent DNA damage phenotype.

Previous studies have identified higher levels of DNA damage in pluripotent cells than their differentiated derivatives (Halliwell *et al.*, 2020; Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018). I wanted to corroborate this result in my system ahead of sending cells for sequencing. To this end, I cultured hPSC lines under pluripotent and differentiated conditions, and assessed levels of DNA damage, using γH2AX as a surrogate for DSB formation.

To validate the sensitivity of γH2AX immunofluorescence, I treated cells with Camptothecin (CPT) as a positive control for DSB induction (Forment and Jackson, 2015). CPT is a selective TOP1 poison which traps TOP1 cleavage complexes, which, in turn, are processed to DSBs, primarily in S-phase, upon encounter with DNA replication forks (Pommier, 2006). All cell lines, in both pluripotent and differentiated conditions, showed an increase in γH2AX foci following CPT treatment, demonstrating the sensitivity of the assay (Figure 3.7 A, B).

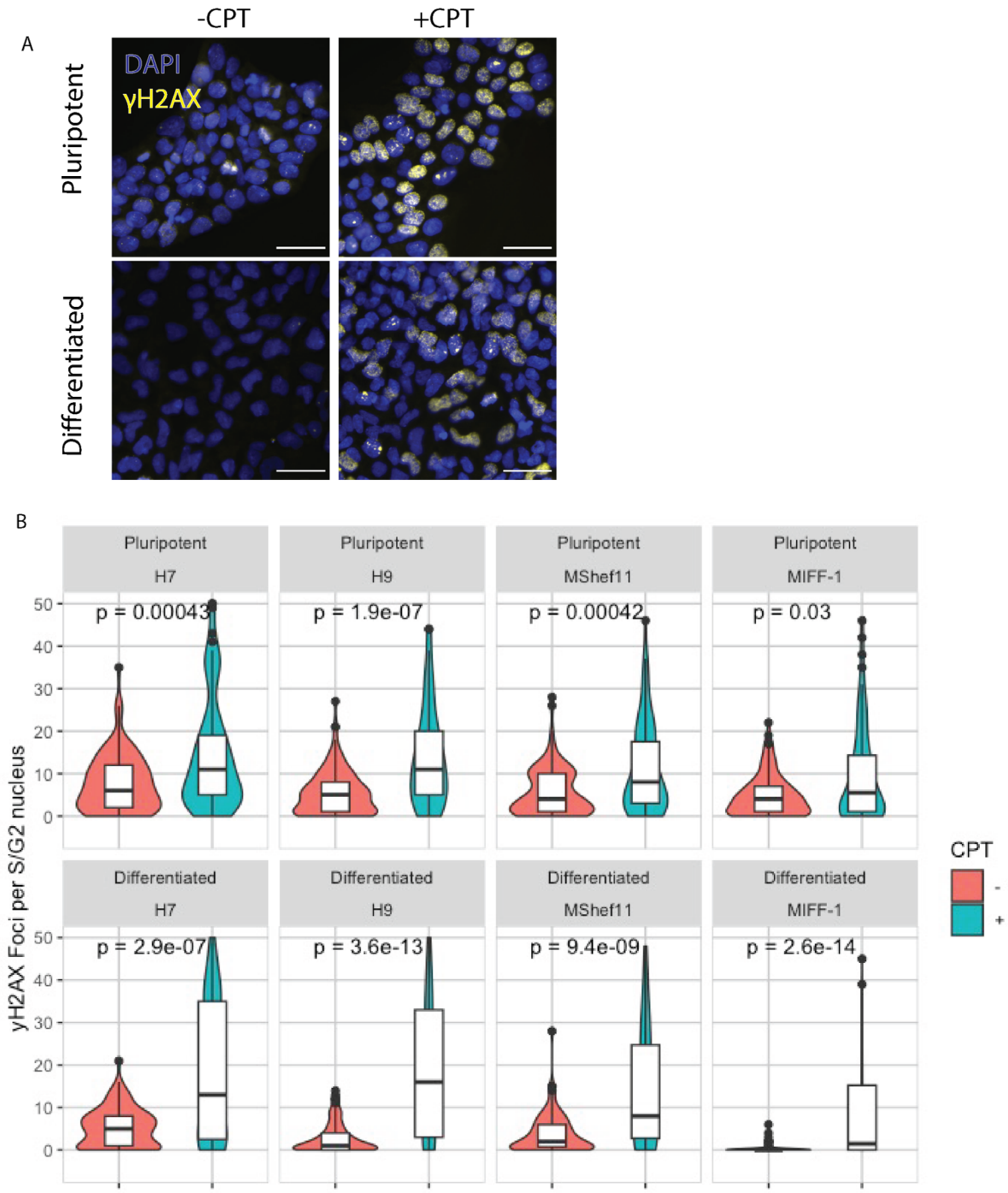


Figure 3.7 γH2AX expression increases on CPT treatment

A) Representative images of H7 cells cultured in pluripotent or differentiated conditions +/- 100nM CPT (scale 40μm). B) Violin plots of γH2AX foci per S/G2 nucleus in pluripotent and differentiated conditions +/- CPT. Overlaid boxplots denote median values, data from 1 experiment (Wilcoxon test, n=200 cells)

Consistent with previous reports, 3 out of 4 lines tested showed higher levels of γH2AX in pluripotent than differentiated states (Figure 3.8 A-C). Surprisingly, MShef11 showed comparable levels of DNA damage in pluripotent and differentiated states.

Previous studies have attributed DNA damage in hPSC to DNA replication. Halliwell and colleagues report increased DNA damage in pluripotent versus differentiated states, when comparing only S/G2 populations of each cell type (J. A. Halliwell *et al.*, 2020), whilst others find DNA damage to be proportional to the fraction of cells in S/G2 (Simara et al., 2017; Vallabhaneni et al., 2018). To determine whether DNA damage is increased in hPSC when comparing only replicative cells, I subset cells in S/G2 phases of the cell cycle based on their DAPI staining intensity (3.8 D). When comparing cells in S/G2, only H7 and H9 showed increased γH2AX expression, whereas MShef11 and MIFF1 expressed comparable levels between cell states (Figure 3.8 E). This discrepancy between cell lines could account for the different DNA damage phenotypes reported in the aforementioned studies.

Importantly, when comparing cells in all phases of the cell cycle, more γH2AX foci were present in 3 out of 4 pluripotent lines when compared with their differentiated derivatives, consistent with previous reports, indicating that the differentiation protocol used here is suitable for investigating differences in DNA damage between cells states.

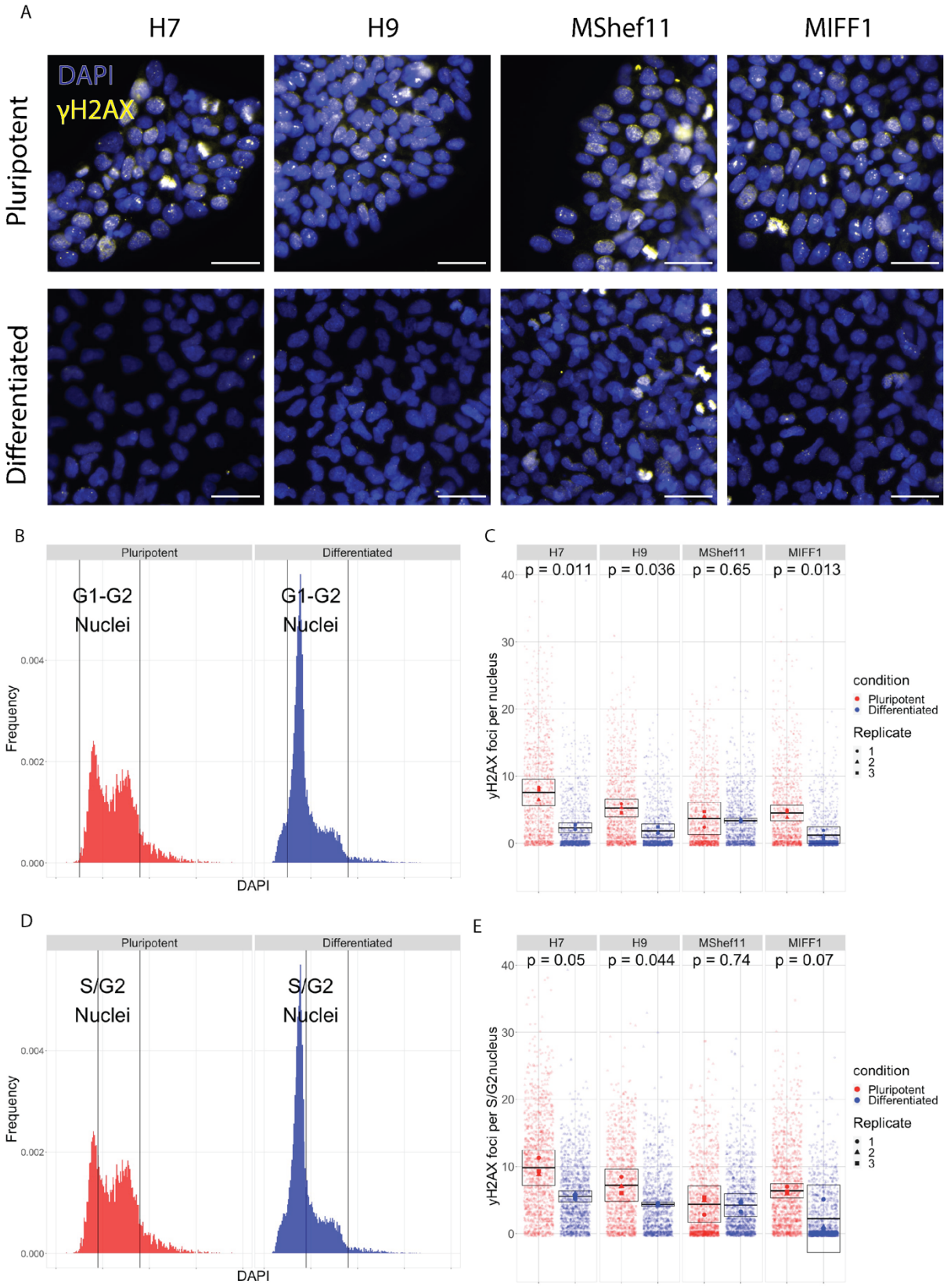


Figure 3.8 γH2AX expression in pluripotent vs differentiated cells

A) Representative γH2AX immunofluorescence images of each cell line cultured in pluripotent or differentiated states (scale 40μm). B, D) Representative gating strategies for data in C and E. histogram plots of nuclear DAPI

integrated intensity in pluripotent and differentiated conditions. Gate in D encompasses all nuclei, gate in F approximates only S/G2 nuclei. C ,E) Super plots of γH2AX foci per nucleus (C) or foci per S/G2 nucleus (E) in pluripotent (red) vs differentiated (blue) conditions. Small points denote individual cells, larger bold points denote mean values of all cells of a biological replicate (n>200 cells). Crossbars represent mean values +/-SD of 3 biological replicates. (Paired t-test n=3 biological replicates)

3.2.4. Confirming a low-oxygen phenotype

Previous studies have reported enhanced growth, higher expression of pluripotency-associated markers, fewer chromosomal breaks and a reduced mutation rate in hPSC cultured under low oxygen (Forsyth *et al.*, 2006; Forristal *et al.*, 2010; Närvä *et al.*, 2013; Thompson *et al.*, 2020). I ultimately sought to compare DSB distributions in hPSC cultured under atmospheric (~20%) versus low (5%) oxygen conditions, but first needed to corroborate a "low-oxygen phenotype".

I selected growth rate as a rapid, inexpensive metric for determining hPSC performance under low-oxygen conditions. I carried out 5-day growth curves of the MIFF-1 hPSC line under atmospheric O_2, versus continuous 5% O_2, cultured using a hypoxic workstation (Figure 3.9).

Growth curves were carried out at 3 increasing seeding densities. Increased cell numbers were observed in cells cultured in 5% O_2 at days, 2, 4 and 5 when seeding at low density (Figure 3.9 A, B). At medium seeding density, 5% O_2 conditions yielded higher cell counts at days 3 and 4, but by day 5 both conditions were comparable, likely due to 5% O_2 condition reaching confluence and therefore plateauing at day 4. At high density, 5% O_2 yielded higher cell numbers by day 2, comparable cell counts at days 3 and 4, and significantly fewer cells by day 5. The apparent reduction in cell number at high density could be a result of cells becoming over-confluent and undergoing apoptosis or could be a technical issue with resolving nuclei in very high-density images (Figure 3.9 A). In summary, 5% O_2 culture yielded largely increased growth rates compared with atmospheric oxygen tension. Having confirmed an enhanced growth rate in low-oxygen cultured cells, I reasoned that this 5% O_2 culture-system was suitable for subsequent sequencing experiments.

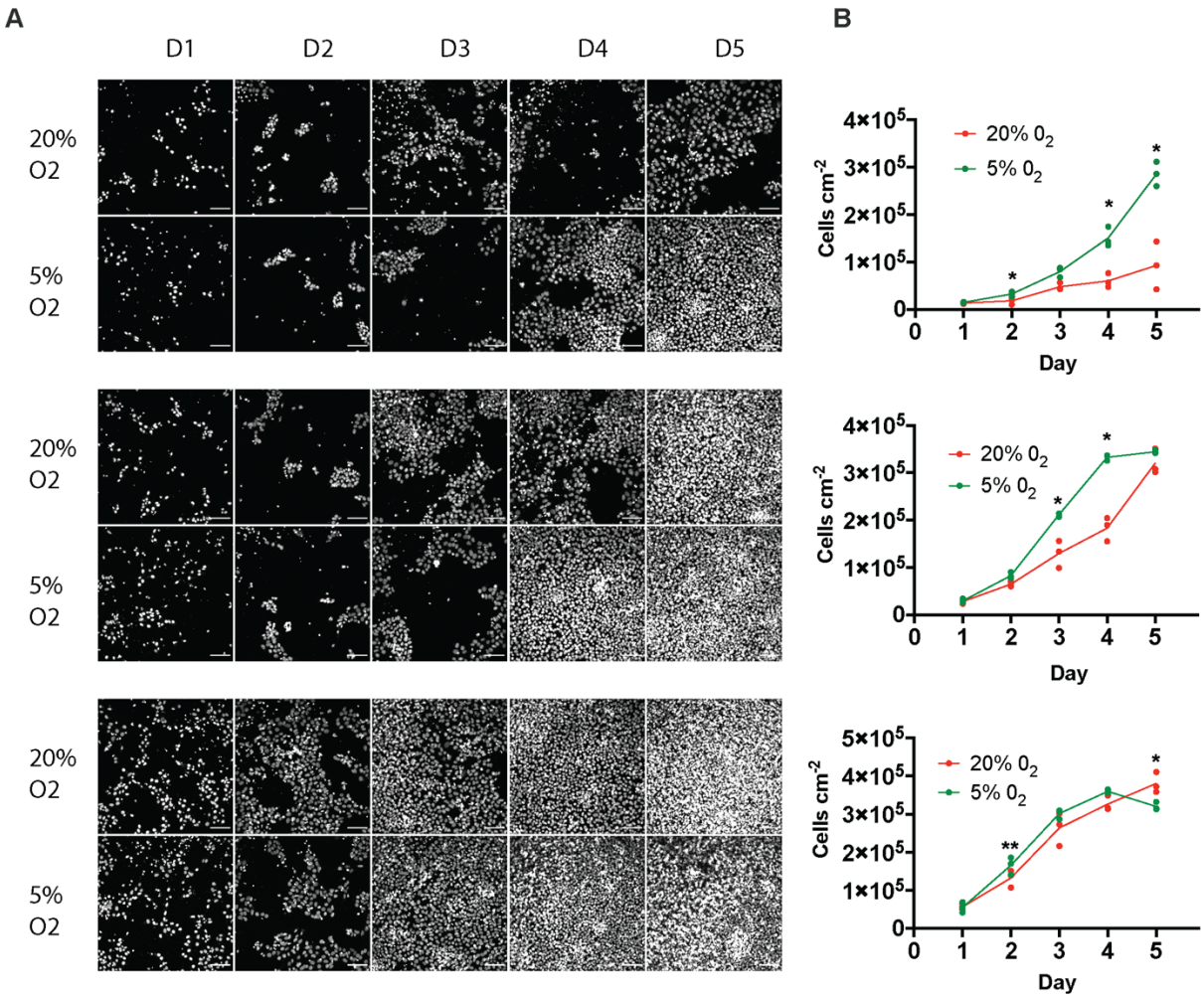


Figure 3.9 20% vs 5% oxygen growth rates

*A) Representative DAPI-stained images of MIFF-1 cells cultured under ~20% oxygen (atmospheric) or 5% oxygen (low), over a 5-day time course, seeded at 11×110^4 cells cm^{-2} (top), 22×10^4 cells cm^{-2} (middle) or 44×10^4 cells cm^{-2} (bottom). B) Quantification of (A). Line position represents mean value of 3 biological replicates, each point represents the cell density of an individual biological replicate. (n=3 biological replicates *p<0.05, **p<0.01. paired t-test).*

3.2.5. Optimising INDUCE-seq seeding protocol for hPSC

To minimise the introduction of artefactual DNA breaks in samples prior to sequencing, INDUCE-seq library preparation uses formaldehyde-fixed cells adhered to poly-D-lysine coated 96-well plates (Dobbs *et al.*, 2022). The standard seeding protocol calls for single cell dissociation of the samples, resuspension in PBS and adherence onto poly-D-lysine coated plates for 10 minutes, followed by fixation. hPSC are inherently sensitive cells, particularly following single cell dissociation (Watanabe *et al.*, 2007), and I reasoned that resuspending in basal culture medium (DMEM) rather than PBS would be more supportive of cell survival. I therefore compared the number of fixed cells in poly-d-lysine coated wells, following resuspension in PBS or DMEM, and observed no significant difference between the two

(Figure 3.10 A, B) and hence determined that seeding and fixing in DMEM is suitable for plate preparation. Cells are seeded at a density of $3x10^5$ cm^{-2} however, the density of fixed cells imaged were less than 50% of this value (Figure 3.10 B). I speculated that hPSC may require longer to adhere to the matrix and therefore tested increasing adherence times from 5-20 minutes and compared cell densities (Figure 3.10 C, D) a progressive increase in cell density was observed as adherence time increased, with maximum cell density at 20 minutes. I therefore took 20 minutes forward as the seeding time for the sequencing experiment.

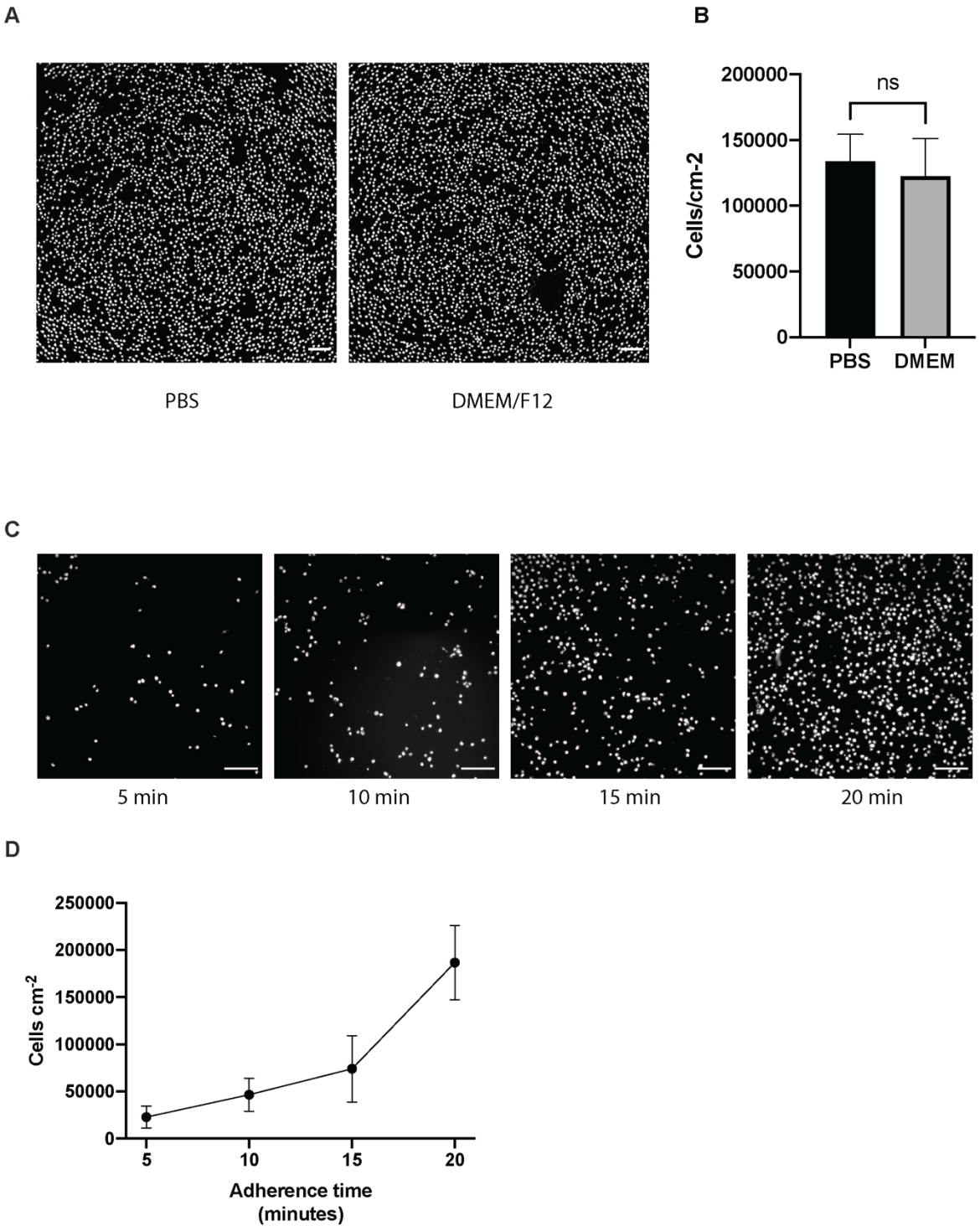


Figure 3.10 Seeding optimisation

A) Representative images of DAPI-stained cells seeded following INDUCE-seq protocol either in PBS or DMEM/F12. Scale=100µm. B) Quantification of cell density in (A). Bars represent mean values of 1 experiment +/- SD (t-test n=36 fields). C) Representative DAPI-stained images of cells seeded in DMEM/F12, allowed to adhere for varying lengths of time prior to fixation. Scale=100µm. D) Quantification of cell density following cell adhesion for the indicated times. Points denote mean values per time point of 1 experiment +/- SD (n=25 fields).

3.2.6. INDUCE-seq experimental setup

Having optimised experimental conditions, sequencing material was prepared as follows: all 5 hPSC lines were seeded in triplicate from independent cultures, as per the established differentiation protocol, and cultured for 4 days post seeding in either pluripotent 20% O_2, pluripotent 5% O_2, or differentiated 20% O_2 conditions. Given low-oxygen culture yields an enhanced growth rate (Figure 3.9), seeding for the 5% O_2 condition was reduced by 33%, to prevent cultures from becoming overconfluent. All conditions were seeded into 6 well plates, with surplus wells for cells pellets, to enable any future parallel sequencing studies.

At day 5, brightfield images were taken of each line in each condition (Figure 3.11 A), and cell densities were approximated from haematocytometer counts (Figure 3.11 B). Notably, across all cell lines, the 5% O_2 condition did not exhibit a stem-like morphology with distended cells, at a much lower density than the 20% O_2 counterparts (Figure 3.11 A). This apparent scarcity of cells was reflected in the cell counts, where the 5% O_2 condition yielded typically less than 50% of the cells obtained from 20% O_2 culture (Figure 3.11 B). This was most pronounced for the H7 line, where there were no recoverable cells in 5% O_2 by day 5 and the experimental condition was therefore lost altogether.

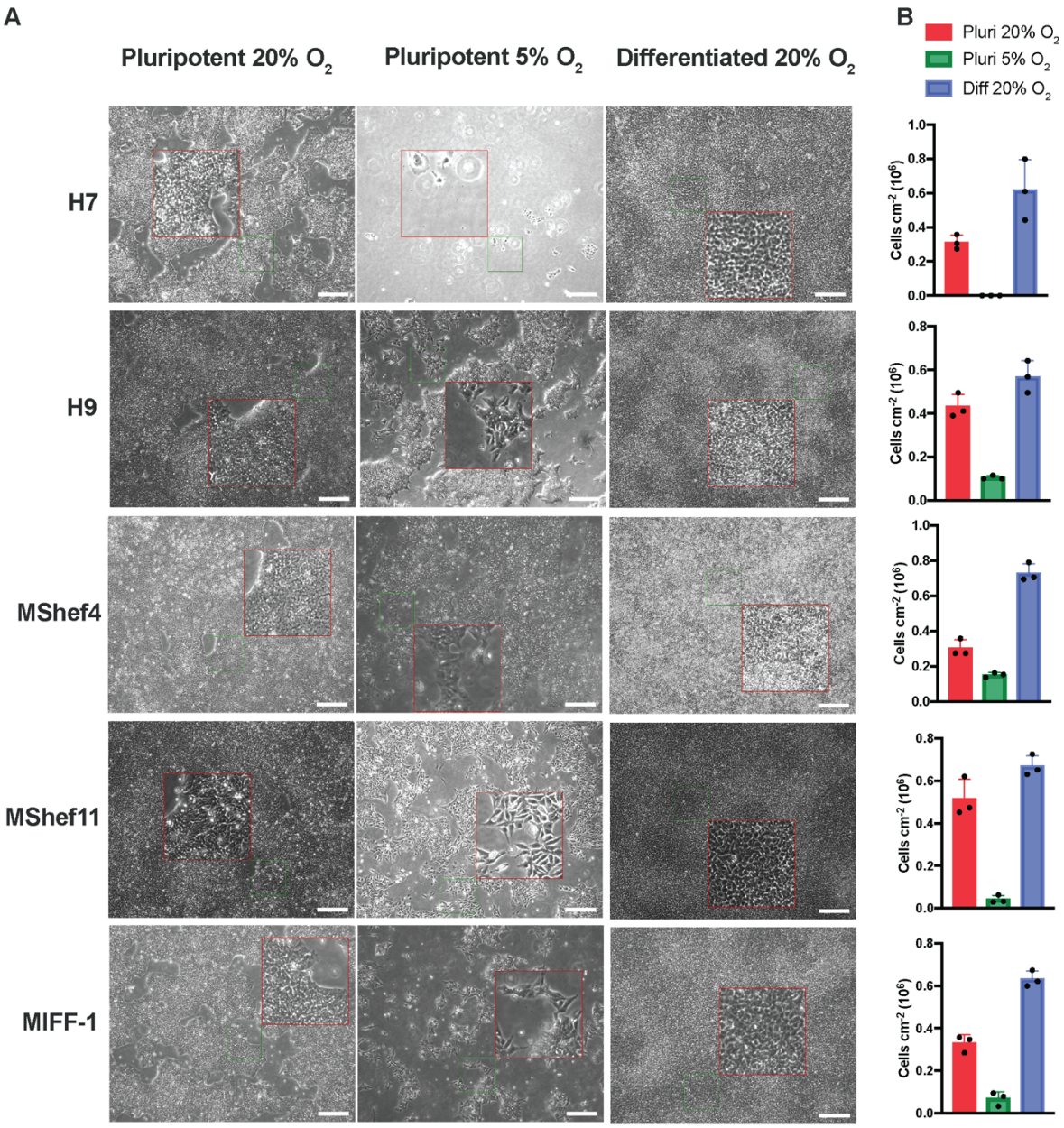


Figure 3.11 INDUCE-seq cell morphology and density

A) Representative brightfield images of each cell line cultured under each condition at day-5. Regions of interest, magnified 2.5× in red boxes. Scale=300μm. B) Cell counts obtained from each cell line in (a), in each condition. data shown are mean values of 3 independent replicates +/- SD.

Despite low cell counts and abnormal morphology, the 5% O_2 condition expressed near 100% OCT4+ve T-ve cells across all lines, akin to the 20% O_2 condition (Figure 3.12 A-C). Moreover, the differentiated condition yielded near 100% OCT4-ve cells across all lines, and ~100% T+ve cells in all lines except MShef4. Whilst only ~70% of MShef4 cells in the differentiated condition were T+ve, nearly all were OCT4-ve. In summary cells of each condition expressed the expected markers.

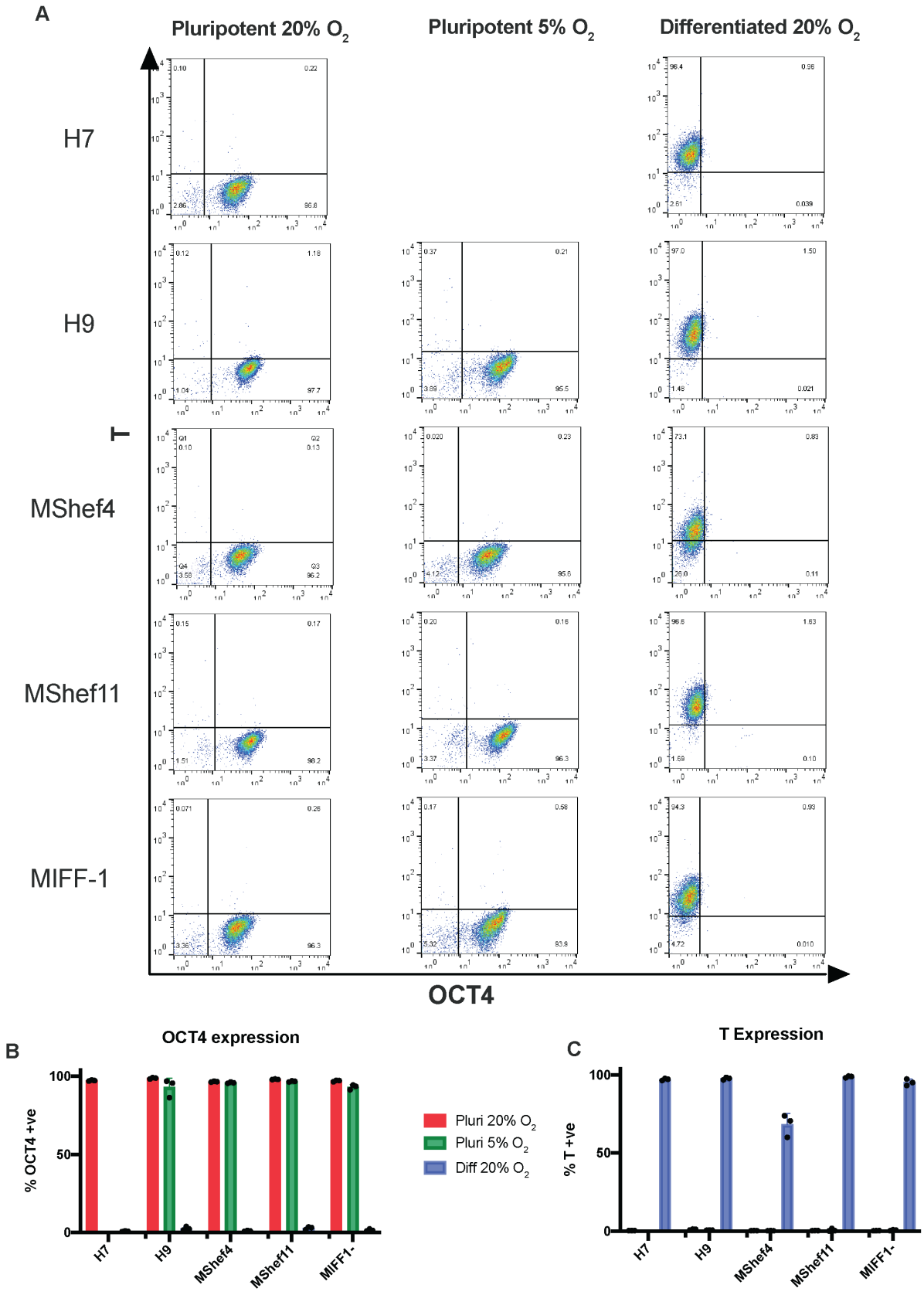


Figure 3.12 INDUCE-seq OCT4/T expression

A) Representative FACS plots of each cell line under each condition at day-5, co-stained for OCT4/T. B, C) Summary data of the percentage of cells for each line (coloured by condition) positive for B) OCT4 or C) T. Bar height represents the mean values of 3 independent replicates +/- DS. Individual points represent values of each biological replicate.

3.3. Discussion

This chapter aimed to characterise hPSC lines and validate experimental conditions ahead of the large scale INDUCE-seq experiment.

3.3.1. Characterisation of Cell lines

I have established the genetic status of all cell lines to be sequenced. This is particularly important for sequencing DSBs, as genetically variant cultures may exhibit phenotypic differences from their wild-type counterparts, most significantly in terms of DNA damage response (DDR) and apoptotic threshold (Avery *et al.*, 2013; Amir *et al.*, 2017). In addition, copy number changes lead to over/underrepresentation of sequencing reads at implicated genomic regions. Indeed, this phenomenon forms the basis of techniques like eSNP-karyotyping, which uses allelic expression ratios from RNA-seq datasets to infer chromosome copy numbers (Weissbein *et al.*, 2016).

As a functional test of cell line's fitness, I determined cloning efficiencies for each of the lines, all of which fell within a, very broadly defined, "normal range" for wild-type hPSC lines. Notably however, MShef11 cloned at higher efficiencies than the other four hPSC lines. This may reflect normal variation in line's cloning efficiency or, alternatively, could be due to a genetic or epigenetic change in this cell line. Given the crudeness of the assay and lack of control known variant clones for each cell line it's not possible from this data alone, to determine the cause of this increased cloning efficiency.

For genetic characterisation, I used a combination of genome-wide and site-specific methods, namely, karyology, qPCR and SNP-array. Cells appeared normal by karyology and qPCR. However, all lines harboured a small number of CNVs as determined by SNP array. MShef11 harboured a duplication of *SINHCAF*, a component of the sin3a-Hdac complex required for rapid proliferation and self-renewal in mESC (Streubel *et al.*, 2017), proposed as a driver of chromosome 12 aneuploidies in germ cell tumours, wherein knockdown leads to cell differentiation and reduced proliferation (Jhuang *et al.*, 2022). Gain of *SINHCAF* may cause the observed increase in cloning efficiency observed in MShef11 and affect cell fitness. Given MShef11 cells were still able to differentiate in my system, the cell line was not excluded from the sequencing experiment, however, this genetic change will be considered in downstream analysis. Crucially, genes on the regions implicated in CNVs have no documented effect on DDR or apoptotic threshold. Moreover, their positions are now known to me and can be accounted for in downstream sequencing analyses.

This genetic characterization, whilst rigorous compared to many published hPSC studies (Halliwell *et al.*, 2020), is by no means exhaustive. Point mutations and other non-recurrent small scale indels would pass undetected by the above characterisation (Merkle *et al.*, 2017; Avior *et al.*, 2019). In order to detect such small-scale changes, additional RNA-seq or whole genome sequencing would be required to detect genic or genome-wide mutations, respectively, both of which carry a significant cost and thus were omitted from this study.

A notable shortcoming of the above characterisation is that it lacks any epigenetic screening. In addition to genetic changes, hPSC have been shown to acquire recurrent epigenetic aberrations over prolonged culture, such as DNA hypermethylation to stress regulator genes or tumour suppressor genes (Calvanese *et al.*, 2008; Konki *et al.*, 2016), X-chromosome re-activation (Bar *et al.*, 2019) or loss of imprinting (Pick *et al.*, 2009; Bar *et al.*, 2017). Whilst X-chromosome inactivation and loss of imprinting can be detected from RNA-seq datasets, changes in methylation status require costly and time-consuming whole-genome bisulphite sequencing or equivalent, and as such, were beyond the scope of this project.

3.3.2. Differentiation protocol optimisation

I have optimised a mesodermal differentiation protocol which yields near 100% OCT4-ve populations for all 5 cell lines, and near 100% T+ve populations for 4/5 lines. Removing FGF2 and TGFβ from culture medium alone is sufficient to drive exit from the pluripotent state in hPSC, however this typically yields neuroepithelial cells, i.e. ectodermal lineage, and can take 6 or more days (Lippmann *et al.*, 2014). Previously our group had used a mesodermal differentiation protocol to study differences in DNA damage abundance between pluripotent and differentiated states (Halliwell et al., 2020), which, in addition to removal of FGF2 and TGFβ, uses activation of WNT signalling via supplementation with the selective GSK3B inhibitor CHIR99021 to direct cells towards a mesodermal fate (Lian *et al.*, 2012).

I titrated CHIR99021 concentration in this protocol, and found the lowest dose of 3μM, used by Lippman and colleagues (Lippmann *et al.*, 2014) yielded persistent, low-level OCT4 expression in certain cell lines, whilst the highest concentration of 10μM, used by Halliwell and colleagues (Halliwell et al., 2020), was toxic to several cell lines, resulting in cell death. This variable response of different cell lines to CHIR99021, likely reflects variable levels of endogenous WNT signalling between lines (Strano *et al.*, 2020). If a given cell line has relatively low levels of endogenous WNT signalling, it may require higher concentrations of CHIR99021 to induce differentiation. Conversely, high doses of CHIR99021 and, therefore GSK3B inhibition, has been documented as cytotoxic in hPSC lines with relatively low S/G2 populations of the cell cycle, whereby GSK3B inhibition upregulates cyclin A expression

prompting premature S-phase entry in and ultimately replication stress and cell death (Laco *et al.*, 2018). 5μM CHIR99021 supplementation appeared to balance high differentiation efficiency with low cytotoxicity across all cell lines and thus was selected for the sequencing differentiation protocol.

3.3.3. Pluripotent cells harbour higher frequencies of γH2AX foci than their differentiated counterparts.

With a differentiation protocol defined, I used γH2AX immunofluorescence to confirm that 3 of 4 cell lines cultured in the pluripotent state, harbour more DSBs than those cultured in the differentiated state, consistent with previous reports (Vallabhaneni *et al.*, 2018; J. A. Halliwell *et al.*, 2020).

Existing studies have generated contradictory results on whether DNA damage is increased in replicating pluripotent cells versus replicating differentiated cells (J. A. Halliwell *et al.*, 2020), or whether increased DNA damage in the pluripotent state is simply proportional to the increased fraction of replicating cells (Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018). Given the differentiation protocol used in this study was based on that used by Halliwell and colleagues (2020), I anticipated DNA damage would be increased in the pluripotent state when assaying only replicating cells. Surprisingly, this was only true for 2 of 4 lines tested, i.e., the H9 and H7 cell lines showed increased damage in S/G2, whereas the MIFF-1 and MShef11 lines showed comparable levels of DNA damage between pluripotent and differentiated states when comparing S/G2 cells. This suggests DNA damage may be cell line-specific.

However, there are several limitations to the techniques used here. Firstly, γH2AX foci presence is only a surrogate for DSB presence. Phosphorylation of H2AX to form γH2AX is one of the first steps in the DDR upon DSB formation (Polo and Jackson, 2011). Detection of breaks using γH2AX is dependent on an in-tact and efficient DDR, and thus assumes that DDR activities are comparable across cell types. The neutral comet assay, which measures the movement of relaxed chromatin, due to DSBs, through agarose upon electrophoresis, is a more direct assay, but it cannot easily be coupled with resolution of cell cycle phases (Ostling and Johanson, 1984).

It is also important to note that gating S/G2 cells based on DNA staining intensity, used both here and by Halliwell and colleagues (2020), is only an approximation. Staining for incorporation of the thymidine analogue, EdU, as used in the studies of Simara and Vallabhaneni (2017; 2018) is more accurate. This deviation in methods may account for conflicting results between studies. Where EdU incorporation is used to mark replicating cells,

a short pulse of EdU, will include only early G2 cells in analysis, whereas using DAPI staining intensity will encompass all G2 cells. This is significant as cells in late G2, have more time to repair replication-associated DSBs induced in S-phase and one would anticipate a lower DNA damage burden accordingly.

Whilst comparing S/G2 cells' DNA damage burden gave variable results, I observed lower γH2AX levels in 3 out of 4 lines tested, when comparing all phases of the cell cycle, indicating a difference in DNA damage phenotype between cell states. Ultimately, this study will focus primarily on the genomic distribution of DSBs rather than absolute numbers.

3.3.4. Validating a low-oxygen phenotype in hPSC.

Cultured hPSC's closest developmental equivalent are cells of the peri-implantation epiblast (Nichols and Smith, 2009). Cells of the epiblast *in utero* are exposed to oxygen tensions far lower than atmospheric conditions under which hPSC are routinely cultured *in vitro* (Jauniaux *et al.*, 1999). Accordingly, culturing hPSC under low-oxygen conditions has been shown to improve growth rates (Forristal *et al.*, 2010), increase expression of pluripotency-associated proteins (Närvä *et al.*, 2013) and reduce levels of spontaneous differentiation in routine culture (Fynes *et al.*, 2014). Most interestingly in the context of this study, low oxygen culture has been shown to yield reduced incidences of chromosomal breaks (Forsyth *et al.*, 2006) as well as a significantly reduced mutation rate at the level of both SNVs and INDELS (Thompson *et al.*, 2020). I therefore wanted to investigate how low-oxygen culture affects the distribution and levels of DSBs in hPSC.

Enhanced growth rate is the most commonly reported effect of low oxygen culture in hPSC across studies (Nit *et al.*, 2021). To confirm a "low-oxygen phenotype" ahead of sequencing, I demonstrated enhanced growth rate of hPSC cultured under continuous 5% oxygen. However, upon preparing samples for sequencing, cells cultured under low oxygen exhibited abnormal morphologies and low cell counts, in spite of uniform OCT4+veT-ve expression.

One explanation for this discrepancy in cell behaviour between preliminary experiments and preparation of sequencing material could be cell seeding density. Namely, for sequencing, cells in the low-oxygen condition were seeded at 1×10^4 cells cm^{-2} to account for an anticipated increase in final cell density, compared with a minimum density of 1.1×10^4 cells cm^{-2} in the preliminary experiment. Given this difference in seeding density is minor, a more likely explanation is hypertonic stress in cells cultured for the sequencing experiment. In order to avoid fluctuations in oxygen concentration in the low-oxygen condition, cells were cultured continuously in a hypoxic workstation, the relative humidity of which is 80%, compared to

~100% in a conventional atmospheric O_2 incubator. The volume of media aspirated from the low-oxygen conditions was noticeably lower than the amount dispensed the previous day, suggesting loss via evaporation and concurrent increase in osmolarity of the culture medium. A crucial difference between the preliminary and sequencing experiments in this regard is the choice of culture vessels. For generating sequencing material, cells were seeded into each well of a 6 well plate, whereas in the preliminary experiment, the inner 60 wells of a 96 well plate were seeded, and the outer 36 filled with excess of PBS which may functionally serve as a water bath, increasing relative humidity within the culture vessel.

Ideally, I would have made adjustments to the 5% O_2 condition, for example by introducing a water bath to the workstation and repeating the experiment. However, sequencing, by our collaborators, could only be carried out in a brief window and therefore samples were sent without amending the 5% O_2 condition. Due to the apparent issues with the 5% O_2 condition, I decided to exclude it from subsequent sequencing analysis.

3.3.5. Validating cell-state of INDUCE-seq starting material

Following populating a 96-well plate with samples for sequencing, surplus cells were assayed for expression of the pluripotency-associated marker, OCT4 and mesodermal marker, T. In both pluripotent conditions, all samples showed high proportions of OCT4+ve, T-ve cells, confirming that cells were undifferentiated. In the differentiated condition, all lines except MShef4 were near-pure populations of OCT4-ve T+ve cells. MShef4, by contrast, harboured only ~70% T+ve cells in the differentiated condition. The T-ve population could be due to incomplete differentiation towards a mesodermal fate, or alternatively, T-ve cells could be further specified in MShef4, as T expression is only transient in mesodermal differentiation (Vidricaire *et al.*, 1994).

Importantly, nearly all MShef4 cells of the differentiated condition were OCT4-ve and had therefore likely exited the pluripotent state. For my purposes, the most important factor in differentiation is exit from pluripotency. It is less important for this study to have a well characterised differentiation endpoint, as I have not sought to map DNA damage in specific differentiated cell lineages *per se*, but rather to use the differentiated cell types as a reference point against which to contrast sites of pluripotent DNA damage.

The work in this chapter, provides a robust differentiation model, which recapitulates the previously observed DNA damage in hPSC. This model will serve to compare DSB locations and prevalence between pluripotent and differentiated cell types, in genetically characterised diploid cell lines in the following chapters.

4. Mapping and annotation of genome wide DSBs in hPSC

4.1. Introduction

Endogenous DSBs can derive from diverse mechanisms, including transcription and replication (Hamperl and Cimprich, 2016), as well as the action of endogenous nucleases (Larsen and Sørensen, 2017) and topoisomerase enzymes (Morimoto *et al.*, 2019). Previous work from our laboratory (J. A. Halliwell *et al.*, 2020) and others in the field (Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018) demonstrated an increased number of DSBs in hPSCs compared to their isogenic somatic cells and my own data in Chapter 3 confirmed these observations in several further hPSC lines. However, for the purpose of identifying the mechanistic cause of this damage, all of these studies are limited by the fact that the methods used, including γh2AX immunofluorescence and the neutral comet assay, lack positional information as to where DNA damage occurs at a genomic level (Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018; J. A. Halliwell *et al.*, 2020). Recent advances in DSB-sequencing technologies have enabled construction of genome-wide DSB maps in diverse cell types (Saayman and Esashi, 2022). From DSB context, such maps have identified sites of TOP2-mediated damage at chromatin loop boundaries (Canela *et al.*, 2017; Gothe *et al.*, 2019), sites of transcriptional pausing (Dellino et al., 2019; Singh et al., 2020) and secondary DNA structures (Szlachta *et al.*, 2020); TOP1-mediated damage at enhancer regions (Hazan *et al.*, 2019) and R-loops (Hidmi and Aqeilan, 2022); transcription-replication collisions in long, actively transcribed genes (Wei *et al.*, 2016; Macheret and Halazonetis, 2018; Wang *et al.*, 2020); and replication fork collapse at low-RPA affinity sequences (Tubbs *et al.*, 2018). No such study has previously mapped DSBs in hPSC. I reasoned that knowledge of DSB location and context could inform mechanistic causes, as in the aforementioned studies, and identify how the biogenesis and location of DSBs differs between cells in pluripotent and differentiated states. In the previous chapter, I therefore generated characterised pluripotent and differentiated cell material for INDUCE-seq, to map endogenous DSBs genome wide. In the following I characterise genome-wide DSB maps in pluripotent and differentiated cells.

4.1.1. INDUCE-seq

Following the advent of CRISPR/CAS9 genome editing technology, several sequencing techniques have been developed, to map the genomic location of DSBs in cells (Rybin *et al.*, 2021). Early efforts, based on γh2AX ChIP, suffered from low sensitivity and resolution (Iacovoni *et al.*, 2010; Georgoulis *et al.*, 2017). Subsequent approaches, based on the

sequencing of repair outcomes following “bait DSB” induction, such as HTGTS (Chiarle *et al.*, 2011), suffer positional biases (Wei *et al.*, 2018), and only identify NHEJ products, hence DSBs processed for repair via the HR pathway pass undetected (Bouwman and Crosetto, 2018). A new generation of DSB-sequencing techniques function by directly labelling DSB ends with adapter sequences either in extracted DNA (Baranello *et al.*, 2014) or *in situ* in permeabilised cells (Crosetto *et al.*, 2013). Labelled DNA fragments are then subjected to amplification and sequencing. Such methods allow individual nucleotide resolution of DSBs, moreover, labelling breaks directly in cells minimises the opportunity for introduction of artefactual breaks during DNA during extraction. However, the PCR-based amplification step used in the library preparation of these methods is prone to introducing bias based on sequence preference, moreover the stochastic nature of PCR amplification means that low-abundance fragments can be lost (Kebschull and Zador, 2015). The result is that the sequencing reads obtained do not reflect the true proportions of DSBs in the starting population of cells.

INDUCE-seq is an *in situ* break labelling method which lacks PCR-based amplification of DNA fragments during library preparation (Dobbs *et al.*, 2022). Briefly, adherent cells are fixed in a 96 well plate and permeabilised. Double-stranded ends of DNA are blunted using a T4 DNA ligase, followed by ligation of functional, full-length P5 sequencing adapters. DNA is next extracted, fragmented by sonication and half-functional P7 adapters are ligated to DNA ends. Fragments are directly hybridized onto an Illumina flow cell, as a means of enrichment. At this stage, only fragments containing the P5 adapter sequence (i.e., DSB ends in the starting population of cells) hybridise to the flow cell. Fragments lacking the P5-sequancing adapter have half-functional P7 adapters at either end, rendering them unable to hybridise and thus are washed off. Bridge amplification generates clonal clusters on the flow cell, which are then sequenced by synthesis. Reads are aligned to a reference genome, with the 5’ position of the sequencing read pertaining to the position of the DSB end (Figure 4.1) (Dobbs *et al.*, 2022). The lack of PCR amplification during library preparation means each read from the flow cell corresponds to an individual DSB end in the starting population of cells, yielding superior signal to noise compared with alternative direct break labelling methods, capturing both rare and abundant DSBs (Dobbs *et al.*, 2022).

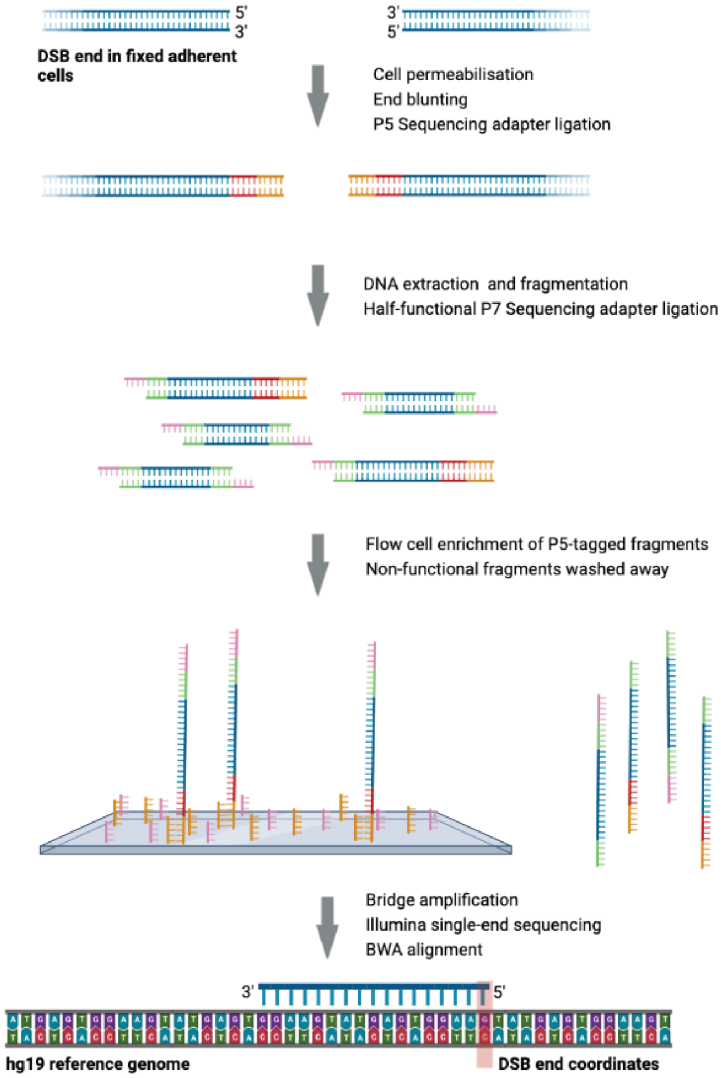


Figure 4.1 INDUCE-seq schematic

DSB ends in fixed cells are directly tagged with P5 sequencing adapters. DNA is then extracted, fragmented and half-functional P7 sequencing adapters ligated. Fragments are hybridized to the flow cell without PCR-based amplification. Only fragments with P5 sequencing adapters can hybridize. Non-functional fragments, labelled post-extraction, are washed away. Each sequencing read pertains to an individual DSB end in the starting population of cells. Mapped DSB coordinate is highlighted in red (bottom). Figure based on (Dobbs *et al.*, 2022).

4.1.2. Annotation of DSB maps

Previous DSB-mapping studies have used annotation of genomic features as well as additional sequencing datasets to inform models for endogenous DSB formation. Chromatin structure is dynamic and modified to permit or restrict transactions such as DNA replication and transcription. This is partly facilitated by post translational modification of histone proteins. The presence of certain histone modifications, as determined by ChIP-seq, can be used to infer chromatin state. For example H3K4me3 is typically associated with active promoter regions (Howe *et al.*, 2017), H3K27ac is associated with active promoters and enhancers

(spicuglia and Vanhille, 2012), H3K36me3 with transcriptional elongation and H3K27me3 with transcriptional repression (Kimura, 2013). Complementary sequencing methods can be used to map physical accessibility of chromatin. The assay for transposase accessible chromatin (ATAC-seq) utilises a Tn5 transposase enzyme, which cleaves DNA exclusively in histone-depleted regions, thereby mapping open chromatin, often associated with active promoter and enhancer regions (Luo *et al.*, 2022). DSB-mapping studies have noted an enrichment of DSBs in open, active chromatin (Dellino *et al.*, 2019; Hidmi and Aqeilan, 2022), with two recent computational studies identifying open chromatin as one of the single best predictors for DSB formation (Mourad *et al.*, 2018; Ballinger *et al.*, 2019). Similarly, annotating DSB maps with transcriptional datasets in the form of RNA-seq or GRO-seq has demonstrated a positive correlation between gene's transcriptional activity and DSB density in multiple cell types (Baranello *et al.*, 2014; Yan *et al.*, 2017; Dellino *et al.*, 2019; Gothe *et al.*, 2019; Sandeep Singh *et al.*, 2020; Ballarino *et al.*, 2022).

I hypothesize that DSBs in hPSC are distributed in a non-random fashion, and that their context can inform their initial cause. In this chapter, I aim to annotate INDUCE-seq data to identify the context of genome-wide DSBs in hPSC, relative to:

i) Genomic features,
ii) Epigenomic features, and
iii) Transcriptional activity,

with the ultimate goal of inferring putative causes of DSBs.

4.2. Results

4.2.1. Pluripotent cells harbour greater numbers of genome-wide DSBs than their differentiated counterparts

All samples prepared in chapter 3 were subjected to INDUCE-seq library-prep, sequencing and alignment to the hg19 reference genome by Broken String Biosciences. As a first-pass coarse means of comparing break prevalence and distribution between pluripotent and differentiated cell types, I plotted break coverage for all canonical chromosomes in an example H9 pluripotent and differentiated sample (Figure 4.2). DSBs appear broadly distributed throughout the genome, excluding blacklisted regions which are not mapped in hg19. Notably, the amplitude of break coverage in the pluripotent sample (red, above) appears greater than that in differentiated sample (blue, below), consistent with the quantitative γh2AX immunofluorescence in chapter 3.


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

Figure 4.2 H9 break coverage

Example ideogram of INDUCE-seq read coverage across all canonical chromosomes. Data from example H9 replicate. Pluripotent coverage in red above chromosomes, differentiated coverage in blue below chromosomes. Read counts normalised to library size (ng DNA).

As INDUCE-seq's library preparation lacks PCR-based amplification of DNA fragments prior to sequencing, each read sequenced corresponds to an individual DSB end in the starting population of fixed cells (Dobbs *et al.*, 2022). The technique is therefore quantitative and, following normalization of the number of sequencing reads obtained against the total mass of DNA sequenced for each sample, allows direct comparison of genome wide DSB numbers between samples (Figure 4.3 A, B). Three of five cell lines analysed (H7, MShef4 and MIFF-1) had significantly higher numbers of DSBs per ng of DNA sequenced in pluripotent than differentiated samples (Figure 4.3 A). Most notably, the mean normalised break numbers obtained in the human foreskin fibroblast (hFF) line CRL2429 was 2.6×10^2 constituting a ~90-fold reduction in DSBs compared with its derivative reprogrammed iPSC counterpart MIFF-1 (mean 2.4×10^4 DSBs/ng) (Figure 4.3 B, A).

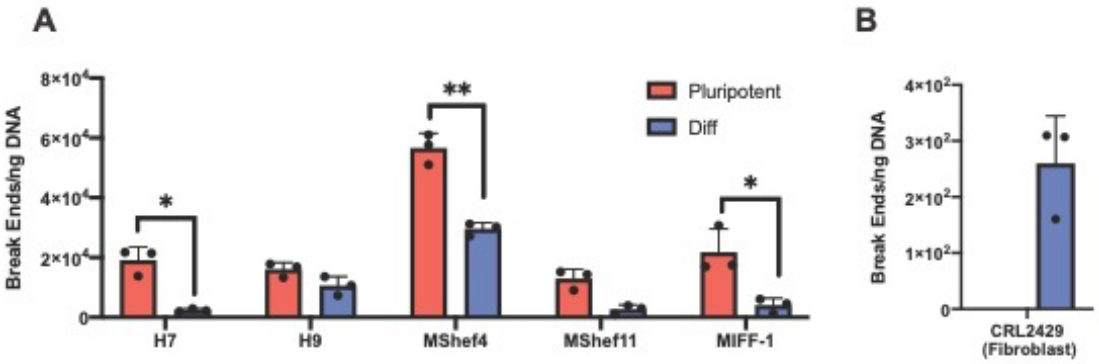


Figure 4.3 Normalised DSB counts in pluripotent and differentiated samples.

*DSB ends sequenced, normalised to sequencing library masses. Bars represent mean values of 3 biological replicates, points denote individual values in A) pluripotent lines and Differentiated derivatives and B) hFF line. NB scales differ between A and B (paired t-test, n= 3 biological replicates *p<0.05, **p<0.01)*

These data are consistent with quantitative assays carried out in chapter 3 as well as previously published work (Ruiz *et al.*, 2015; Vallabhaneni *et al.*, 2018; Halliwell *et al.*, 2020) and serve as a useful quality control step, corroborating that cell types exhibit the expected patterns in terms of relative break frequencies.

4.2.2. DSBs are enriched in open, active chromatin in pluripotent cells

Next, I sought to annotate genetic, epigenetic and transcriptomic features to INDUCE-seq datasets, to determine the context of DSBs and ultimately infer putative causes. Ahead of carrying out annotation studies, I carried out further quality control steps to verify that break datasets were reproducible between replicates of a given sample. To this end, I split the hg19 reference genome into 100kb sequential bins and counted DSBs in each bin. Pearson's correlation coefficients were calculated for each replicate of each cell line using these binned data. Pearson's correlation coefficients were consistently > 0.8 for replicates of a given cell line and condition, indicating good reproducibility (Figure 4.4). Reads from replicate samples were therefore pooled ahead of subsequent annotation studies.

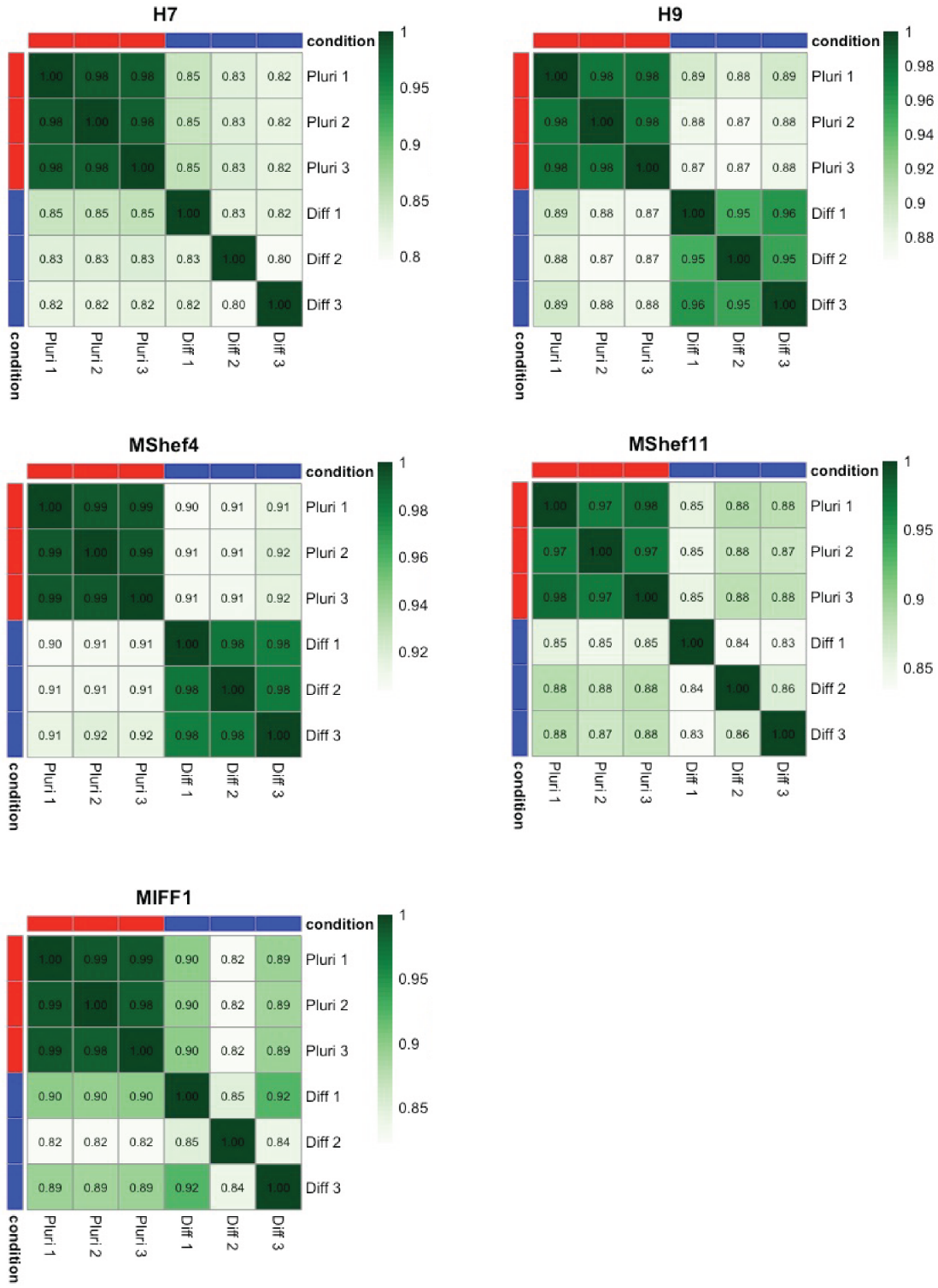


Figure 4.4 INDUCE-seq QC: Replicate heatmaps

Heatmaps of Pearson's correlation coefficients for DSB counts in 100kb bins. Pluripotent condition red, differentiated blue. Pearson's coefficient values in cells, scales indicated to right.

To determine whether DSBs are significantly enriched in a genomic region of interest, it is necessary to calculate the observed (i.e., real) overlap, as well as the expected (randomly

distributed data) overlap. To this end, genomic coordinates of DSBs (of which each is length 1bp) can be shuffled at random (excluding a mask of regions which are not mapped in hg19) and the number of times the randomised dataset overlaps a given region of interest can be calculated. By dividing the observed (real) overlaps by expected (random) overlaps for a region of interest, we can measure enrichment, whereby values >1 indicate enrichment of DSBs at a feature, and values <1 represent depletion of DSBs (Figure 4.5). Randomly shuffling datasets many times, allows estimation of a random distribution of overlaps, and calculation of P-values, to infer statistical significance, when comparing real and randomised datasets. Such analysis is commonly known as a randomization or permutation test (Edgington, 1980).

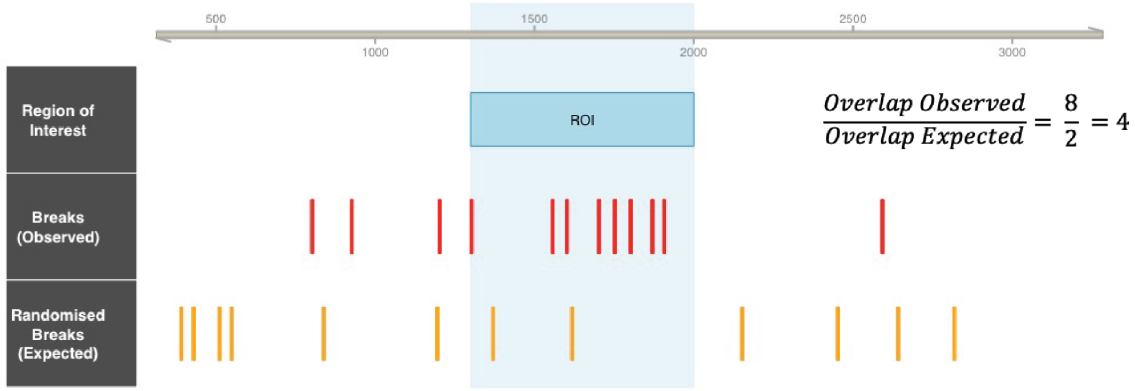


Figure 4.5 Schematic of enrichment calculation

Schematic of break enrichment at genomic features. Individual DSBs for a given sample (red) are intersected with a region of interest (blue) to obtain the observed overlap frequency. Break coordinates are shuffled at random within chromosomes (orange) and intersected with the region of interest to obtain the expected overlap frequency. Observed/Expected values can be used to measure enrichment over random of DSBs at a given feature.

In permutation tests, the estimated P-value is calculated by:

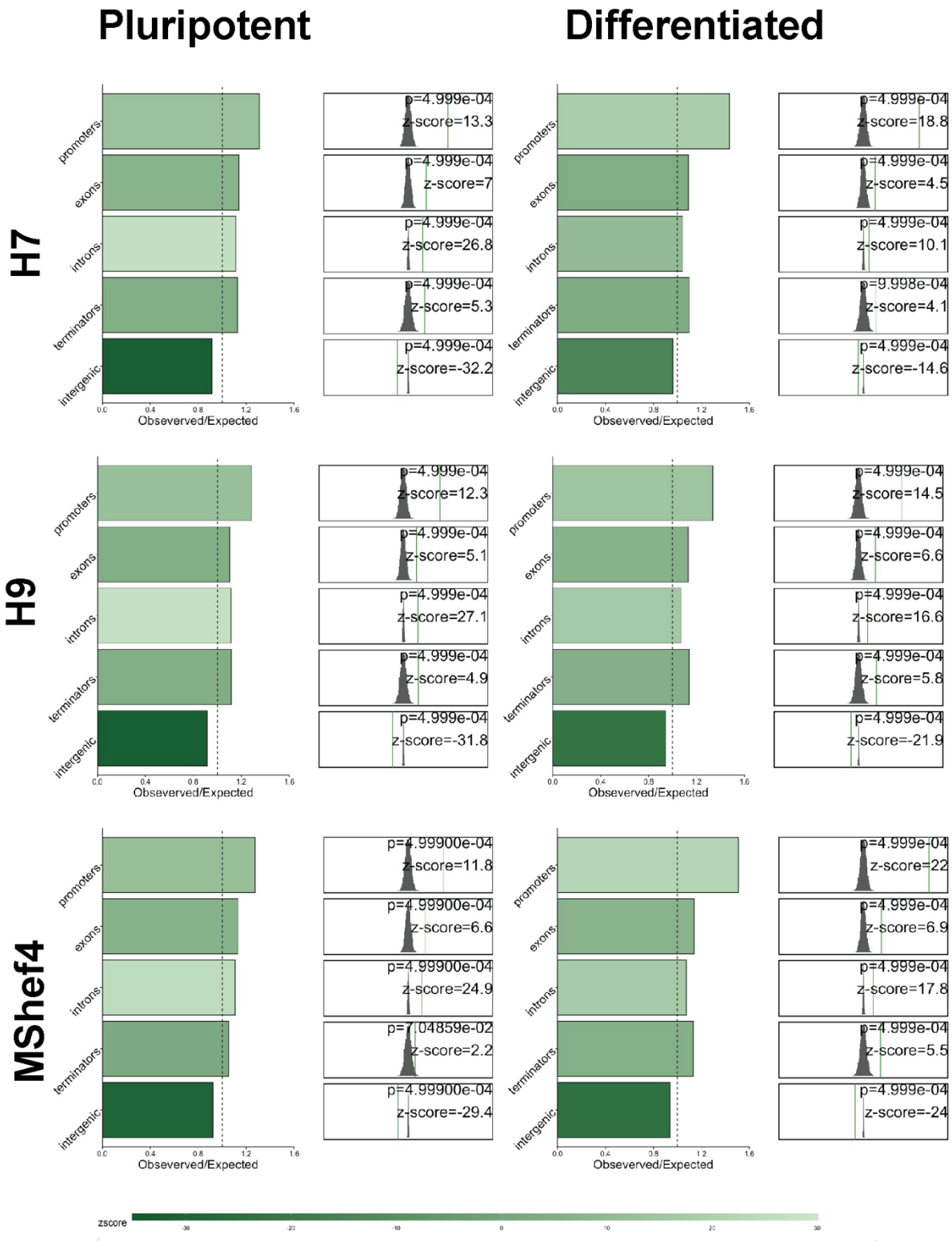
$$\frac{1 + N\ Permuted\ frequencies\ more\ extreme\ than\ observed\ frequencies}{N\ Permutaitons}$$

The minimum size of the estimated P-value is therefore limited by the number of permutations carried out (Hutson and Yu, 2021). Thus, I chose to carry out relatively a high number of permutations (10^4) to increase P-value accuracy. As a typical pooled pluripotent sample contains ~3 $\times 10^7$ DSB ends, permutation of all reads required more computing resources than were available to me. I therefore randomly sampled 10^5 break ends from each sample and, given all break ends are the same length (1bp) across all samples, permuted 1 dataset only to calculate random distributions for each feature. To determine enrichment of DSBs in samples at genomic features, I obtained RefSeq promoter, exon, intron and terminator

coordinates from UCSC. I categorised any genomic region not overlapping the aforementioned features as intergenic and carried out permutation tests for each cell line in pluripotent and differentiated conditions (Figure 4.6). All genic features, most notably promoter regions, were significantly enriched for DSBs in all cell lines in both pluripotent and differentiated conditions, consistent with DSB mapping studies in other cell types (Dellino *et al.*, 2019; Hazan *et al.*, 2019; Wang *et al.*, 2020). Intergenic regions were universally significantly depleted of DSBs across samples, raising the possibility of transcription-associated DNA-damage.

Quantifying break enrichment across all annotated genomic features is a crude means of analysis, particularly considering gene expression and regulation is cell-type specific (Cramer, 2019). A more precise analysis would entail defining the epigenetic state of specific genomic features to determine active and inactive regions. I sought to determine whether DSBs were preferentially enriched in specific chromatin states. To this end, I used a published sequencing dataset of H9 cells cultured under pluripotent conditions (Yan et al., 2020). The sequencing dataset comprises ChIP-seq for histone modifications H3K27ac, H3K27me3, H3K36me3 and H3K4me3, ATAC-seq to map open chromatin, and DRIP-seq, which uses the S9.6 antibody to immunoprecipitate R-loops (Ginno *et al.*, 2012).

Whilst annotating DSBs on individual histone modifications can provide some information as to the epigenetic context of DSBs, it is more informative to combine available datasets of epigenetic features with genomic features to define functional “chromatin states”. I used epigenetic sequencing reads from the Yan dataset (Yan et al., 2020) to define chromatin states. For this I used Spectacle software (Song and Chen, 2015), a modified version of the commonly used ChromHMM software (Ernst and Kellis, 2010), which employs spectral learning to reduce run-time (Figure 4.7 A). Annotating chromatin states with RefSeq genomic features allowed manual naming of chromatin states based on the presence of genomic and epigenetic features (Figure 4.7 B-D). I calculated enrichment of DSBs for the H9 pluripotent dataset, as in Figure 4.6. DSBs appeared enriched in all chromatin states with open chromatin (marked by ATAC-seq signal), but this was most pronounced in chromatin states also high in marks associated with active transcription H3K27ac and H3K4me3, within 2kb of a TSS and lacking repressive chromatin modification H3K27me3, demonstrating a strong association between break density and open, actively transcribed chromatin (Figure 4.7 B-F).


Pluripotent
Differentiated
H7
H9
MShef4
promoters
exons
introns
terminators
intergenic
Obseverved/Expected
p=4.999e-04
z-score=13.3
z-score=7
z-score=26.8
z-score=5.3
z-score=-32.2
z-score=18.8
z-score=4.5
z-score=10.1
p=9.998e-04
z-score=4.1
z-score=-14.6
z-score=12.3
z-score=5.1
z-score=27.1
z-score=4.9
z-score=-31.8
z-score=14.5
z-score=6.6
z-score=16.6
z-score=5.8
z-score=-21.9
p=4.99900e-04
z-score=11.8
z-score=6.6
z-score=24.9
p=7.04859e-02
z-score=2.2
z-score=-29.4
z-score=22
z-score=6.9
z-score=17.8
z-score=5.5
z-score=-24
zscore

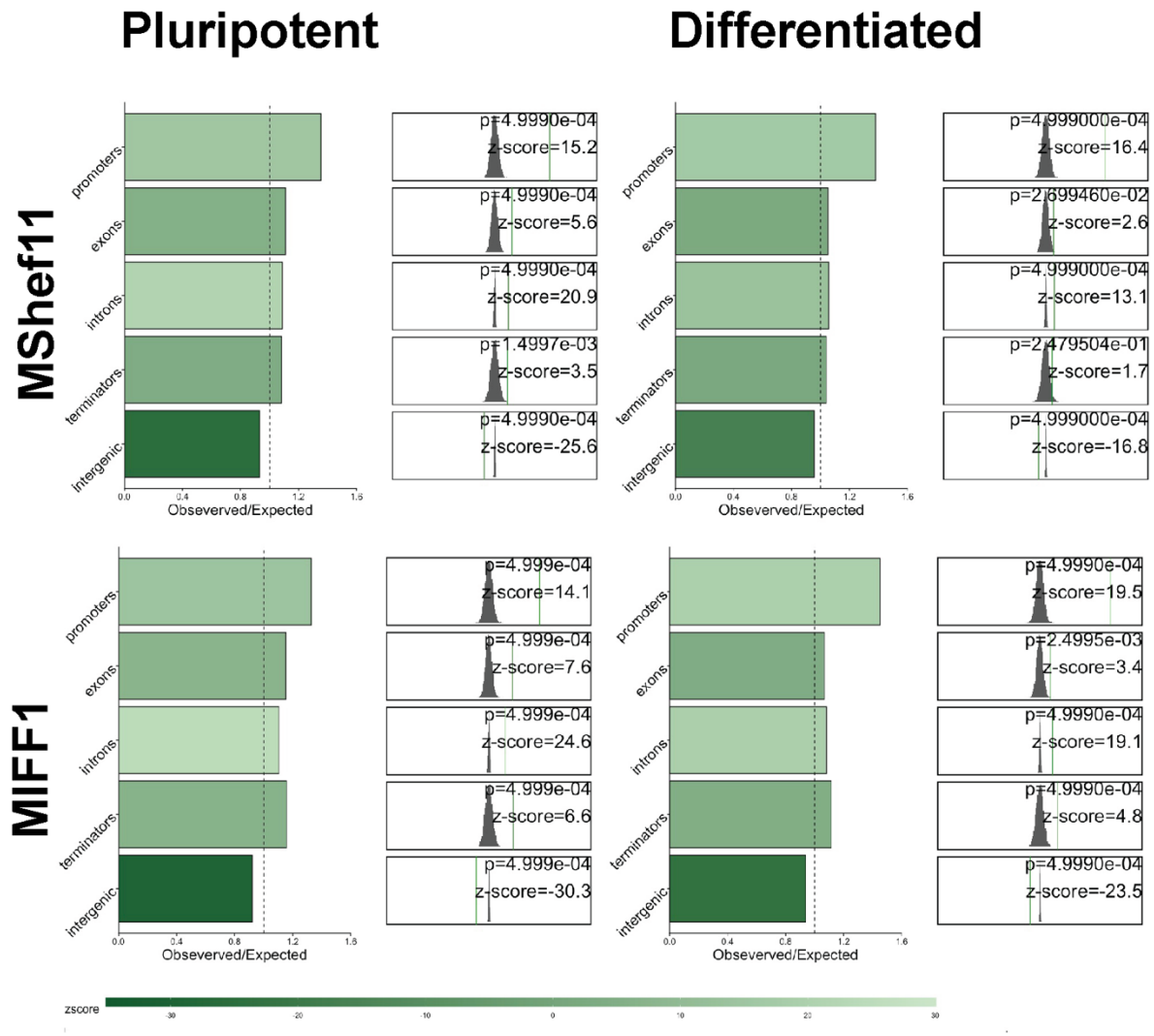


Figure 4.6 Enrichment of DSBs in Genomic Features

10^5 DSBs were randomly sampled from each condition. Downsampled DSBs were randomly permuted on a per-chromosome basis 10001 times. Barplots (left) of Observed/Expected frequencies of intersection between DSBs and genomic features, colored by z-score (z-score=observed events-mean permuted events/SD). Dashed line x-intercept at 1, the threshold for enrichment/depletion of DSBs. Expected frequencies taken as the mean intersection of 10^4 randomised datasets. Histograms (right) show distribution of permuted break overlaps in grey, versus the observed number of break overlaps (green line). Scales are +/-60% of mean permuted values. Adjusted P-values and z-scores overlaid. Minimum adjusted P value=4.999e-04.

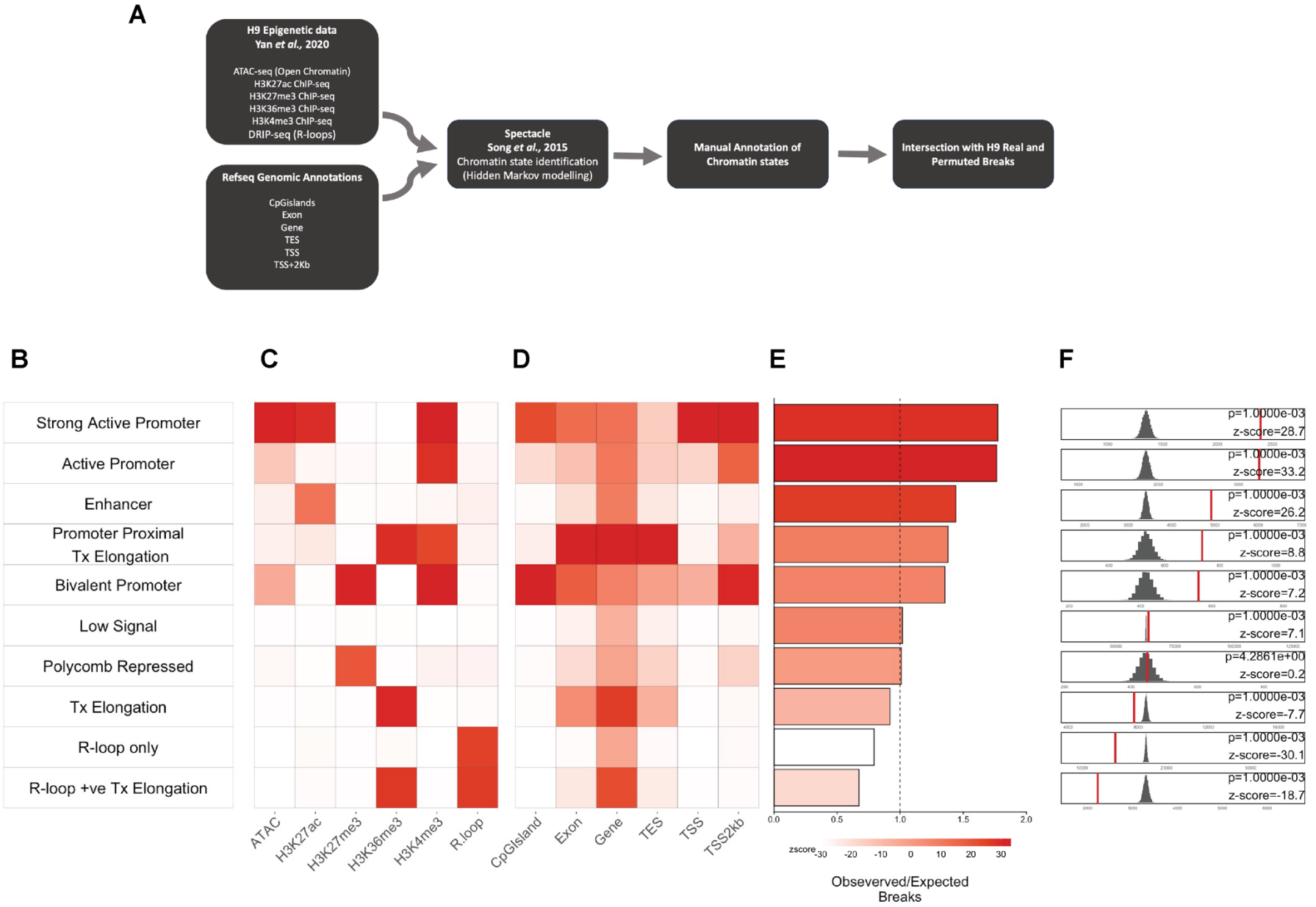

A
H9 Epigenetic data
Yan et al., 2020
ATAC-seq (Open Chromatin)
H3K27ac ChIP-seq
H3K27me3 ChIP-seq
H3K36me3 ChIP-seq
H3K4me3 ChIP-seq
DRIP-seq (R-loops)
Refseq Genomic Annotations
CpGislands
Exon
Gene
TES
TSS
TSS+2Kb
Spectacle
Song et al., 2015
Chromatin state identification
(Hidden Markov modelling)
Manual Annotation of Chromatin states
Intersection with H9 Real and Permuted Breaks
B
Strong Active Promoter
Active Promoter
Enhancer
Promoter Proximal Tx Elongation
Bivalent Promoter
Low Signal
Polycomb Repressed
Tx Elongation
R-loop only
R-loop +ve Tx Elongation
C
ATAC
H3K27ac
H3K27me3
H3K36me3
H3K4me3
R.loop
D
CpGIsland
Exon
Gene
TES
TSS
TSS2kb
E
0.0
0.5
1.0
1.5
2.0
zscore
-30
-20
-10
0
10
20
30
Obseverved/Expected Breaks
F
p=1.0000e-03
z-score=28.7
p=1.0000e-03
z-score=33.2
p=1.0000e-03
z-score=26.2
p=1.0000e-03
z-score=8.8
p=1.0000e-03
z-score=7.2
p=1.0000e-03
z-score=7.1
p=4.2861e+00
z-score=0.2
p=1.0000e-03
z-score=-7.7
p=1.0000e-03
z-score=-30.1
p=1.0000e-03
z-score=-18.7

Figure 4.7 DSBs are Enriched in open and active chromatin in H9 cells

A) Schematic for defining and determining enrichment in chromatin states. Epigenetic sequencing data, in H9 cells from Yan et al., 2020, was used to define chromatin states, using Spectacle software. States were manually annotated (B) based on chromatin marks and overlap with genomic features (C,D), respectively. The intersections of DSBs with each of these states, alongside with 10001 randomly permuted datasets, were calculated to determine enrichment over random. B) Annotations of each chromatin state C) Heatmap of chromatin marks in each state. Darker color indicates higher probability of observing chromatin mark in state. D) Heatmap of overlap between genomic features and chromatin states, individual color scales applied to each genomic feature for visualization. E) Observed/ Expected frequencies of DSBs in each chromatin state, Bars colored by z-score. F) Distributions of permuted samples overlap with chromatin states (grey histograms), observed overlap values indicated by red lines. Adjusted P-values and z-scores indicated on right of plot (n=10001 permutations). Minimum adjusted P value=1.000e-03.

4.2.3. R-loops are specifically depleted of DSBs in H9 hESCs

R-loops have been implicated as a driver of genome damage in diverse cell types via numerous mechanisms, including deamination of the ssDNA component (Basu *et al.*, 2011), flap-endonuclease attack of flap structures flanking the hybrid (Sollier *et al.*, 2014; Jurga *et al.*, 2021), and interference with DNA replication forks (Tuduri *et al.*, 2009; Helmrich *et al.*, 2011; Kotsantis *et al.*, 2016; Stork *et al.*, 2016; Hamperl *et al.*, 2017; Zhang *et al.*, 2022). Surprisingly, the two R-loop-high chromatin states identified by Spectacle software, were significantly depleted for DSBs, compared with permuted datasets (Figure 4.7 B-F), contradictory to a recent study in MCF7 cells (Hidmi and Aqeilan, 2022). Of the three reported mechanisms of R-loop induced damage in the literature, only deamination of the ssDNA component would yield DNA breaks overlapping the R-loop. Attack of the flap regions by endonucleases, or indeed interference with oncoming replication forks, should cause DNA breakage adjacent to the R-loop structure. To determine whether R-loop adjacent regions in the pluripotent genome are enriched for DSBs, I used a consensus H9 pluripotent R-loop peak set from (Yan et al., 2020), and analysed enrichment of DSBs over the R-loop peak itself, as well as peak-flanking regions of increasing size (Figure 4.8 A). Given the abundance of R-loop peaks in the identified set ($n=3.6\times10^5$), I limited the size of the maximum flanking region to 1kb, to minimise overlap with neighbouring R-loops and to avoid generating a file which encompasses the majority of the genome (assuming even distribution of R-loop peaks, 10kb flanking regions would cover hg19 reference genome ~2× (Church *et al.*, 2011)) (Figure 4.8 A). Consistent with the Chromatin state analysis, annotating DRIP-seq peaks reveals mapped R-loops are significantly depleted for DSBs, as are their 10bp and 100bp flanking regions (Figure 4.8 B). 1000bp flanking regions showed a very modest enrichment for DSBs ($p=0.0308$), which, given their breadth of coverage genome-wide, may simply reflect being a large region of R-loop-negative genome. Together with the chromatin state analysis, these data demonstrate that mapped R-loops in H9 hESC cells are specifically depleted of DSBs, and thus these R-loops likely do not drive DSB formation in H9 cells.

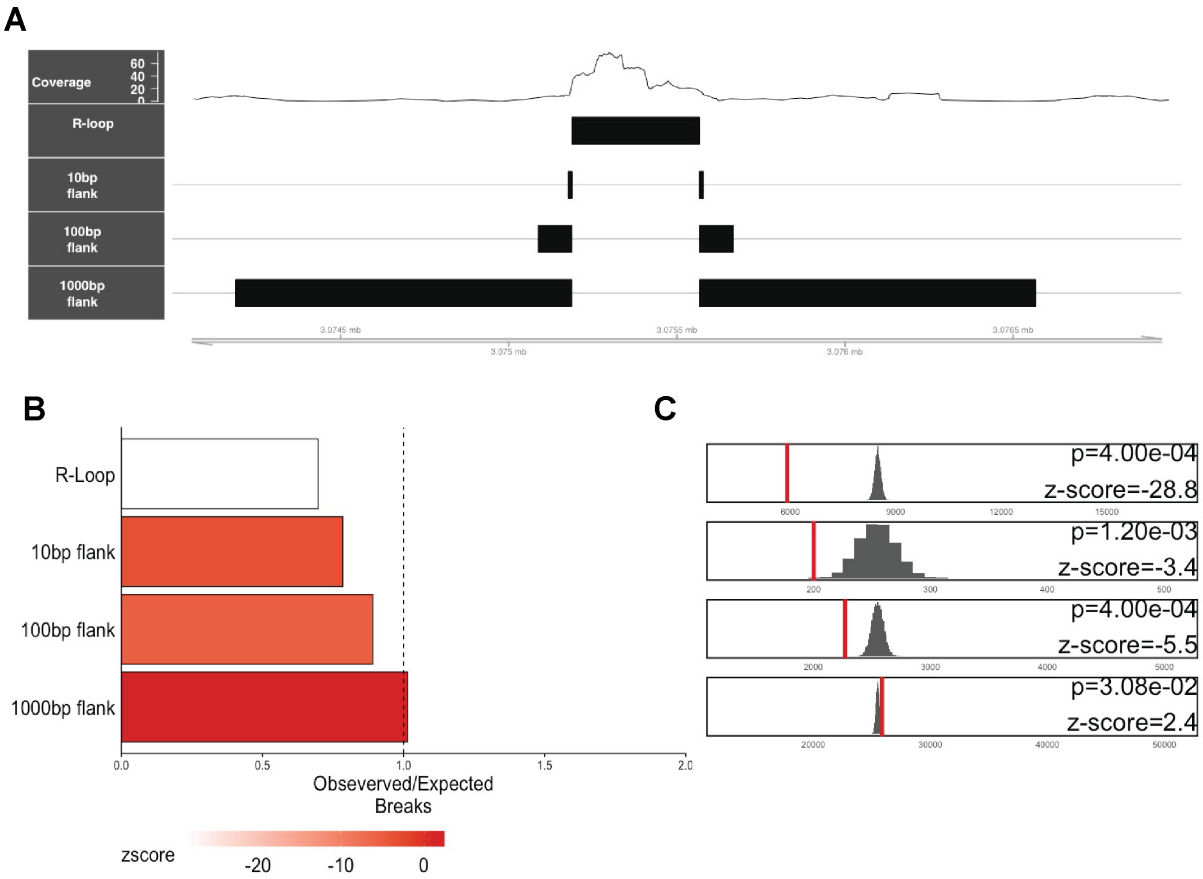


Figure 4.8 R-loops are break-poor regions in H9 hPSC

A) *Example genomic interval, showing H9 pluripotent DRIP-seq coverage (top), R-loop peak below, and flanking regions of increasing size, used for intersection with H9 pluripotent break data and permuted datasets. B) Observed/Expected frequencies of DSBs in R-loops and flanking regions, bars colored by z-score. C) Distributions of permuted samples overlap with chromatin states (grey histograms), observed overlap values indicated by red lines. Adjusted P-values and z-scores indicated on right of plot (n=10001 permutations). Minimum adjusted P value=4.000e-04.*

4.2.4. DSBs are associated with transcription in H9 hESC

The chromatin states with the greatest enrichment of DSBs in H9 were transcriptionally active open chromatin, most commonly promoter proximal. Several groups have identified transcription as a driver of endogenous DNA damage, often in a replication-dependent manner (Helmrich *et al.*, 2011; Kotsantis *et al.*, 2016). Indeed DSB mapping studies have reported a high break density in highly expressed genes (Wei *et al.*, 2016; Dellino *et al.*, 2019; Gothe *et al.*, 2019; Wang *et al.*, 2020; Ballarino *et al.*, 2022). To determine whether break density is associated with gene expression levels in hPSC, I took a published RNA-seq dataset in H9 cells (Yan et al., 2020), and used transcripts per million (TPM) values to analyse only expressed genes (mean TPM>1). I categorised all expressed genes into quartiles based on expression levels, Q1 being the lowest and Q4 the highest (Figure 4.9 A), and quantified DSB density throughout gene bodies, defined as Ensembl transcript start to end coordinates. This

analysis revealed that break density along gene bodies increases with expression level, with genes of the highest expression quartile harbouring a significantly higher break density than genes of all other quartiles (Figure 4.9 B).

Transcription is commonly reported to induce DNA damage via collisions with oncoming DNA replication forks, causing stalling and cleavage of the fork (Hamperl and Cimprich, 2016). DNA replication and transcription are spatiotemporally separated in both prokaryotes and eukaryotes, minimising the chances of interference between the two processes (Meryet-Figuiere *et al.*, 2014). However, longer genes take longer to transcribe and therefore the likelihood of transcription interfering with DNA replication is higher in longer genes, even inevitable in genes exceeding 800kb which typically take more than one iteration of the cell cycle to transcribe (Helmrich *et al.*, 2011). Consistent with this observation, studies in neural progenitor cell types reported increased breakage in longer genes (Wei *et al.*, 2016, 2018; Wang *et al.*, 2020). To determine whether gene length is associated with DSB density in hPSC, I plotted H9 gene body DSB density against gene length and noted a significant correlation between the two parameters, i.e. longer genes have higher break densities (Figure 4.9 C). Together with the increased gene body DSB density in highly expressed genes, correlation of gene length with DSB density may be indicative of DNA damage resulting from transcription-replication collisions.

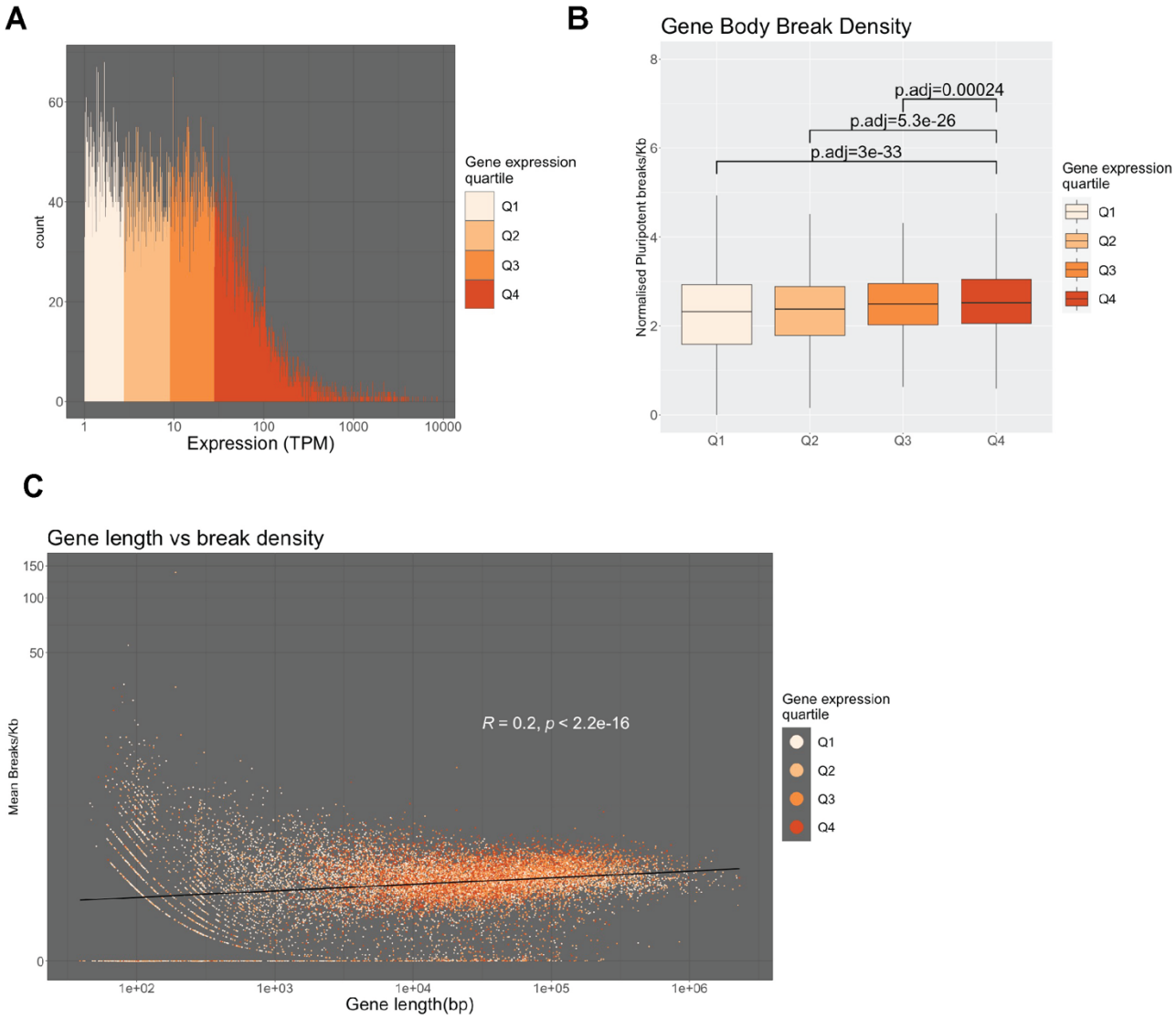


Figure 4.9 DSBs are enriched in bodies of long actively transcribed genes

A) Histogram of gene expression levels, coloured by expression quartile. B) H9 Break density across expressed gene bodies split by expression quartile (Q1 lowest, Q4 highest). Boxes represent interquartile range, mid-line represents median value. outliers not shown. (Wilcoxon test, n=5428 genes (Q1) or n=5249 genes (Q2:Q4)). C) Break density vs length of expressed genes (Pearson correlation n=20995 genes).

A body of work implicates the action of topoisomerase enzymes in generating DSBs at sites of transcriptional initiation at active enhancers (Hazan *et al.*, 2019), and promoters (Dellino *et al.*, 2019; Sandeep Singh *et al.*, 2020) to alleviate supercoiling of DNA. Chromatin state analysis revealed the greatest enrichment of DSBs was in states frequently overlapping, or proximal to the gene TSS (Figure 4.7 B-F), moreover promoters showed the highest break density of any genomic feature (Figure 4.6). I next sought to determine whether promoter break density was proportional to gene expression level in hPSC. To this end, I defined promoter regions as +1kb/-2kb of the Ensembl transcript start (Ray *et al.*, 2022). I then quantified DSB densities in promoters of genes in each expression quartile, and, consistent

with reports in other cell types (Baranello *et al.*, 2014; Yan *et al.*, 2017; Dellino *et al.*, 2019; Gothe *et al.*, 2019; Sandeep Singh *et al.*, 2020; Ballarino *et al.*, 2022), promoter DSB density increased with gene expression level (Figure 4.10 A).

To more closely inspect how DSBs are distributed around TSS in H9 cells, I next generated metagene plots of mean DSB coverage in the 5kb region either side of genes' TSS (Figure 4.10 B). Mean break density increases from 5kb upstream up to the TSS in genes of each quartile. Immediately following the TSS however, is a marked drop in break coverage for ~500bp, following which a second peak of DSBs, with a steady decay is visible in all gene quartiles (Figure 4.10 B). This marked increase in DSB density downstream of the TSS could stem from transcription-replication conflicts, akin to gene body DSBs, however the distribution of DSBs, most notably upstream of the TSS is strikingly similar to TOP2 induced DSBs reported elsewhere (Dellino et al., 2019; Singh et al., 2020).

The unexpected, conspicuous dip in DSB coverage around the TSS is scarcely reported in published DSB-mapping literature. One study mapping endogenous DSBs in HeLa cells noted a dip in break coverage in the ~200bp immediately surrounding the TSS, which they attribute to RNAPII occupancy of chromatin (Singh et al., 2020). Other studies have reported that decreased break coverage in the immediate vicinity of the TSS is a feature of CpG-containing promoters (Ballarino *et al.*, 2022), and postulate the dip in break coverage may be due to altered minor groove width of CpG-rich DNA exhibiting an increased resistance to spontaneous breakage (Yang *et al.*, 2015). Neither of these hypotheses are substantiated with data and as such this phenotype warrants further investigation.

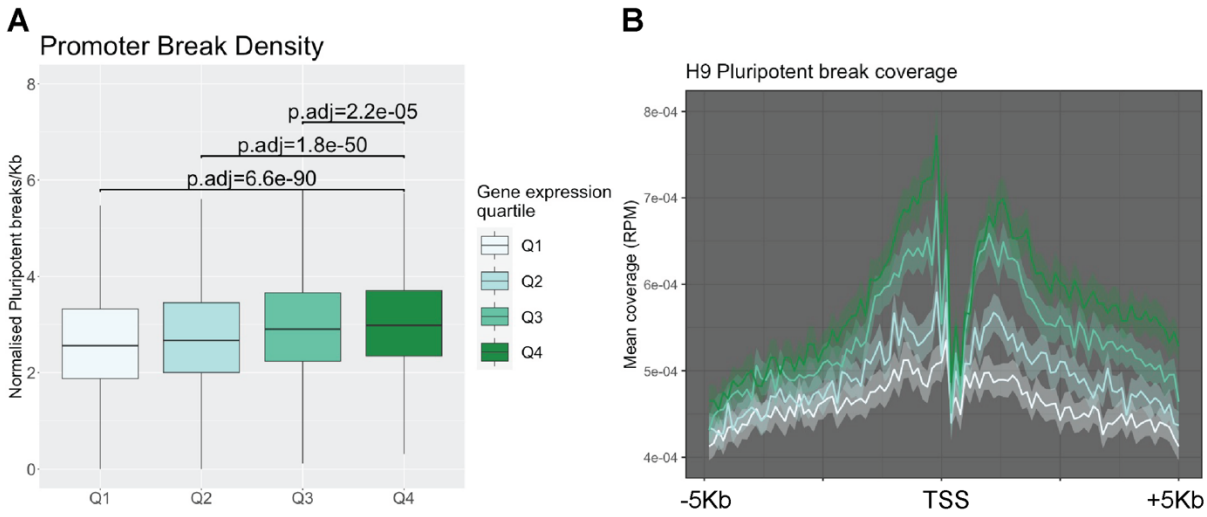


Figure 4.10 DSBs peak before TSS of highly expressed promoters

A) H9 Break density in expressed gene promoters (TSS+1kb/-2Kb) split by expression quartile (Q1 lowest, Q4 highest). Boxes represent interquartile range, mid-line represents median value. Outliers not shown.(Wilcoxon test, n=5428 promoters (Q1) or n=5249 promoters (Q2:Q4)). B) Metagene profile of mean H9 Break coverage (RPM) around the gene's TSS. Lines coloured by gene expression quartile as in (A).

4.2.5. Unbiased clustering of promoter regions reveals 4 distinct DSB patterns in hPSC.

The dip in DSB coverage observed near the TSS in H9 cells is poorly documented in other cell types, and, in instances where it *is* reported, no mechanistic basis for this reduction in damage has been established (Ballarino et al., 2022; Singh, Shih, et al., 2020; Yang et al., 2015). I therefore sought to investigate this phenomenon further. As other groups propose that a protected TSS is a feature of CpG island-containing promoters (Ballarino et al., 2022; Yang et al., 2015), I categorised genes based on presence or absence of a CpG island in the TSS +/- 500bp, and plotted DSB coverage around the TSS for genes of each gene category (Figure 4.11). CpG+ve promoters appear to have increased coverage throughout the 10kb plotted region, when compared with CpG-ve promoters. CpG-ve promoters have a single apparent peak downstream of the TSS, whereas CpG+ve promoter break coverage peaks once immediately upstream of TSS and again ~500bp downstream of the TSS, largely comparable to the pattern reported by Yang and colleagues (Yang *et al.*, 2015).

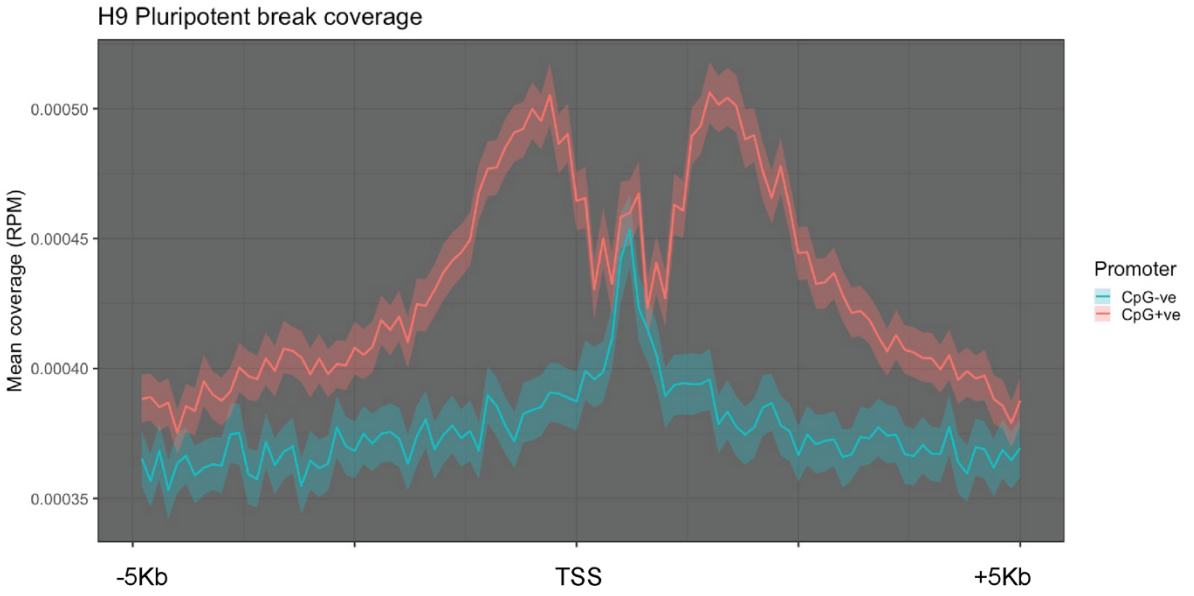


Figure 4.11 DSB coverage in CPG+ve vs CPG-ve promoter regions in H9 hESCs

Mean H9 pluripotent INDUCE-seq coverage for CpG +ve (red n=13489) and CpG-ve (blue n=8226) promoter regions.

Based on these data, bimodal DSB coverage around the TSS with a relatively protected TSS, would appear to be a feature of CpG+ve promoters. However, this method of analysis has two major limitations: i) It makes assumptions as to which genes harbour a protected TSS and ii) metagene plots are composite coverage plots representing mean coverage of thousands of promoter regions, and as such do not detect heterogeneity in DSB coverage amongst

individual genes of a plotted group. I therefore selected an alternative approach to investigate this phenotype, whereby promoter regions are first categorised based on their break distribution, then annotated to determine what drives protection of TSS in a sub-set of genes. To this end I used deepTools software (Ramírez *et al.*, 2016) to plot heatmaps of the promoter regions of all expressed genes in H9 pluripotent cells and carry out unbiased clustering of promoter regions based on the distribution of DSBs (Figure 4.12 A). Clustering of promoters and potting of heatmaps revealed 4 distinct clusters in H9 pluripotent samples:

Cluster 1: Low-DSB TSS, high-DSB flanking

Cluster 2: High-DSB TSS, low DSB flanking

Cluster 3: Low-DSB TSS, low-DSB flanking

Cluster 4: Very low DSB

To determine whether this damage patterning is common to pluripotent lines or unique to H9 cells, I plotted break coverage across the identified promoter clusters in the remaining 4 hPSC lines, and each appeared broadly comparable indicating four distinct promoter damage patterns are detectable in hPSC and are reproducible across different cell lines (Figure 4.12 B).

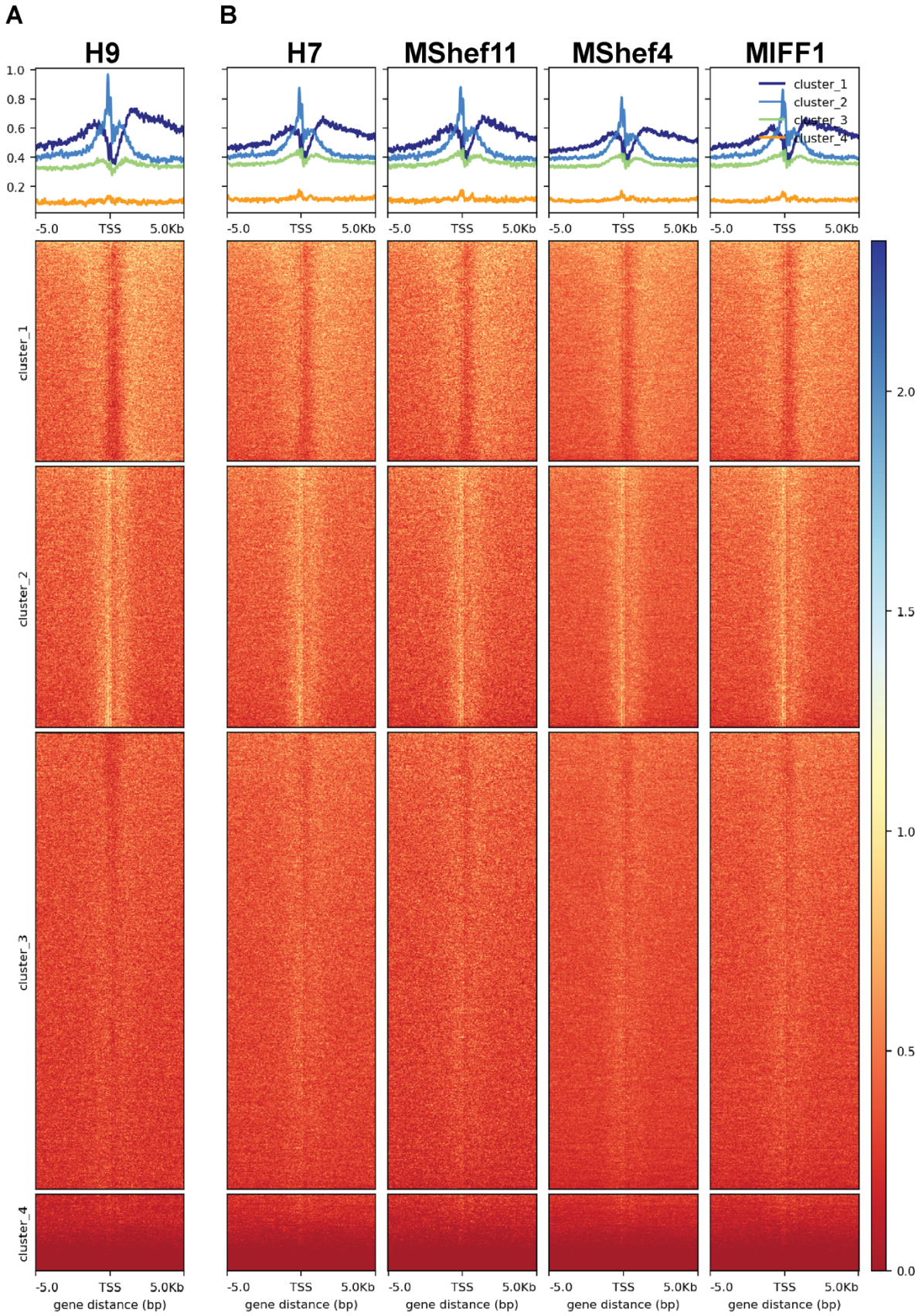


Figure 4.12 Unbiased clustering of pluripotent promoters

Metagene profiles (upper) and heatmaps (lower) of break coverage in pluripotent samples +/-5kb of TSS. TSS of expressed genes in H9 cells +/-5kb clustered based on break distribution in H9 pluripotent samples (A). B) Break coverage in remaining pluripotent samples, clustered as in (A). scales in reads per million.

4.2.6. Investigation of promoter cluster features

In the previous section I identified four clusters of promoters in hPSC, with distinct DSB distributions. The starkest difference amongst promoter clusters was between clusters 1 and 2, which exhibit near-inverse distributions of DSBs around the TSS. Specifically, cluster 1 harbours high DSB density in the regions flanking the TSS, but with a relatively protected TSS. Conversely, cluster 2 harbours a high density of DSBs at the TSS with relatively low DSB coverage in the adjacent regions (Figure 4.12 A). I aimed to annotate genes in each cluster with the aim of finding distinguishing features between regions to inform putative mechanisms of breakage or indeed protection at TSSs.

Previously, categorising genes based on the presence of CpG islands in the 1kb region straddling the TSS, revealed CpG+ve promoters have a similar pattern of breakage to promoters of cluster 1 (Figure 4.11). To determine whether CpG island presence distinguishes between clusters, I determined the proportion of CpG+ve promoters in each cluster (Figure 4.13). Whilst cluster 1 has the highest proportion of CpG+ve promoters, cluster 2 is still comprised of >60% CpG+ve promoters, higher than the global average, demonstrating that the presence of a CpG island alone in the promoter of genes is insufficient to yield a protected TSS. Clusters 3 and 4 showed lower CpG content than the global average (Figure 4.13).

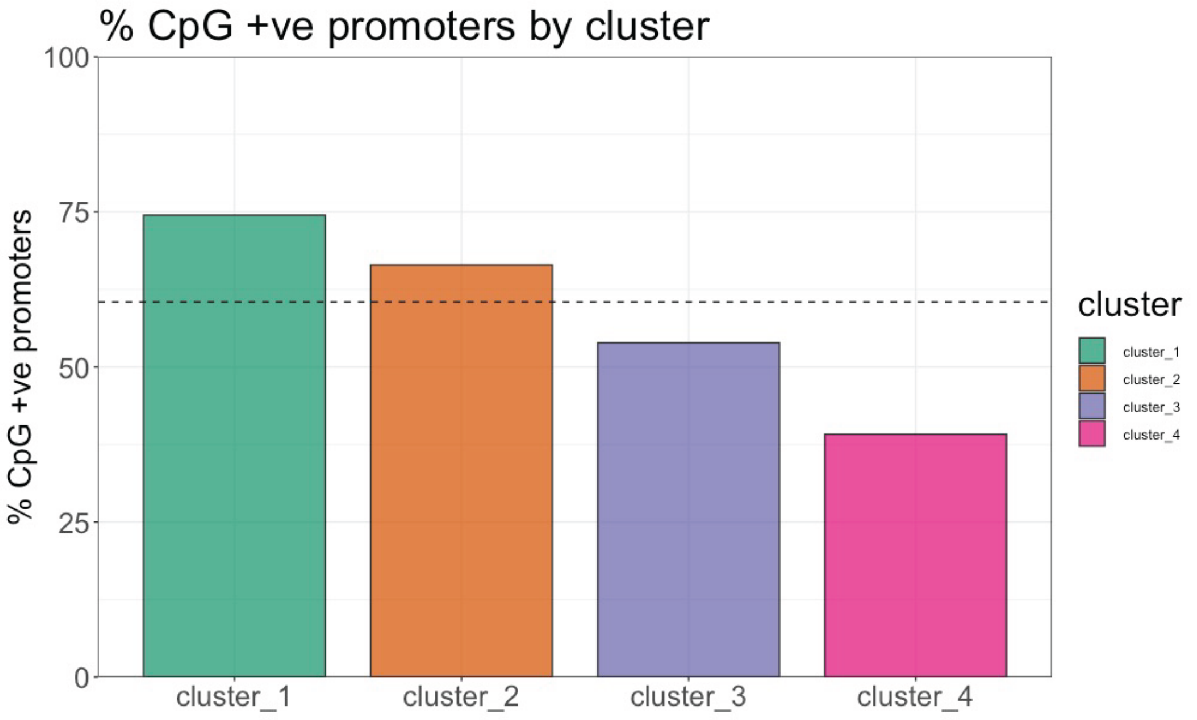


Figure 4.13 Presence of CpG islands alone does not distinguish between promoter clusters

% CpG +ve promoter containing regions in each cluster. Each promoter classified on the presence of a CpG island +/- 500bp from the TSS. Dashed line represents the global percentage of CpG+ve expressed promoters in H9 cells.

Genes harbouring CpG islands in their promoters are generally highly expressed (Deaton and Bird, 2011). As gene expression level is associated with DSB density in hPSC at both promoter regions and gene bodies, I next looked at expression levels of genes in each cluster. Cluster 1 was the most highly expressed, with clusters 2:4 exhibiting decreasing levels of expression, consistent with, and mirroring the proportion of CpG+ve promoters in genes of each cluster (Figure 4.14 A). Genes of cluster 1 are also longest (Figure 4.14 B), which may account for high level of damage downstream of the TSS. However, this does explain the protected TSS or the high DSB density upstream of the TSS. Clusters 3 and 4, which have the lowest break coverages, are typically lower expression and shorter in length than genes of the other clusters, consistent with their lower DSB coverage across the plotted region (Figure 4.14).

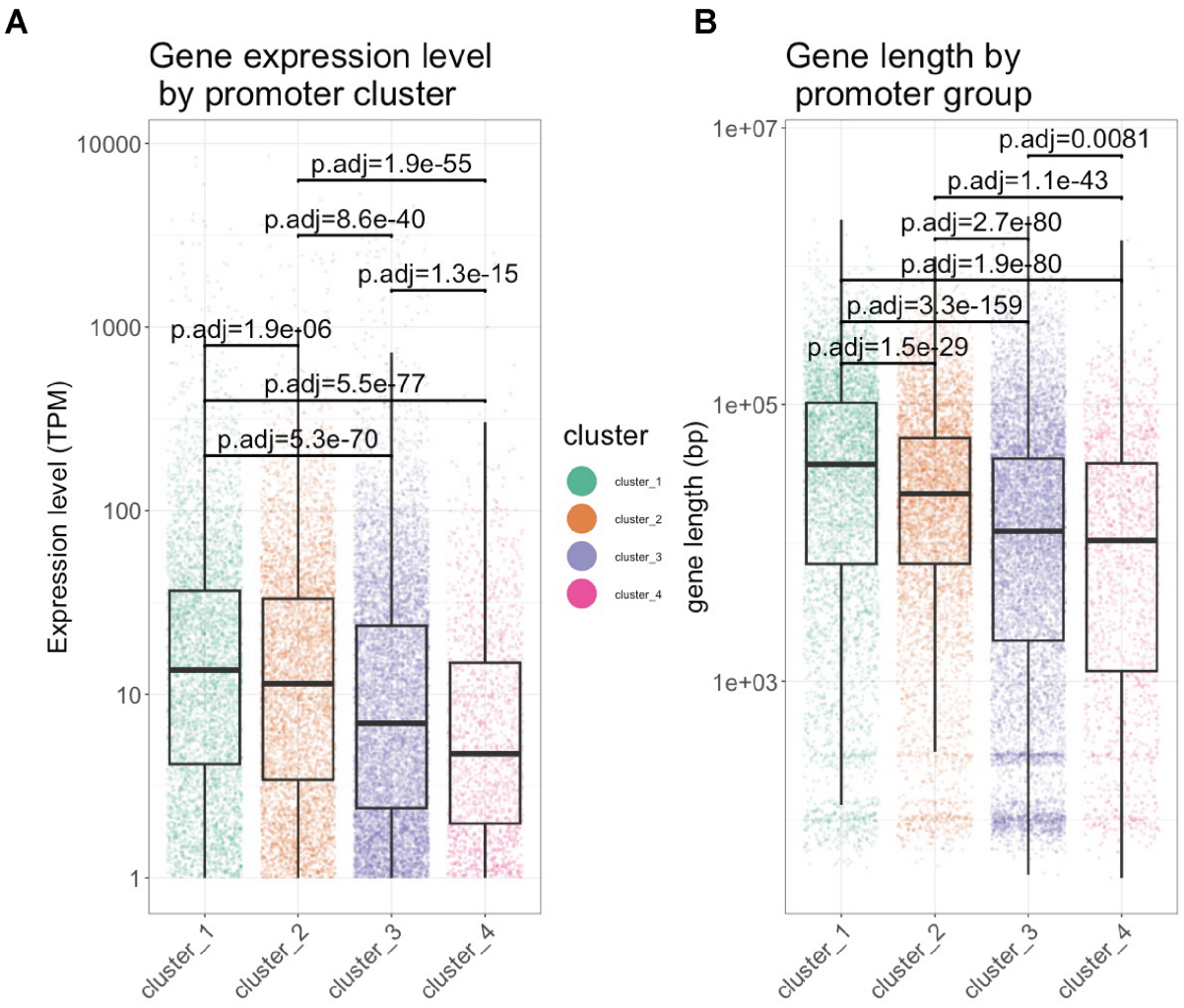


Figure 4.14 Gene expression level and length by cluster

boxplots of A) expression levels and B) lengths of genes in each cluster. Overlaid colored dots represent individual genes. Wilcoxon test (n= 4556 (cluster_1), 5410 (cluster_2), 9440 (cluster_3), 1589 (cluster_4).

I next hypothesized that gene type and therefore regulation may distinguish between clusters. To address this, I first looked at the GENCODE (Frankish *et al.*, 2019) biotype annotations of the constituent genes of each cluster. Clusters 1 and 2 had similar gene biotype compositions, both predominantly comprised of protein coding genes (Figure 4.15 A, B). The majority of genes in cluster 3 are also protein coding, whereas cluster 4 is predominantly non-protein coding genes (Figure 4.15 A, B).

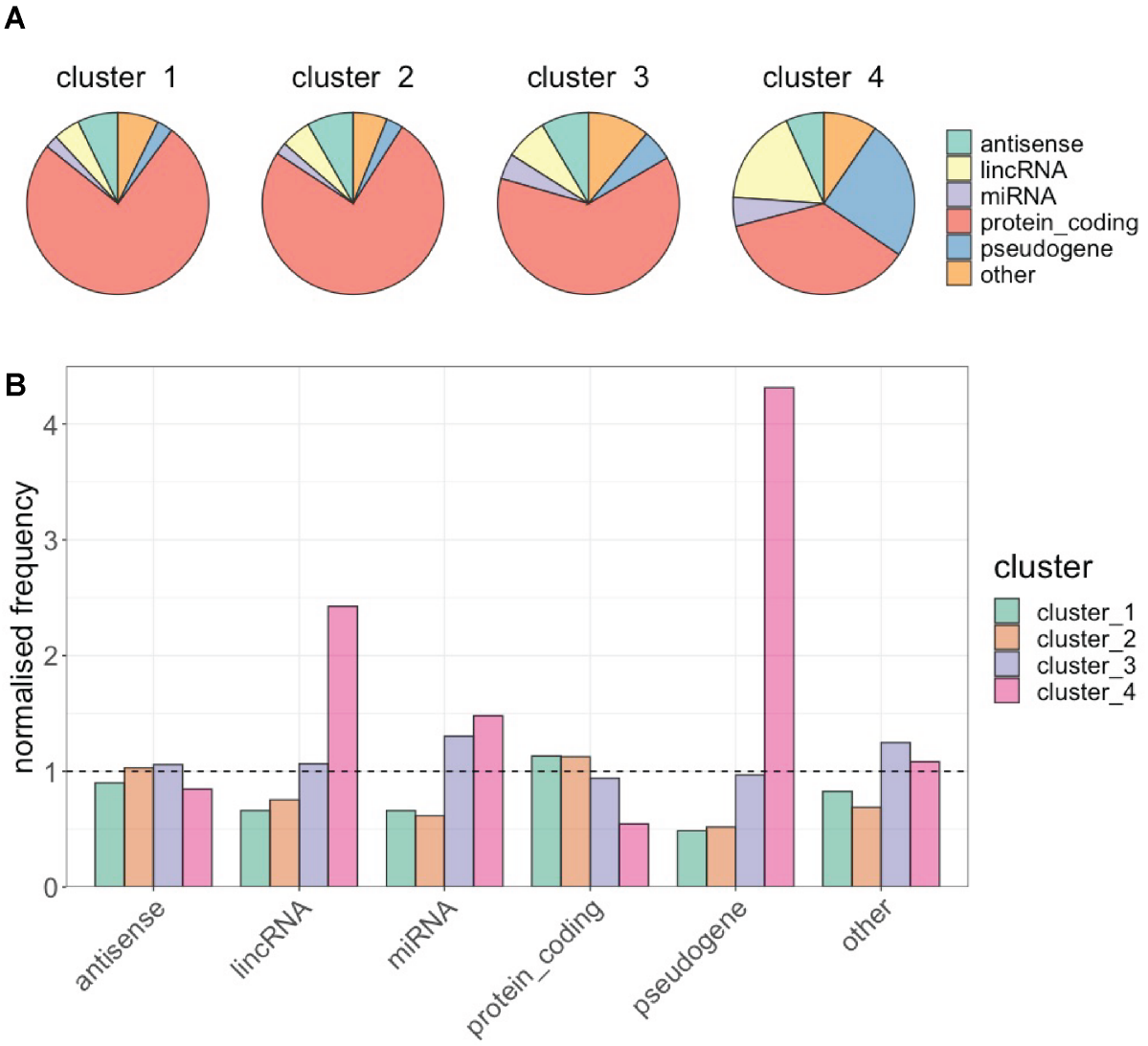


Figure 4.15 Cluster Genes biotype composition

A) Pie charts showing GENCODE gene biotype annotation proportions of constituent genes in each cluster. B) Normalised frequency (cluster frequency/global frequency) of each biotype, colored by cluster. Dashed y intercept represents global frequency.

Whilst clusters 1 and 2 are both largely comprised of protein coding genes, the roles of these gene products and ultimately regulation of genes may still differ. To determine whether clusters are enriched for different ontologies of genes, I carried out biological process GO-enrichment analysis on the constituent genes of each cluster. Assessing the commonality in enriched GO-terms between clusters, revealed that clusters 1 (n=526 enriched GO terms) and 2 (n=123 GO terms) are enriched for largely distinct GO-terms, with only 19 enriched terms in common (Figure 4.16 A). Plotting the ten most significantly enriched GO terms for each gene cluster, revealed that cluster 1 is enriched for genes associated with maintenance of the stem cell population, but also processes typically associated with differentiation including wnt

signalling and neural development. Cluster 2 by comparison is enriched for biological processes likely common to replicative cells, such as DNA damage repair and ncRNA processing (Figure 4.16 B).

Figure 4.16 GO-Enrichment analysis of clusters

A) Venn diagram of enriched GO-terms (P<0.01) in each cluster. B) Dot plots of top 20 biological process GO-terms enriched in each cluster. GeneRatio=Genes hits in GO-term/total number of input genes.

As biological process ontologies of genes in clusters 1 and 2 were distinct, I reasoned that their regulation was also distinct and could give clues as to the origin of their divergent break patterns. I therefore next looked to see if transcription factor binding sites (TFBS) were differentially enriched between gene clusters. To this end I used CiiiDER software (Gearing *et al.*, 2019), to analyse differential enrichment of TFBS between promoters of genes in clusters 1 and 2. The analysis revealed a total of 102 TFBS enriched in cluster 1 over 2, and 514 TFBS enriched in cluster 2 over 1. The differential enrichment of TFBS between clusters 1 and 2 demonstrates the potential for differential regulation of genes in each cluster. However, it is important stress that this analysis is limited to the presence of binding site, not transcription factor binding, which would require further analysis of ChIP-seq or cut and run datasets.

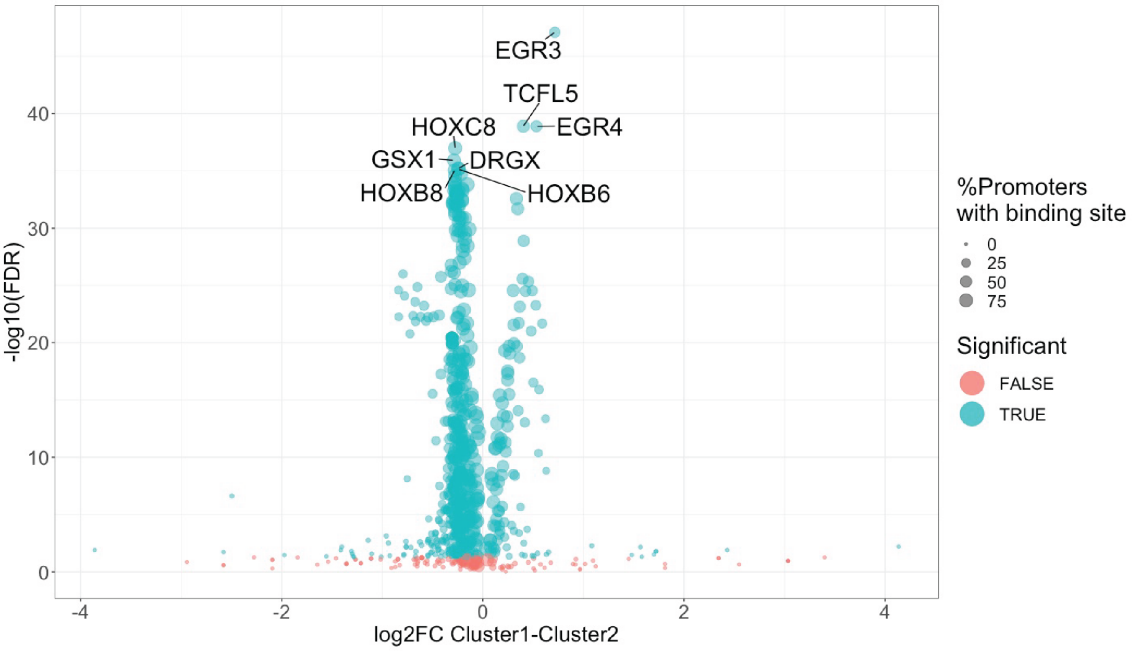


Figure 4.17 Transcription factor binding site differential enrichment cluster 1 vs cluster 2

TSS +/- 500bp of genes in each referenced against JASPAR2020 TFBS database. Volcano plot of TFBS enrichment in Cluster1 vs Cluster 2 genes. TFBS with log2FC>0.4 and -log10(FDR)>35 annotated. point size represents the percentage of promoters within each cluster harboring TF binding site.

I next wanted to consider which mechanisms of gene regulation could contribute to differential distribution of DSBs around genes' TSS, one such regulatory mechanism previously associated with DSB formation is RNAPII pausing (Dellino et al., 2019; Singh et al., 2020). Promoter proximal RNAPII pausing is a mechanism of transcriptional regulation whereby RNAPII transcription transiently halts following a brief initiation phase of 20-60nt transcription (Abuhashem *et al.*, 2022). Promoter proximal pause release is essential for transcription of

nearly all genes in mESC, however genes' degree of transcriptional pausing, i.e. rate of transcription initiation versus pause release, is variable (Jonkers *et al.*, 2014). Release of paused RNAPII to yield productive transcription elongation has been proposed to require cleavage of DNA by TOP2 (Bunch *et al.*, 2015). More recently, a DSB-mapping study in MCF5 cells identified the most fragile promoter regions as being enriched for paused genes and demonstrated a requirement for TOP2-mediated cleavage for gene expression (Dellino *et al.*, 2019). Subsequently Singh and colleagues identified TOP2-dependent break accumulation at RNAPII pause sites throughout the genome, irrespective of annotated TSS (Singh, et al., 2020). Considering the above data in support of differential regulation of genes in clusters 1:4 and, given that transcriptional pause release is documented to require DNA cleavage, I posited that differential break distribution could be due to differences in transcriptional pausing levels between gene clusters. To investigate this, I took a published RNAPII ChIP-seq dataset from a study in H9 cells (Lyu *et al.*, 2018) and calculated RNAPII pausing indices for genes of each cluster (Figure 4.18 A). Genes of cluster 4 had, on average, the lowest pausing index. Clusters 1 and 3 had comparable pausing indices. Strikingly, cluster 2 (high-damage TSS, low damage flanking) had a significantly higher pausing index than all other clusters, raising the possibility that damage immediately flanking the TSS is dependent on promoter-proximal transcriptional pausing, as reported in MCF7 cells (Dellino *et al.*, 2019).

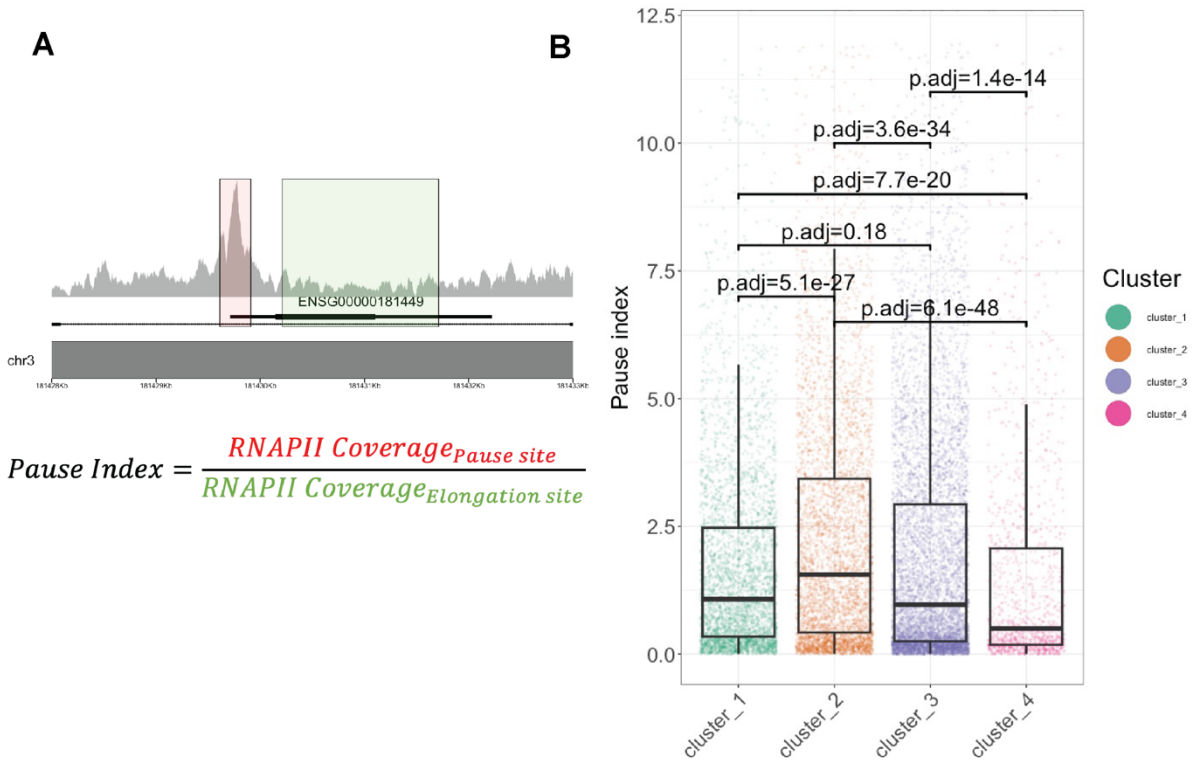


Figure 4.18 High pausing index in cluster 2 genes

A) Example of transcriptional pause site (red box TSS+300bp/-100bp) and elongation site (green box +500:2000bp) on SOX2 gene. RNAPII coverage in grey histogram. Pause index taken as RNAPII coverage in pause site/ RNAPII coverage in elongation site. B) Gene pause index by cluster. n= 3980 (cluster 1), 4848 (cluster 2), 7494 (cluster 3), 1227 (cluster 4).

Calculating pausing index as in (Figure 4.18) is contingent on accurate TSS annotation. Throughout this study, TSS were defined as the start coordinates of Ensembl genes (i.e. the first exon), a method commonly in the literature (Ray *et al.*, 2022). Notably, manual inspection of RNAPII coverage around certain Ensembl annotated TSS, reveals a sharp peak in RNAPII coverage downstream of the apparent TSS (based on first exon position) but overlapping with the start of a shorter transcript variant (Figure 4.19 A-B). RNAPII coverage does not map TSS position, however, RNAPII coverage is *typically* highest around the TSS of transcribed genes (Noe Gonzalez et al., 2021).This accumulation of RNAPII over start coordinates of alternative transcripts may represent transcriptional pausing at TSS preferentially used in H9 cells suggesting the TSS annotations used here are inaccurate.

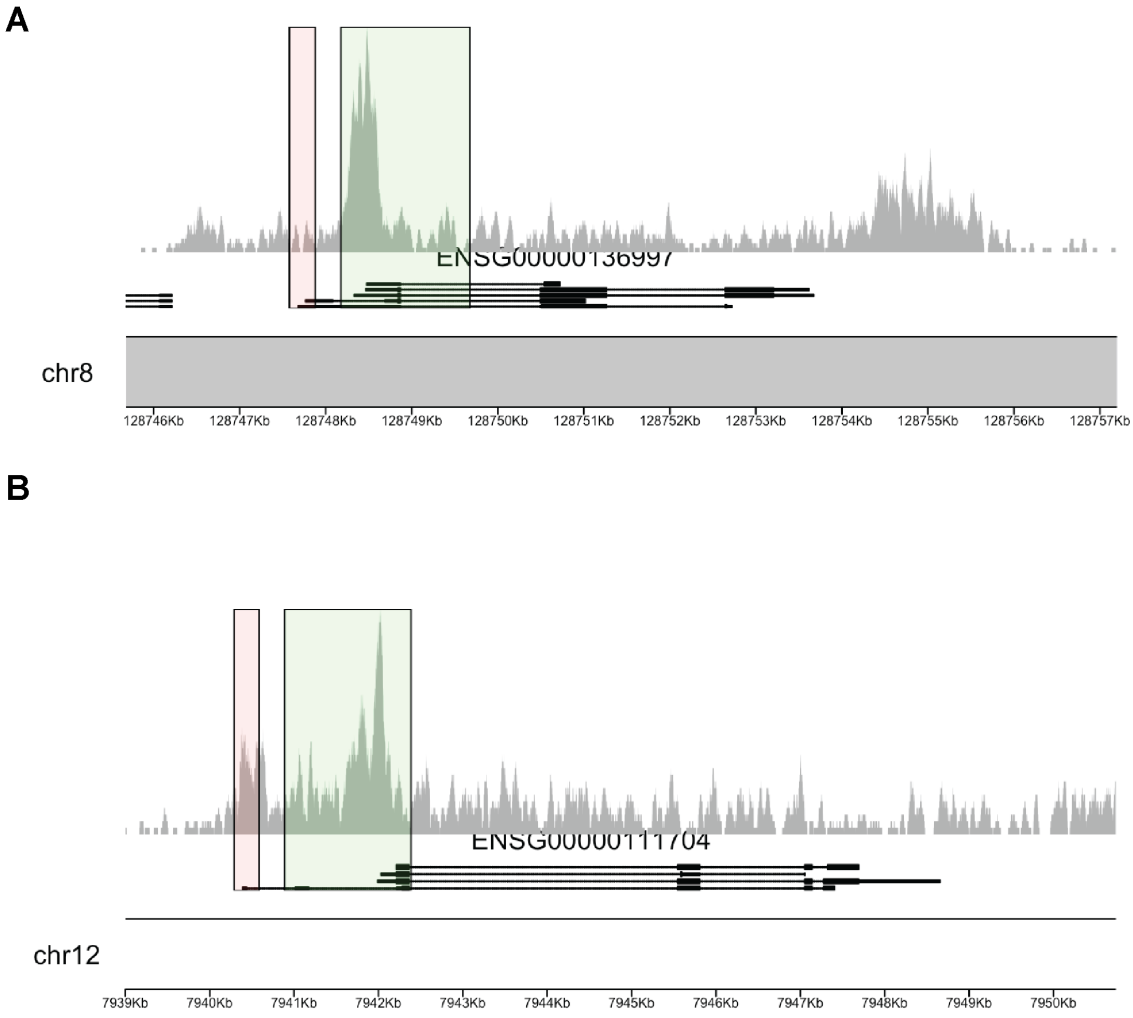


Figure 4.19 Example genes with RNAPII peak downstream of annotated TSS

RNAPII coverage (top). Nominal pause site in red box, nominal elongation site in green box. Ensembl transcripts shown below (A) MYC, B) NANOG

A possible consequence of inaccurate TSS annotation is that differential DSB distribution around TSS in genes of each cluster, simply reflects differences in position of the genuine TSS relative to the Ensembl-annotated TSS. Of particular concern is cluster 1, for which the apparent protected TSS with high-DSB flanking regions could simply reflect an averaging effect of TSS being mis-annotated upstream of the genuine TSS in some genes and downstream in others. To address whether there is systematic deviation between RNAPII maxima and annotated TSS amongst clusters, I randomly sampled 12 genes from clusters 1 and 2 and manually scored the position of RNAPII maxima relative to the TSS by genome browser visualization (Table 4.1). 10/12 genes in cluster 1 had RNAPII peaks overlapping the annotated TSS versus 7/12 genes in cluster 2. Whilst this is by no means statistically conclusive, genes of both clusters had TSS with overlapping RNAPII peaks and TSS with

up/downstream peaks. This gives no cause to believe that TSS are systematically inaccurately annotated in genes of one cluster versus another, i.e. differential break coverage relative to TSS is likely a *bona fide* biological feature, not an artefact of analysis.

Table 4.1 RNAPII maxima relative to annotated TSS

RNAPII coverage was visualized in IGV over annotated TSS used in this study. Maxima overlapping with the pause site are marked "0", upstream "-", downstream "+".

Cluster 1		**Cluster 2**	
Gene name	**RNAPII maxima position**	**Gene name**	**RNAPII maxima position**
VEZT	**0**	***TFIP11***	**0**
AKAP11	**-**	***GCAT***	**0**
SLC45A4	**0**	***EIF3L4***	**+**
JARID2	**-**	***MICAL***	**+**
MAP3K9	**0**	***AK6***	**0**
ZNF195	**0**	***SLC1A3***	**+**
HMGA1	**0**	***RPL31***	**+**
KIF13A	**0**	***PIGB***	**0**
DENND4C	**0**	***NDUFS5***	**0**
PHF21A	**0**	***SEMA4C***	**+**
FGD4	**0**	***KIF5C***	**0**
TCF12	**0**	***NUDT9***	**0**

4.3. Discussion

In this chapter, I have generated the first genome-wide map of endogenous DSBs in hPSC lines and their differentiated counterparts. I have annotated INDUCE-seq break maps in hPSC with genomic, epigenomic and transcriptional data to show that:

1) DSBs are specifically enriched in open, active chromatin in hPSC
2) Mapped R-loops are genomic cold-spots of DNA damage in hPSC
3) hPSC suffer transcription-associated damage in promoters and gene bodies.
4) Pluripotent promoter regions exhibit four distinct patterns of break coverage, distinguished by RNAPII pausing.

4.3.1. DSBs are enriched in open, active chromatin in H9 hESC

Prior to annotation studies, I mapped DSBs in each sample, finding DSBs to be widespread throughout the genome. Annotation of genic features is the simplest form of annotation and, as RefSeq genomic annotations are not cell-type specific, they are applicable to all samples in the study (O'Leary *et al.*, 2016). Annotating DSBs with RefSeq features revealed enrichment of DSBs at all genic regions, most notably promoters, in all pluripotent and differentiated samples, consistent with break-mapping studies in other cell types, and suggestive of transcription-associated DNA damage (Baranello *et al.*, 2014; Yan *et al.*, 2017; Dellino *et al.*, 2019; Gothe *et al.*, 2019; Sandeep Singh *et al.*, 2020). To investigate whether certain genes were predisposed to DSBs based on epigenetic context, I next used published sequencing datasets to define chromatin states, revealing enrichment of DSBs at open, active chromatin, most notably active promoters and, to a lesser extent, enhancers in H9 hESCs, consistent with studies in other cell types (Mourad *et al.*, 2018; Ballinger *et al.*, 2019; Dellino *et al.*, 2019; Hidmi and Aqeilan, 2022).

Whilst blacklisted regions with poor mappability were excluded from this analysis, more subtle differences in regional mappability could still yield artefactual spikes and troughs in the data. In ChIP-seq experiments, use of an input sample, i.e., gDNA prior to immunoprecipitation, can be used to determine background levels of read coverage, which serve as a control to determine enrichment in the immunoprecipitated sample. No such control sample was included here; indeed, to the best of my knowledge, no such control has been used in previous DSB-mapping studies. However, one could conceivably use an in-silico approach to determine mapping biases. Bioinformatic tools such as ART (Huang *et al.*, 2012), synthesise reads from a reference genome with platform-specific error rates. Such reads could then be subjected to the same QC and alignment pipeline as INDUCE-seq to determine regional biases in read

alignment efficiency. INDUCE-seq read coverage could then be normalised based on the calculated regional mappability. It should be noted that such an approach would still fall short of identifying biases introduced at the point of break labelling. For instance, ChIP fragments and input samples are both enriched for euchromatin (Rozowsky *et al.*, 2009) and it is possible that similar biases exists in INDUCE-seq or indeed other DSB mapping techniques.

4.3.2. Mapped R-loops are genomic cold-spots of DNA damage in hPSC

There is a large body of literature demonstrating that pathological R-loops cause DSBs in diverse cell types, particularly highly proliferative cells (Sollier and Cimprich, 2015). Very few studies to date have mapped DSBs and R-loops in parallel to determine levels of enrichment. Notably however, a recent DSB-mapping study reported accumulation of DSBs at R-loop peaks, proposed to be a result of TOP1 activity, rather than conflict with DNA replication forks (Hidmi and Aqeilan, 2022). Conversely, Promonet and colleagues identified replication-dependent R-loop driven damage in the terminators of converging genes, specifically in the absence of TOP1 (Promonet *et al.*, 2020). To investigate the contribution of R-loops to genome damage in hPSC, I accessed a published DRIP-seq dataset (R-loop mapping) (Yan et al., 2020), which I used alongside other ChIP-seq datasets to define the aforementioned chromatin states. Surprisingly, I observed that the two R-loop high states were significantly depleted for DSBs. As R-loop mediated transcription-replication collisions or indeed flap-endonuclease attack of R-loops would cause DNA damage *adjacent to* as opposed to *overlapping* the R-loop, I analysed DSB enrichment in R-loop peaks as well as their immediately flanking regions. Again, I noted a significant depletion of DSBs in these regions, demonstrating that mapped R-loops from (Yan et al., 2020), are *bona fide* genomic cold-spots of DNA-damage.

This unexpected result could be due to effective resolution of pathological R-loops in hPSC. Similar is seen in cancer cells, which are typically highly proliferative and employ a battery of mechanisms to resolve R-loops, to facilitate rapid DNA replication, including upregulation of R-loop resolving helicases or components of the Fanconi Anaemia pathway to resolve R-loops ahead of DNA replication forks, and overexpression of R-loop resolving chromatin remodellers (Boros-Oláh *et al.*, 2019; Prendergast *et al.*, 2020). Indeed, hPSC are also highly proliferative cells with concurrent high levels of transcription (Efroni *et al.*, 2008) and it is likely that they employ some of the same mechanisms to ensure resolution of potentially genotoxic R-loops. Competent resolution of genotoxic R-loops in wild-type, unchallenged cells would be consistent with a recent break-mapping study in lymphoblastoma cells which noted that DSB enrichment in R-loop forming sequences was only detectable under conditions of replication stress, in the absence of FMRP protein, which the authors proposed to resolve R-loops

(Chakraborty *et al.*, 2020). By contrast, Hidmi and Aqeilan recently demonstrated a marked increase in break density at R-loop peaks, mapped by DRIP-seq, and used TOP1 knockdown to demonstrate the requirement of TOP1 for DSBs in gene bodies but not promoters of R-loop positive genes (Hidmi and Aqeilan, 2022). These conflicting results likely reflect differences between cell types in R-loop metabolism.

An alternative explanation for the depletion of DSBs in mapped R-loops in this study, is that the published DRIP-seq dataset used here, may not map pathological R-loops. Firstly, it is worth mentioning that DRIP-seq data generated by Yan and colleagues lacks a crucial control condition whereby DNA is RNase H treated, to specifically degrade the RNA component of R-loops, prior to immunoprecipitation (Yan et al., 2020). The S9.6 antibody used in DRIP-seq has affinity for dsRNA as well as DNA:RNA hybrids (König *et al.*, 2017; Hartono *et al.*, 2018). It is therefore vital to demonstrate sensitivity of S9.6-mapped R-loops to RNase H degradation to distinguish genuine R-loop signal from noise (Vanoosthuyse, 2018). I also noted that no R-loop peaks called from this dataset overlapped with promoters of commonly used positive control loci: *RPL13A* and *CALM3* (data not shown) (Sanz and Chédin, 2019), raising concerns over the accuracy of the DRIP-seq dataset used here. Another caveat in this dataset is that S9.6-based methods of R-loop mapping have been shown to preferentially enrich broad benign R-loops, in gene bodies, whilst methods using catalytically dead RNase H (e.g. R-ChIP) enrich for genotoxic short R-loops associated with promoter-proximal pausing (Castillo-Guzman and Chédin, 2021). A more comprehensive map of R-loops in H9 cells would require complementary R-ChIP and DRIP-seq (Chen *et al.*, 2017). Determining the enrichment of DSBs at distinct R-loop classes, would allow more robust conclusions to be drawn about R-loops and their contribution to DSBs in hPSC. However, at the time of writing, no R-ChIP or equivalent dataset is publicly available for any of the lines used in this study and as such, this analysis was omitted from the current study. Together, based on the analysis of the available data, I concluded that DNA damage is not associated with DRIP-mapped R-loops in H9 hESC.

The specific depletion of DSBs in mapped R-loops could reflect a role for R-loops in DSB repair. R-loops have been shown to form at sites of DSB induction, where they are proposed to recruit components of the HR pathway for faithful repair of the lesion (Ohle *et al.*, 2016; Gómez-Cabello *et al.*, 2022). Thus conceivably, R-loop negative DSBs may persist longer and hence be more abundant.

4.3.3. Highly expressed genes have higher DSB densities in promoter regions and gene bodies.

In light of observed DSB enrichment in open, active chromatin, I next sought to determine whether breakage was associated with gene expression in hPSC. Categorising genes based on expression level revealed higher densities of DSBs in promoters of highly expressed genes. This association between promoter break density and expression level of genes is very commonly reported in the literature and attributed to active transcription (Baranello *et al.*, 2014; Yan *et al.*, 2017; Dellino *et al.*, 2019; Gothe *et al.*, 2019; Sandeep Singh *et al.*, 2020; Ballarino *et al.*, 2022). Studies probing the mechanistic basis of promoter-specific accumulation of damage have implicated the action of TOP2, facilitating transcriptional pause release, or transcription-replication collisions to this damage. Dellino and colleagues demonstrated that DSBs are enriched at promoter-proximal RNAPII pause sites and that proper TOP2 function is essential for transcription of these genes (Dellino *et al.*, 2019). Shortly after, others demonstrated that pause sites are enriched for DSBs irrespective of the presence of an annotated TSS and that break enrichment was dissipated in TOP2B KO cells (Singh et al., 2020). Zampetidis and colleagues propose transcription-replication collisions as the source of DSB accumulation in promoter regions and use transcription elongation inhibition with DRB as a means of validation (Zampetidis *et al.*, 2021). However, as DRB specifically inhibits RNAPII elongation, this does not exclude the possibility of TOP2-driven breakage facilitating RNAPII pause-release. Whilst transcription-replication collisions have been mechanistically demonstrated to cause breakage along the length of transcriptional units (Macheret and Halazonetis, 2018; Tena *et al.*, 2020; Michel *et al.*, 2022; Zhang *et al.*, 2022), specific enrichment of DSBs in promoter regions has thus far only been mechanistically attributed to the action of TOP2, thus it is possible that expression-proportional break accumulation in promoters in H9 cells is due to the action of topoisomerase enzymes facilitating RNAPII pause release.

Quantifying break densities across gene bodies also revealed higher break density in bodies of highly expressed genes. Increased gene expression level does not universally correlate with gene body break density in the literature. For instance Singh and colleagues, identified TOP2-induced damage proportional to expression level in promoter regions, but saw no effect of expression level on gene body breakage (Singh et al., 2020). Several HTGTs-based studies noted increased breakage in gene bodies of actively transcribed genes (Wei *et al.*, 2016; Michel *et al.*, 2022), particularly under conditions of replication stress (Tena *et al.*, 2020), suggesting transcription-replication collisions may be causative. Most notably, Macharet and Halazonetis (2018) recently demonstrated the direct role of transcription-replication collisions

in gene body breakage. Using replication and transcription mapping, they identified that cells with a short G1-phase activate intragenic replication origins which are unable to restart following replication stress, unless transcription is inhibited beforehand, indicating that transcription causes fork collapse. Subsequent HTGTS revealed marked enrichment of DSBs in highly transcribed genic domains replicated by these oncogene-induced replication origins (Macheret and Halazonetis, 2018). Like cancers, hPSC have an abbreviated G1-phase of the cell cycle (Becker *et al.*, 2006) and are known to suffer constitutive replication stress (Halliwell *et al.*, 2020). Thus, it is possible that increased break density in the bodies of highly transcribed genes is due to collisions with replication forks emanating from “oncogene-induced” origins. In further support of this hypothesis, I also noted increased gene body break density in longer genes. Longer genes take longer to transcribe, rendering spatiotemporal separation of transcription and replication more difficult. Accordingly longer genes have been shown to suffer high transcription-replication collision mediated break frequencies driving chromosomal fragility (Helmrich *et al.*, 2011).

4.3.4. Pluripotent promoter regions exhibit four distinct patterns of DSB coverage, distinguished by RNAPII pausing.

Plotting mean coverage of DSBs over promoter regions revealed a conspicuous, rarely reported decrease in breakage in the ~500bp immediately downstream of the TSS, proposed by some to be a feature of CpG island containing promoters (Yang *et al.*, 2015; Ballarino *et al.*, 2022). Instead of assuming that this dip was a feature of CpG islands, I carried out unbiased clustering of promoters based on break distribution and identified 4 distinct classes common to all five hPSC lines used in this study. Most interestingly, classes 1 and 2 exhibited near-inverse break coverage profiles, with class 1 suffering high levels of genome damage upstream and downstream of a protected TSS, whist cluster 2 harboured a sharp peak of DSBs near the TSS with relatively low breakage in the flanking regions. I reasoned that differential DSB patterning in promoter clusters was caused by different mechanisms and looked for distinguishing features between promoters of each cluster. Firstly, I noted that cluster 2 promoters, with a sharp peak of DSBs over the TSS, contained CpG islands in >60% of their promoters, arguing against the suggestion that CpG island presence in promoters yields a protected TSS (Yang *et al.*, 2015; Ballarino *et al.*, 2022). Genes in cluster 1 were more highly expressed and longer than genes in cluster 2, consistent with high break density downstream of the TSS, but failing to account for the protected TSS. Clusters 1 and 2 were comprised largely of protein coding genes, however GO-enrichment analysis revealed largely independent biological functions were enriched in each sample, moreover TSS-proximal

regions were differentially enriched for transcription factor binding sites suggesting genes of each may be differentially regulated.

I next looked at promoter proximal pausing of RNAPII, as a potential mechanism of differential regulation and DNA breakage between clusters. Gene transcription is comprised of three steps: initiation, elongation, and termination, each of which is controlled as a means of regulating gene expression (Noe Gonzalez *et al.*, 2021). Promoter-proximal pausing occurs between initiation and elongation steps and is often detectable as an accumulation of RNAPII 20-60nt downstream of the TSS. Mechanistically, it entails CDK9-mediated phosphorylation of elongation-inhibitory factors NELF and DSIF9 causing their dissociation from RNAPII and a concurrent conformational change, tilting RNA within the active site into the catalytically active conformation, resulting in pause release and productive elongation (Abuhashem *et al.*, 2022). Whilst CDK9-mediated pause release is proposed to be necessary for expression of > 95% of genes in mammals (Jonkers *et al.*, 2014), several studies have proposed relief of DNA supercoiling by TOP2 cleavage, to be a necessary step for pause-release specifically in genes with a high pause index (accumulation of RNAPII in pause regions relative to elongation regions) (Bunch et al., 2015; Dellino et al., 2019; Singh et al., 2020). Reminiscent of this, I found genes of cluster 2 (DSB peak over TSS), to have a significantly higher pause index than genes of the other clusters indicating possible TOP2-mediated damage. Alternatively, promoter-proximal RNAPII pausing is known to generate short, genotoxic R-loops which are not detected by DRIP-seq, which may yield damage in cluster 2 gene promoters (Chen *et al.*, 2017; Castillo-Guzman and Chédin, 2021). A final possibility is that SSBs generated during demethylation to facilitate transcriptional activation, are processed to DSBs (Wu *et al.*, 2021; Wang *et al.*, 2022; Ray *et al.*, 2022). I propose high break coverage over the TSS of cluster 2 genes results from high levels of RNAPII pausing at these genes, either mediated by TOP2 or R-loops. Proper resolution of the mechanistic cause of cluster 2 promoter DSBs would require further rounds of sequencing: INDUCE-seq in TOP2-depleted cells, to determine the contribution of TOP2 to cluster 2 DSBs, and R-ChIP to map promoter-proximal R-loops. Owing to time constraints, neither were adopted here.

The origin of the conspicuously protected TSS in cluster 1 genes remains elusive. It is tempting to speculate that, as highly expressed genes, promoter regions of genes in cluster 1 may be bound by protein complexes which confer protection to the TSS proximal region. In the eventuality of TOP2-induced damage this could act by restricting topoisomerase access to the underlying DNA template. Alternatively in a model of R-loop induced damage, this could be achieved by facilitating rapid pause-release of RNAPII into active elongation, to prevent formation of R-loops, or indeed recruitment of R-loop resolving factors. Further work is required to determine the mechanistic basis of this phenotype.

Ahead of embarking on further characterisation of clusters, additional validation could be carried out to increase confidence in the biological validity of the clusters identified here. To test the robustness of clusters, one could down sample the data set at random and attempt to repeat clustering using the same parameters. If the clusters are robust biological features, one would expect the same clusters to be identified, comprising the same genes. If the algorithm identified different clusters following down sampling, the clustering approach may be inappropriate and alternative clustering methods, or cluster numbers, would be necessary to identify bona fide biological clusters. Given the reproducibility of gene clusters across cell lines, such an approach was omitted here.

In summary the work in this chapter reveals DSB accumulation in transcriptionally active regions of the genome, implicating transcription-associated damage in hPSC. This knowledge is of immediate use in the following chapter, where I identify and validate causes of DSBs in damage hotspots corresponding to sites of recurrent structural rearrangement in hPSC.

5. Identification and validation of DNA damage hotspots yielding genetic variants in hPSC

5.1. Introduction

In the previous chapter I annotated DSB maps in hPSC with genomic, epigenomic and transcriptomic data, finding transcription to be the unifying feature in DSB enrichment at a global scale. In this chapter I focus on unbiased identification of sites of DSB enrichment, herein referred to as *DSB hotspots,* with a view to determining which, if any, of these hotspots overlap with sites of recurrent genetic changes observed in hPSC cultures and ultimately determining the mechanistic cause of breakage at these sites.

5.1.1. Advantages of hotspot identification

Hotspot calling has several advantages over genome-wide profiling of DSBs. Firstly, hotspot calling can be carried out in an unbiased manner. Genome-wide annotation, as carried out previously, determines DSB enrichment at specific pre-defined regions, which requires assumptions as to where DSBs may be enriched. By contrast, identification of DSB hotspots can be carried out in the absence of any annotation and therefore be centred entirely around the break data.

Another advantage of hotspot identification is that, identifying regions of high break density relative to the global average, should mitigate the effect of noise in the data. By labelling breaks *in situ* and lacking PCR-based amplification, INDUCE-seq minimises opportunity for introduction of noise during library preparation (Dobbs *et al.*, 2022). However, there are still several potential sources of noise in the data, firstly, via fragmentation of genomic DNA in apoptotic cells (Nagata *et al.*, 2003). During apoptosis, a network of nine endonuclease enzymes is activated, which digest cellular genomic DNA into nucleosomal fragments (Kawane and Nagata, 2008). As INDUCE-seq maps DSBs on a population scale, a subpopulation of apoptotic cells could contribute regular inter-nucleosomal DSBs, potentially masking genuine individual endogenous DSBs. As such, noise caused by apoptosis-induced breaks would be regularly distributed between nucleosomes and identifying DSB hotspots could overcome the noise. Alternatively, noise due to mis-mapping of breaks could be introduced by DNA end resection at homologous repair (HR) intermediates. In HR, 5’ ends of DSBs are resected to yield 3’overhangs of ssDNA which are required for subsequent strand

invasion and homology directed repair (Mehta and Haber, 2014). Short-range resection is first achieved by MRE11-RAD50-NBS1 (MRN) in complex with CtIP, which resects ~300 bases (Sartori *et al.*, 2007), and subsequent long-range resection is carried out by EXO1 or BLM/DNA2, typically yielding several kb of 3' overhang (Nimonkar *et al.*, 2011). INDUCE-seq library preparation requires blunting of DNA ends prior to sequencing adapter ligation. Specifically, T4 DNA polymerase synthesizes DNA to fill 5' overhangs, or digests 3' overhangs (Dobbs *et al.*, 2022). The result of this, in the case of resected HR intermediates, is that the sequencing read pertaining to the break, maps to a region distal to the endogenous break location, the distance of which is dependent on the extent of resection (Figure 5.1). As resection distance is variable and increases as a function of time (Nimonkar *et al.*, 2011; Canela *et al.*, 2016), genomic regions harbouring recurrent frequent breaks should still be identifiable as hotspots.

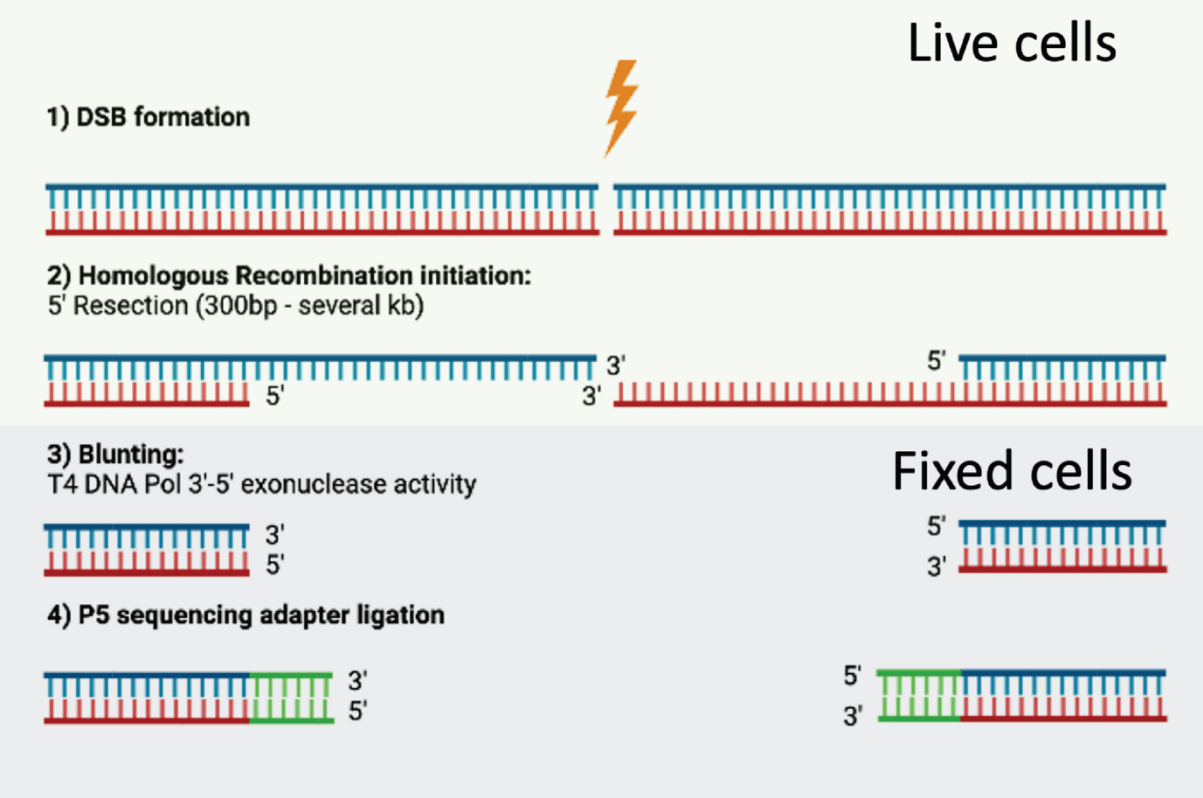


Figure 5.1 Effect of 5' resection on INDUCE-seq break mapping

1) An endogenous DSB is formed in live cells. 2) if DSB ends are processed for repair by HR, short range resection (~300 bases) followed by long-range resection (several kb) resect 5' ends of DNA to yield 3' overhangs. 3) In fixed cells during INDUCE-seq library preparation, 3' overhangs are specifically digested to yield blunt DSB ends. 4) P5 sequencing adapters are ligated to the newly blunted ends. The 5' nucleotide adjacent to the P5 adapter is mapped as the break end, at a location distal to the initial DSB.

Finally, grouping DSB data into genomic intervals >1bp in length enables identification of regions which are commonly broken in samples of a given condition and specifically enables

statistical comparison of region-specific DSB frequency between samples. Therefore, calling hotspots will allow identification of regions of high DSB density which are common to all hPSC lines and specifically higher DSB density in pluripotent than differentiated cells.

5.1.2. Methods of hotspot identification in DSB-mapping studies

Identifying genomic regions of enrichment has long been established in the field of epigenetics, where sites of significant accumulation of overlapping ChIP-seq reads are identified in a process known as *peak calling* (Park, 2009). Many different algorithms are available for peak calling in ChIP-seq datasets, which differ in their approaches to identify potential peaks and calculate statistical significance (Thomas *et al.*, 2017). Despite being a chromatin assay, INDUCE-seq is not ChIP-seq like. ChIP-seq reads correspond to fragments of protein-bound DNA, with often overlapping or immediately adjacent reads representing the same binding site of a given protein. By contrast, each INDUCE-seq read corresponds to a DSB end in the starting population of cells. Endogenous DSB positions are, by definition, 1 nucleotide in length and are very rarely overlapping. INDUCE-seq was developed for the detection of off-target CRISPR/Cas9 cut sites in genome editing, (wherein breaks accumulate at the exact same nucleotide position (Dobbs *et al.*, 2022)), and has rarely been used for detection of endogenous breaks (Fletcher *et al.*, 2022). As such, there is no off the shelf method for detection of DSB hotspots using this type of data.

Previous DSB-mapping studies have made use of standard ChIP-seq peak calling algorithms for detection of DSB hotspots. HTGTS-based studies in neural progenitor cell types have used the SICER algorithm, which identifies regions of read accumulation in-non overlapping windows, then merges identified windows within a user-defined distance of one another, into broad peaks (Zang *et al.*, 2009). Using this approach, HTGTS studies have identified low numbers of broad (~0.1-2Mb) DSB hotspots in neural progenitor cell types, with breakage attributed to transcription-replication collisions in long actively transcribed genes (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020).

By contrast, direct break-labelling studies using peak-calling approaches, predominantly opt for the MACS2 algorithm (Lensing *et al.*, 2016; Mourad *et al.*, 2018; Tubbs *et al.*, 2018; Gothe *et al.*, 2019; Chakraborty *et al.*, 2020), which uses sliding windows of 600bp length by default, to identify regions of enrichment on opposing strands, and takes the distance between the maxima as a putative peak (Zhang *et al.*, 2008). The default mode of analysis harnesses the fact that, in ChIP-seq, immunoprecipitated DNA fragments are sequenced as short reads in a 5'-3' orientation. The sequencing reads are invariably shorter than median fragment size and therefore aligned reads are, on average, shifted upstream of the binding site on the forward

strand of DNA and downstream on the reverse strand which, applied to DSB mapping, could select specifically for double-ended DSBs. However many break mapping studies stipulate that cross correlation between strands is not required, mitigating this effect (Lensing *et al.*, 2016; Tubbs *et al.*, 2018; Gothe *et al.*, 2019). Notably Dellino and colleagues use an alternative peak calling algorithm (*Homer findPeaks*) to independently identify narrow peaks on forward and reverse strands, and only call hotspots where peaks on opposing strands are within 5kb of one another, thereby specifically selecting for double-ended DSBs and likely biasing towards identification of the TOP2-mediated damage described (Dellino *et al.*, 2019). MACS2-based peak calling, under default settings, is biased towards identification of narrow peaks (Zang *et al.*, 2009), which likely contributes to MACS2-based studies universally identifying narrow DSB hotspots (Lensing *et al.*, 2016; Mourad *et al.*, 2018; Tubbs *et al.*, 2018; Gothe *et al.*, 2019; Chakraborty *et al.*, 2020), often attributed to topoisomerase-induced damage at specific, short genomic features (Gothe et al., 2019; Szlachta et al., 2020). Thus, existing methods for DSB hotspot identification from break-mapping data require presumptions on the size of hotspots, likely skewing the output.

5.1.3. Replication stress and chromosomal fragility

Common fragile sites (CFS) are regions of the genome prone to incomplete replication and breakage across multiple cell types, they are defined by a sensitivity to replication stress (Glover *et al.*, 1984), and typically replicate late during S-phase, proposed to heighten sensitivity to replication perturbation, owing to a lack of time to recover stalled forks prior to mitosis (Durkin and Glover, 2007). Replication stress can yield DNA breaks at CFS independent of transcription (LeTallec *et al.*, 2013). However, work in the previous chapter noted a strong association between transcriptional activity and DSB enrichment in hPSC.

There are two principal models for chromosomal fragility stemming from transcription-associated replication stress. The first is via transcription-replication collision, as proposed by Helmrich and Ballarino, who demonstrated that fragility of long genes harbouring CFS, was dependent on concurrent DNA replication and transcription and that such fragility was exacerbated in RNase H knockdown, implicating R-loops at transcription replication collisions as the cause (Helmrich *et al.*, 2011). Alternatively, stalling of the replication fork upon transcription replication collision may cause uncoupling of the replicative helicase, from the fork, generating long stretches of ssDNA prone to secondary structure formation, ultimately yielding a more potent block to replication (Sinai *et al.*, 2019).

The second model of chromosomal fragility stemming from transcription-associated replication stress is via modulation of DNA replication origins. Several studies have

demonstrated that, prior to origin firing, processive transcription can reposition (Gros *et al.*, 2015b), or even displace (Macheret and Halazonetis, 2018) licensed origins. Over long transcriptional units, this can yield long stretches of origin-poor DNA and consequently stalled forks are not readily rescued by convergence with neighbouring forks. Consistent with this model, a recent study in lymphoblastoid cells demonstrated that fragile sites in long genes were origin-poor and relied on individual unidirectional forks to replicate megabase scale regions of DNA, explaining an increased sensitivity to aphidicolin (Brison *et al.*, 2019). Another study by Le Tallec and colleagues showed that, whilst CFS predominantly overlapped with large genes, break accumulation often extended beyond gene bodies in various human cell lines, and did not correlate with transcriptional activity, suggesting transcription replication collisions were not causative (LeTallec *et al.*, 2013). In either model, it is apparent that transcription, at some level, exacerbates replication stress-driven chromosomal fragility (Ji *et al.*, 2022).

I hypothesize that DSBs recur at certain genomic loci, specific to cell types, and that the location of these DSBs in hPSC correspond to sites of recurrent structural rearrangements. In the following work, I plan to identify recurrent DSB hotspots in hPSC, and determine causes of breakage yielding genetically variant cells. Specifically, I aim to:

i) Develop an unbiased method for the identification of DSB hotspots in hPSC
ii) Compare hotspot breakage between pluripotent and differentiated states
iii) Identify hotspots corresponding to sites of structural rearrangements in hPSC
iv) Determine and experimentally validate the cause of genetic variation at candidate sites.

5.2. Results

5.2.1. Comparison of methods for DSB hotspot identification from INDUCE-seq data

As peak callers previously used in DSB hotspot identification are biased in terms of size and form of DSB enrichment captured, I sought to develop an unbiased approach for identifying regions of DSB enrichment in samples. In the following section, I compare three methods for identification of DSB hotspots in samples. Existing methods of peak-calling require user input of a scale at which to search for regions of enrichment (Starmer and Magnuson, 2016). Whilst scale inevitably affects the output, for the purpose of comparing methods in the first instance, I used a fixed interval length of 1kb across all methods. Similarly, the threshold above which

regions are called as hotspots affects output, but for the purpose of comparison, has been set at the top five percent of regions genome wide. Both scale and threshold value are optimized later in the chapter.

My first approach was to divide the reference genome into sequential bins, of a fixed length, 1kb (i.e. 1-1000 bp, 1001-2000 bp etc.), and count the number of breaks which fall in each bin (Figure 5.2 A). All bins are of equal size and therefore sub-setting the most broken regions simply entails identifying the top 5% of bins based on break count (Figure 5.2 B).

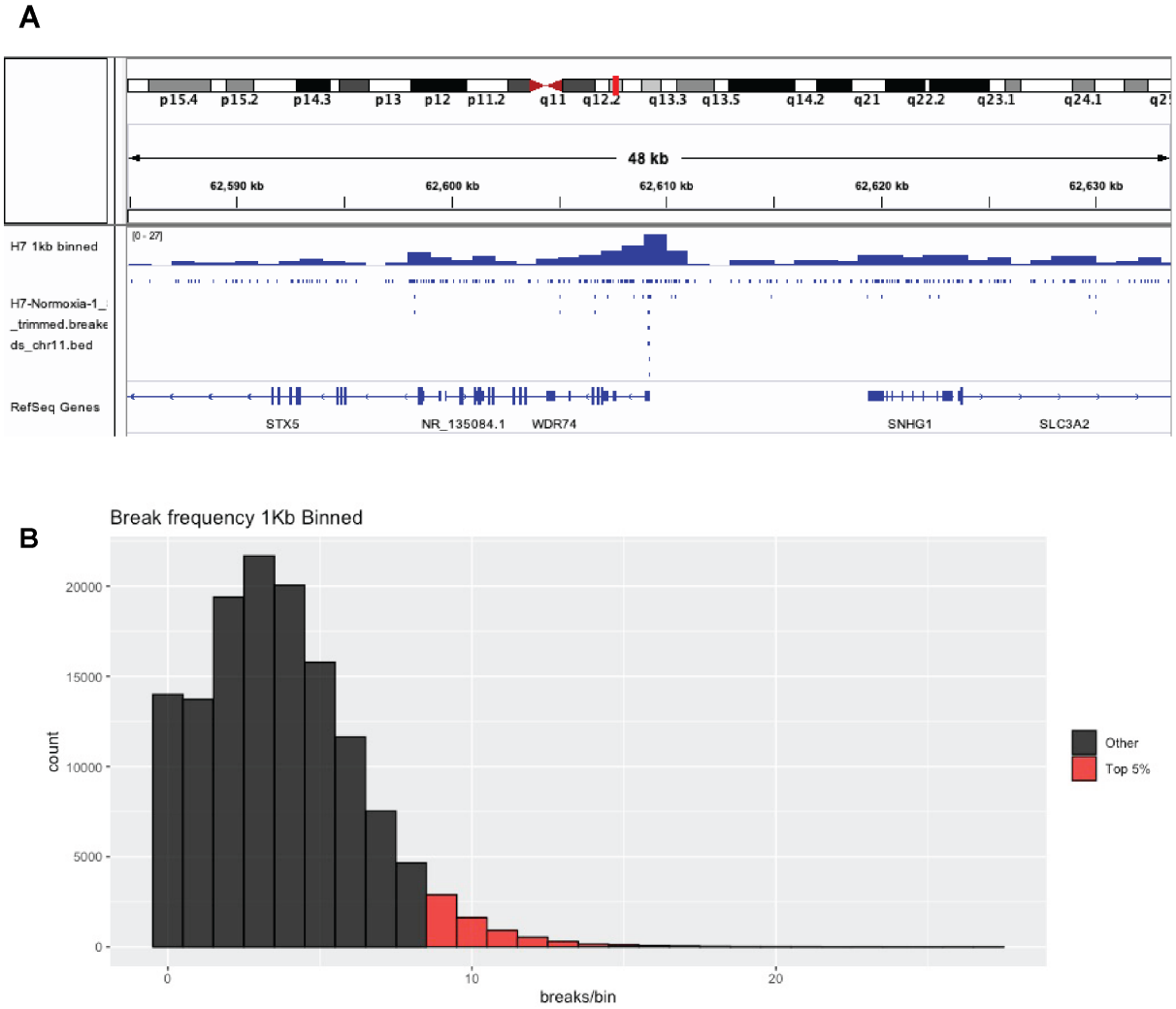


Figure 5.2 Genome binning for hotspot identification

> A) *Representative genome browser track of a 48kb region on Chr11 including observed accumulation of breaks in promoter region of WDR74. Top track, 1kb binned data, height represents the number of breaks overlapping with each interval. Lower track shows individual DSB ends in a sample H7 pluripotent dataset B) Histogram of break density distributions in individual 1kb bins (data from Chromosome 11 only in sample H7 pluripotent dataset). Top 5% based on break count in red.*

A notable shortcoming of this approach is that the boundaries of individual bins are arbitrary, and not centred around the data. The result of this is that inevitably, in some instances, bin

boundaries bisect regions of high-break density (Figure 5.3), potentially masking regions of enrichment and rendering it unsuitable for my purposes.

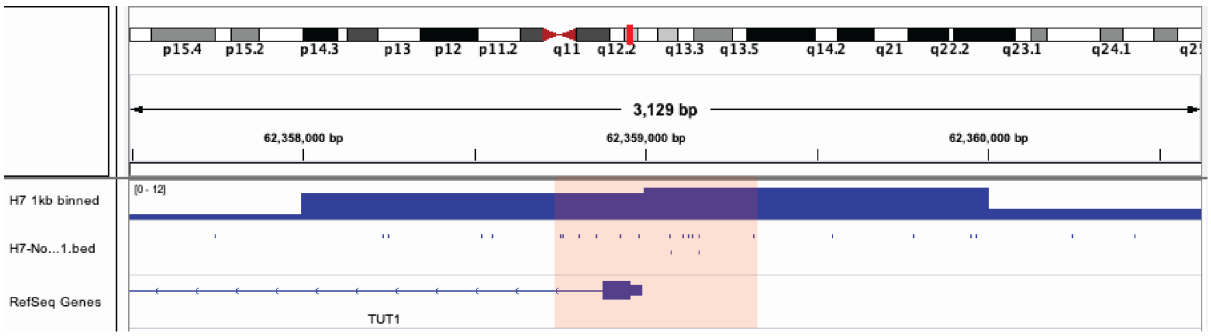


Figure 5.3 Bin boundary bisecting region of high-break density

Representative genome browser track of region of high-break density at TUT1 promoter (highlighted in red box) being split across two 1kb bins.

As binning draws arbitrary boundaries, not dictated by the data, I next sought an approach which is data-centric. To this end, I used the *merge* function of the *bedtools* suite to combine breaks within a fixed distance *d* of one another into one interval (Figure 5.4) (Quinlan and Hall, 2010).

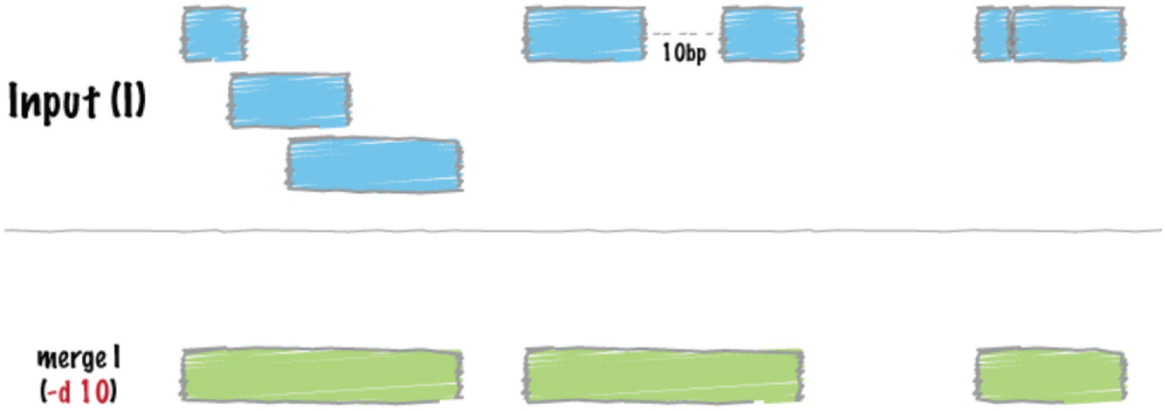


Figure 5.4 Bedtools merge schematic

Multiple regions in input data (blue) are merged into one larger interval (green) if the distance between them is less than 'd' (10bp in the above example). Figure adapted from bedtools.readthedocs.io (Quinlan and Hall, 2010)

As with the binned approach, the number of individual breaks which have contributed to the formation of the merged intervals is variable. However, in contrast to the binned approach, the length of individual intervals is also variable using the merge method (Figure 5.5 A). Identifying hotspots from the merged intervals is therefore less straightforward. Sub-setting the top 5% of intervals based on break count preferentially selects longer intervals, with often relatively low break densities (Figure 5.5 B). By contrast, sub-setting intervals based on their break density exclusively selects intervals composed of either unmerged breaks with intervals of

length 1bp, or two immediately neighbouring breaks with a length of 2bp. Each of these interval types have a de facto minimum break density of 1000 breaks/kb, which far exceeds that of any broader merged intervals, meaning wider regions of beak accumulation pass below the threshold (Figure 5.5 C).

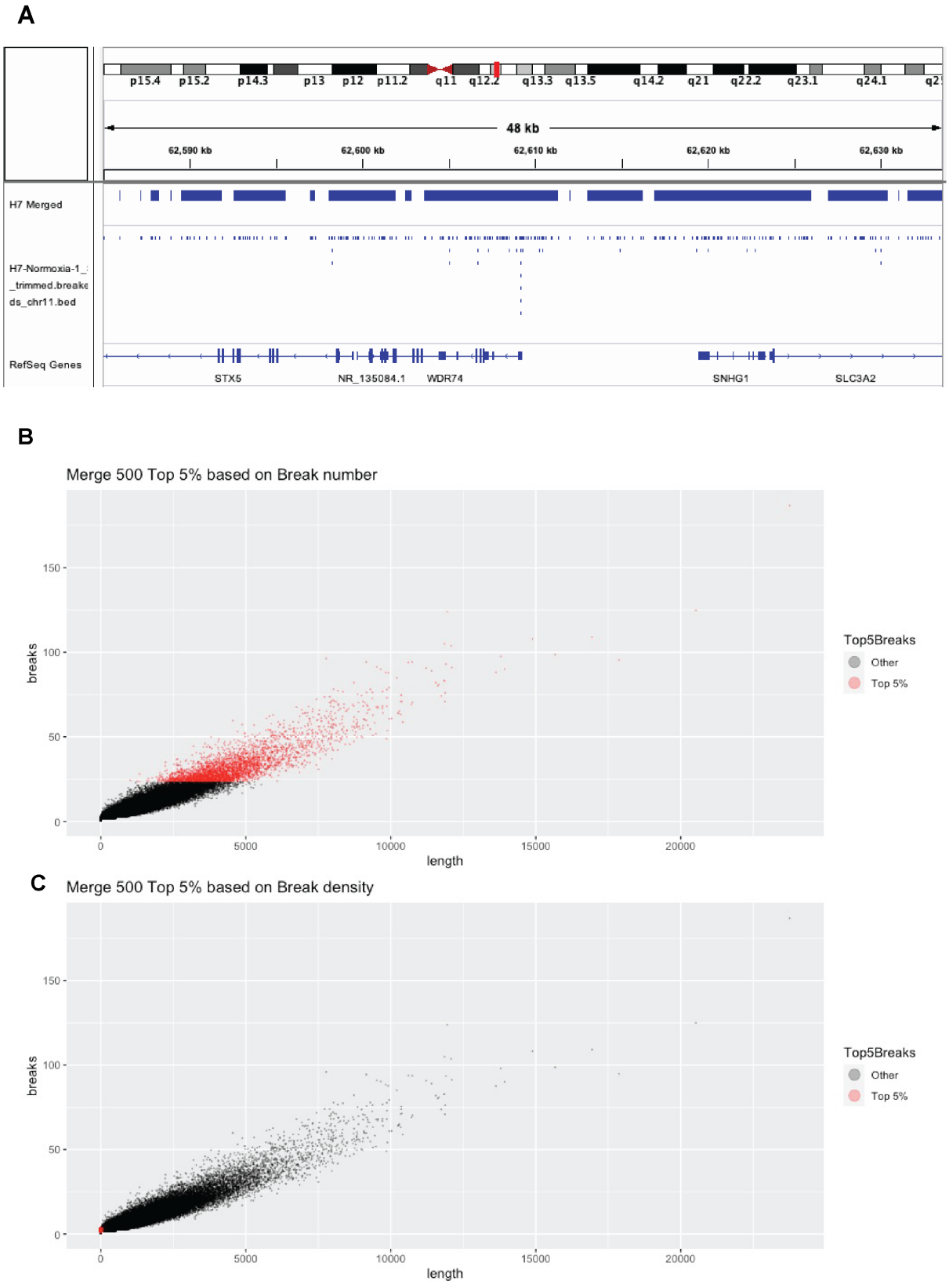


Figure 5.5 Issues in sub-setting hotspots from merged intervals

A) Representative genome-browser track of merged intervals with a 500bp merge distance, 'd' (upper) and individual breaks in H7 sample pluripotent dataset (below). B-C) Scatter plots of merged interval break count vs length of all merged intervals in H7 sample pluripotent dataset, top 5% highlighted in red based on B) break number, or C) break density.

To navigate the issue of break density selecting extremely short intervals, I conceived a transformation to compensate for extreme break density values in merged intervals of 1bp or 2bp lengths. *Bedtools merge* functions by scanning distance *d* upstream and downstream of individual breaks, and, if a second break is present within distance *d*, it is amalgamated into the merged interval. For each interval produced via the merge command, *Bedtools* has scanned, distance *d* upstream and downstream of the extremities of the interval (Figure 5.6 A). I therefore proposed normalising by the total distance *Bedtools* has scanned, for a given interval to calculate "Adjusted break density"

$$Adjusted\ break\ density = \frac{Breaks}{merged\ interval\ length\ + 2d}$$

Sub-setting the top 5% of intervals based on adjusted break density yields a wide distribution of different length intervals (Figure 5.6 B, C) and selects high-density regions whilst excluding, the extreme break density intervals observed in unmerged breaks (Figure 5.6 B, D).

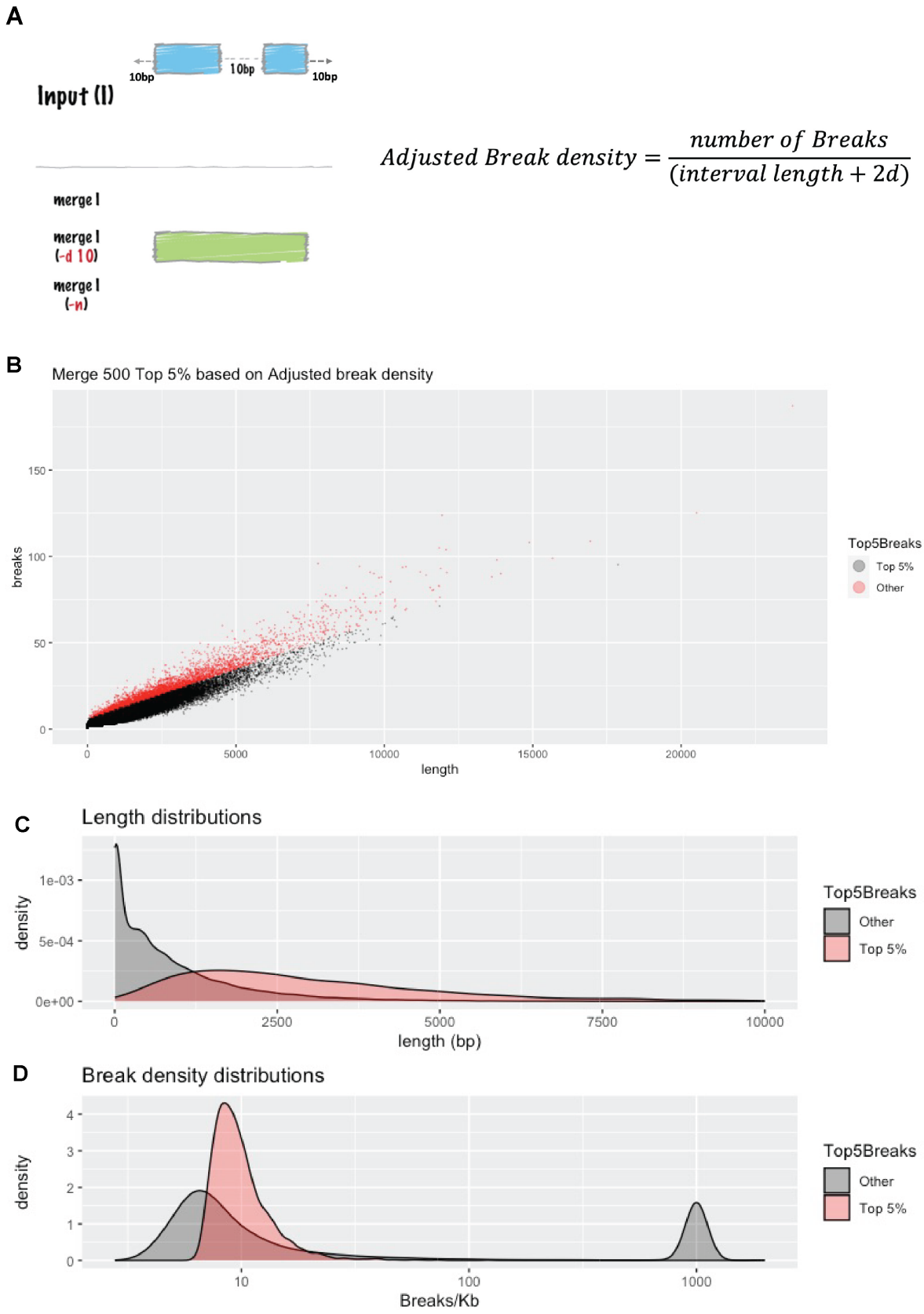


Figure 5.6 Adjusted break density for merged hotspot identification

A) Schematic of merge function, and rationale for adjusted break density calculation, adapted from bedtools.readthedocs.io (Quinlan and Hall, 2010). *For each DSB, bedtools merge scans length 'd' (10bp in the*

schematic) upstream and downstream of that DSB, hence for each interval generated by Bedtools merge, a total of the interval length + 2d has been scanned. B) Scatter plot of merged interval break count vs length of all merged intervals in sample H7 pluripotent dataset, top 5% based on adjusted break density highlighted in red. C,D) Density plots of interval length distributions (C) and break density distributions (D) in Top 5% of intervals (red), based on adjusted break density, or remaining intervals (grey).

A merge-based approach for identifying DSB hotspots, however, is dramatically affected by changes in DSB density between samples. In a sample with more DSBs sequenced, there is a higher likelihood of those DSBs being incorporated into a merged region. For example, the sample H7 pluripotent dataset used in the previous figures comprises ~10^7 DSBs, of which ~85% are incorporated into a merged interval of two or more DSBs. By contrast, the derivative differentiated dataset comprises ~10^6 DSBs. As a result of lower DSB density genome-wide, only ~25% of breaks in the differentiated sample are incorporated into merged intervals, when applying the same *d* value (Figure 5.7 A). This also effects the final size of merged intervals, with the low-break differentiated sample yielding typically far shorter intervals than the high-break pluripotent sample (Figure 5.7 B), meaning hotspots identified would not be readily comparable between the two conditions.

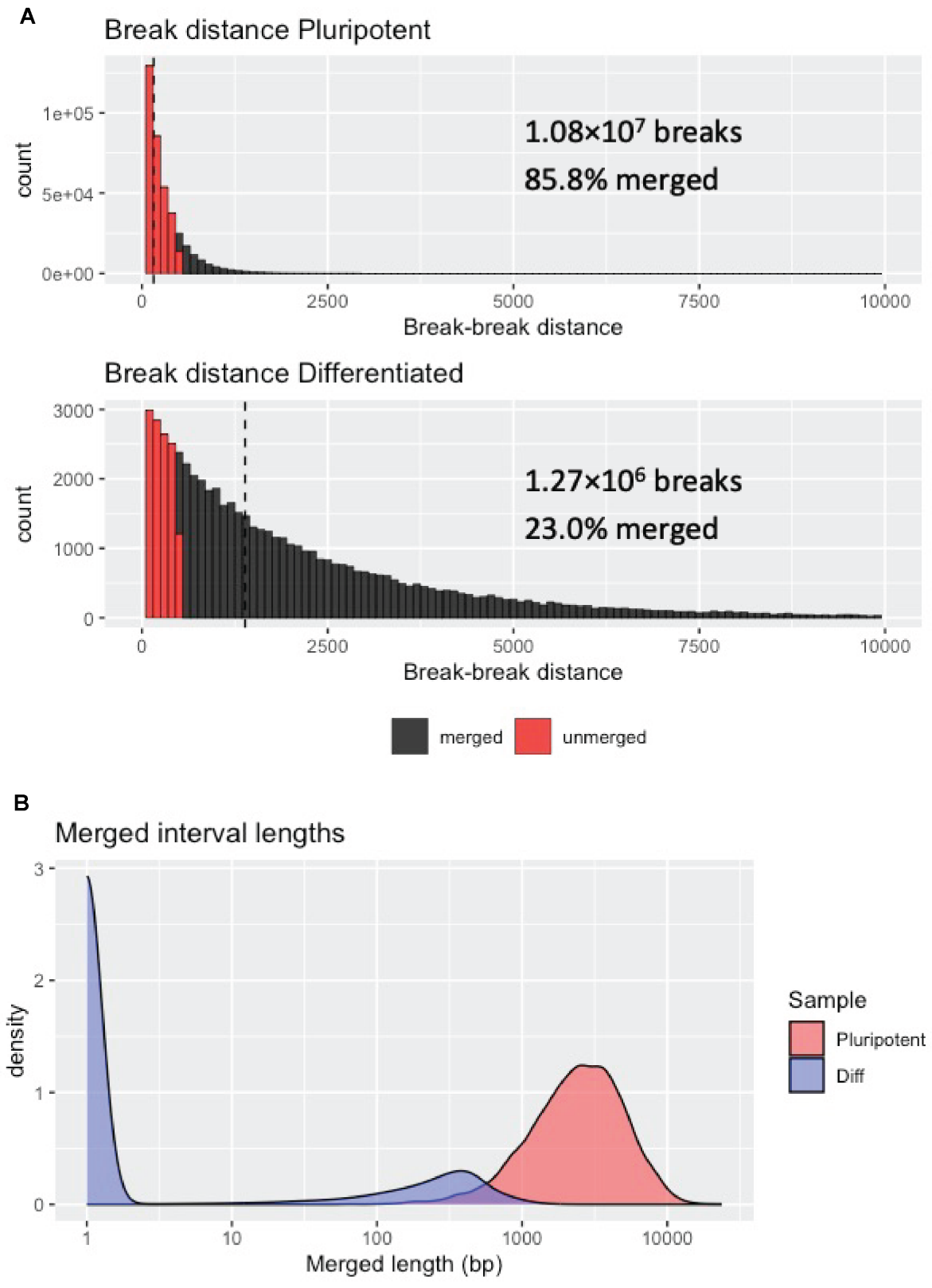


Figure 5.7 Total sample DSB count affects merge frequency and interval length

A) Histograms of inter-break distances in pluripotent and differentiated samples, distances below 'd' of 500bp (therefore merged) in red. Dashed lines represent median break-break distances B) Density plots of merged interval lengths from pluripotent (red) and differentiated (blue) samples.

Given the shortcomings of the binning method and the merge method, I developed a third strategy which uses sliding windows to largely overcome the limitations of the two previous approaches. In this third approach, I divide the genome into equal sized, overlapping windows of 1kb, offset by a smaller *slide* distance of 25bp (Figure 5.8 A). As with the binning method, DSBs are counted within each overlapping window (Figure 5.8 B) and, as all windows are of equal length, sub-setting hotspots simply entails thresholding windows on break count (Figure 5.8 C). By offsetting windows by a relatively short slide distance, the likelihood of bisecting genuine regions of break enrichment is reduced, compared to the sequential bins approach. Moreover, unlike the merge method, at the point of calculating break density and thresholding, windows are the same fixed size across all samples, and thus directly comparable between datasets with different break frequencies.

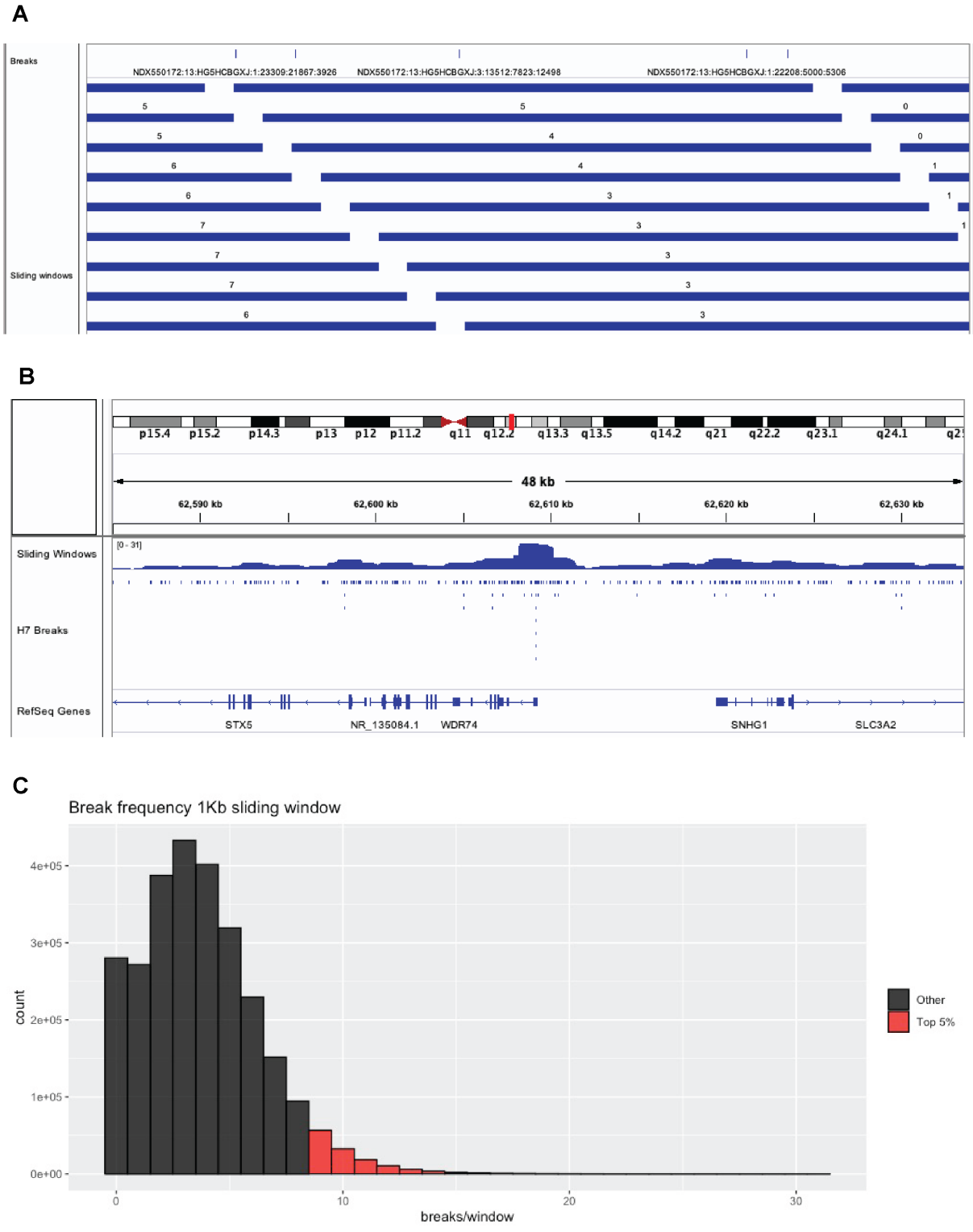


Figure 5.8 Sliding windows for hotspot identification.

A) Representative genome browser track of individual breaks (upper) and 1kb sliding windows offset by 25bp (lower) break counts in each window annotated. B) Representative genome browser track of break counts in each sliding window (upper, height represents break density) and individual breaks in sample H7 pluripotent dataset (lower). C) Histogram of break count distributions of sliding windows across chromosome 11 in a sample H7 pluripotent dataset. Top 5% of windows based on break count highlighted in red.

To compare the hotspots called by each of the three methods, I calculated overlap between the top 5% of intervals on chromosome 11 in an example pluripotent dataset, identified using each approach (Figure 5.9. A, B). Manual inspection of hotspots via genome browser shows similar regions are identified by all methods (Figure 5.9. A). Quantifying overlap of hotspots between methods reveals the vast majority (3450/3567) of hotspots identified by merging were also identified by sliding windows, whilst all 5247 hotspots identified via binning were also detected by the sliding window approach (Figure 5.9. B). This is unsurprising, as the sliding window approach is, in essence, a higher resolution version of sequential binning. Notably, the sliding windows approach calls more hotspots than either of the other two methods: Sliding windows are overlapping, offset by only 25bp, therefore, for a 1kb window length, there are 40-fold more intervals quantified than with equivalent sized sequential bins. Following thresholding, many of the sliding windows sub-set will be overlapping in cases of broad DSB enrichment, whilst in narrower regions of DSB enrichment, there may be one or very few overlapping sliding windows which pass the threshold. Post-thresholding, overlapping windows are amalgamated into a single hotspot, but given the difference in window numbers contributing to each hotspot, the end result is a larger number of final hotspots identified via sliding windows than in a sequential bin approach.

In summary, I have compared three different methods for unbiased identification of DSB hotspots from INDUCE-seq data. There is a high proportion of overlapping regions identified by the three methods. For analysis of my dataset, I selected a sliding window approach due to its increased resolution relative to sequential binning, and its insensitivity to total break count, relative to a merging approach.

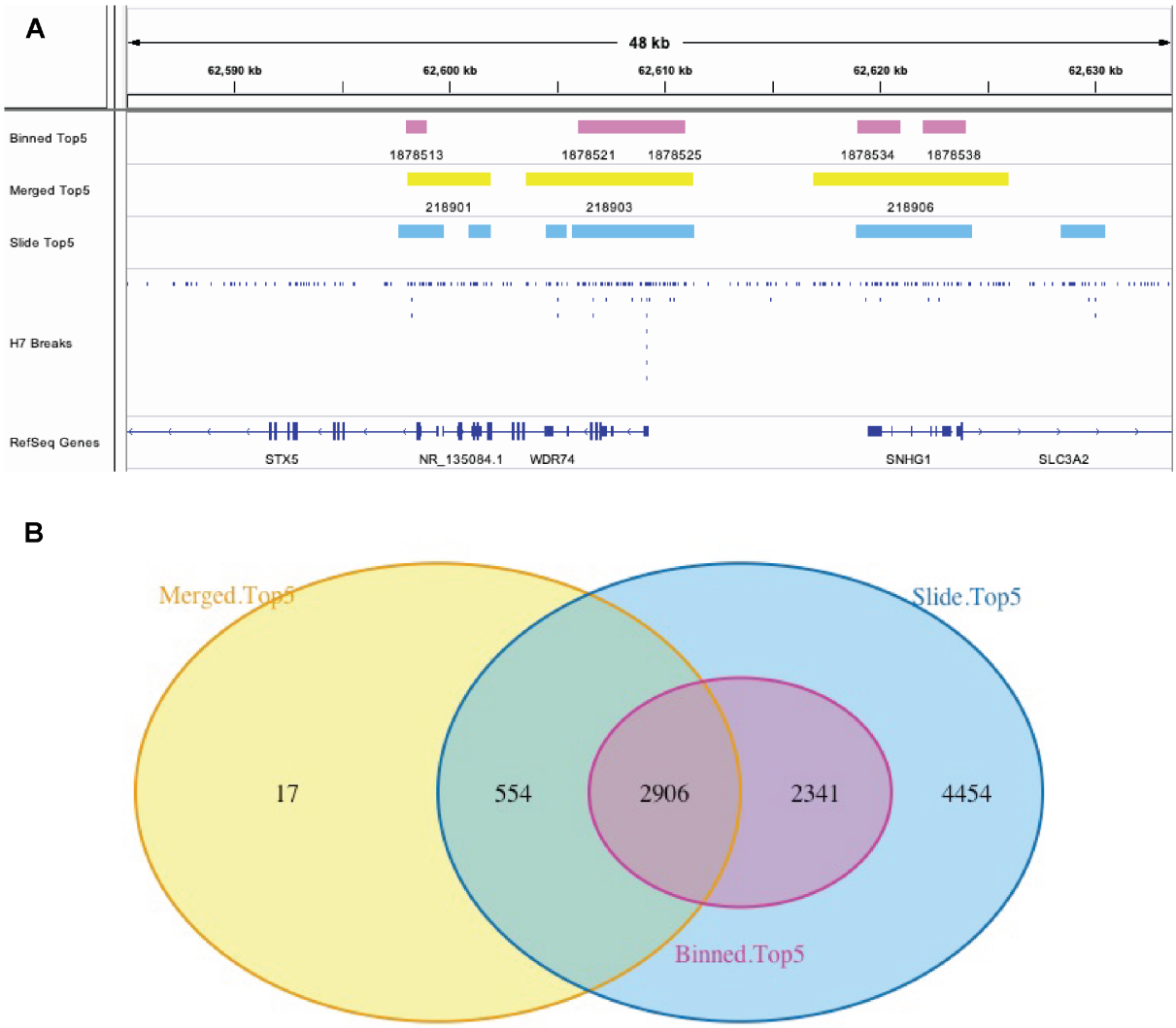


Figure 5.9 Comparison of hotspots identified using three methods

A) Representative genome browser track of hotspots identified after thresholding via sequential binning (pink), merging (yellow) and sliding windows (blue). Individual breaks in H7 sample dataset below. B) Venn diagram of hotspots identified by each method.

5.2.2. Optimising sliding window parameters for DSB hotspot identification

I selected a sliding window-based approach for the identification of DNA damage hotspots. I next sought to optimise the parameters of the sliding windows, starting with window size. I quantified breaks over variable window sizes from 25bp-1000bp, each with a fixed slide distance of 25bp, to visually examine the effect on data (Figure 5.10). Window size has a clear effect on the granularity of the data, with smaller window sizes identifying smaller, “sharp”

peaks, and larger window sizes identifying broader regions of enrichment which are less apparent using a smaller window size.

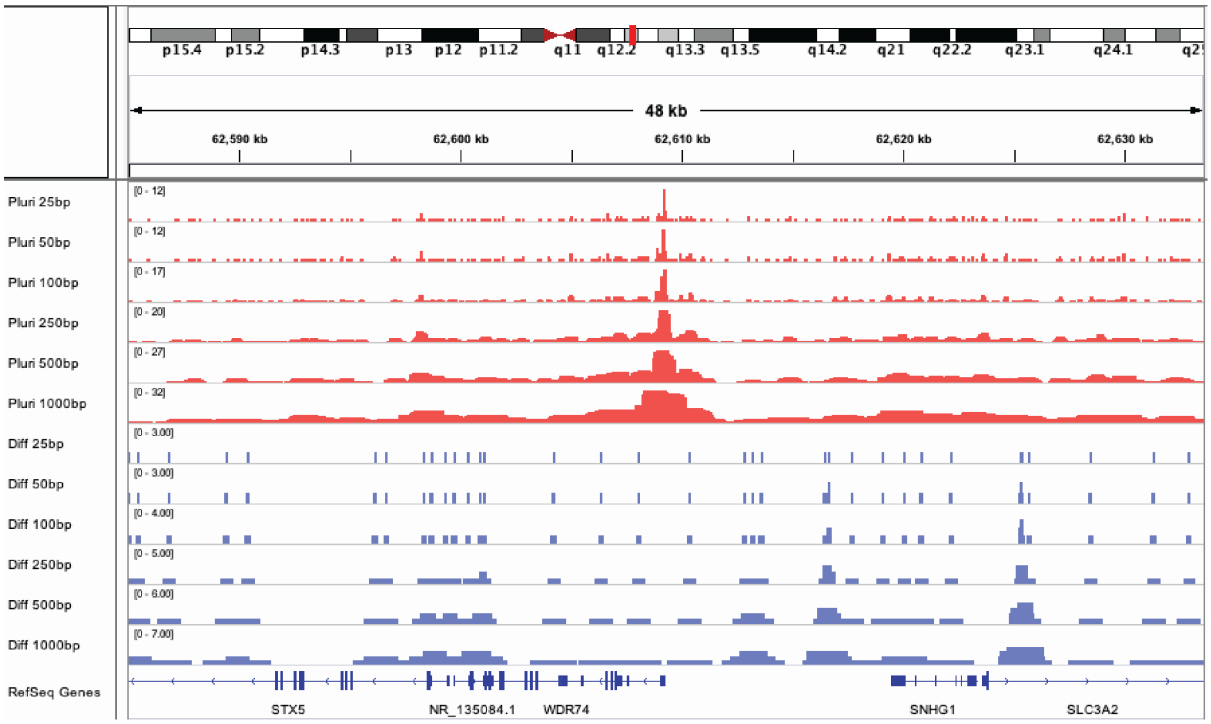


Figure 5.10 Window size affects granularity

Representative genome browser track of example pluripotent (red) and differentiated (blue) break data counted in sliding windows of increasing length (25bp-1000bp).

Based on the appearance of the data, window length is very likely to affect the final size of hotspots called. Importantly, the form of a DNA damage hotspot, in terms of length and break density, is unknown in hPSC or their differentiated derivatives. Previous break-mapping studies which have looked for DSB hotspots have relied on off the shelf methods of peak calling, designed specifically for ChIP-seq datasets. HTGTS-based break mapping studies in neural cell types identify very broad regions of recurrent break clusters, over megabase scale regions (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020) using SICER (Zang *et al.*, 2009). By contrast, direct break-labelling studies in various cell types report hotspots of ~200bp-1kb in length, using MACS2 (Zhang *et al.*, 2008). For both peak calling methods, the scale of the expected peaks or search window is specified by the user. None of the aforementioned studies provide a rationale for the scales used in peak calling methods and thus likely make presumptions as to the nature of their respective hotspots.

To avoid biasing results with presumptions on hotspot form, I aimed to empirically determine the optimal window size for hotspot identification in my datasets. I first looked at the effect of

window size on threshold pass rate in sample pluripotent and differentiated datasets, applying an arbitrary threshold of the top fifth percentile based on break count (Figure 5.11 A-C). In the pluripotent sample, threshold pass rate generally increases with window size, this effect is less pronounced in the differentiated sample (Figure 5.11 C). In both differentiated and pluripotent samples, the lowest window size of 25bp yields a threshold pass rate of far less than the arbitrary threshold of 5%. This poor threshold pass rate at lower window sizes is due to a low range in data on a discrete scale. That is: individual breaks are non-divisible and are therefore either present in a window or not. This yields an integer count by which to threshold. As such, applying a threshold to select the top 5% of breaks is based on an integer value for break count and therefore reflects a percentage less than 5%. Thus, window size directly affects the number of hotspots called in a given dataset. Specifically, smaller window sizes typically yield lower numbers of hotspots.

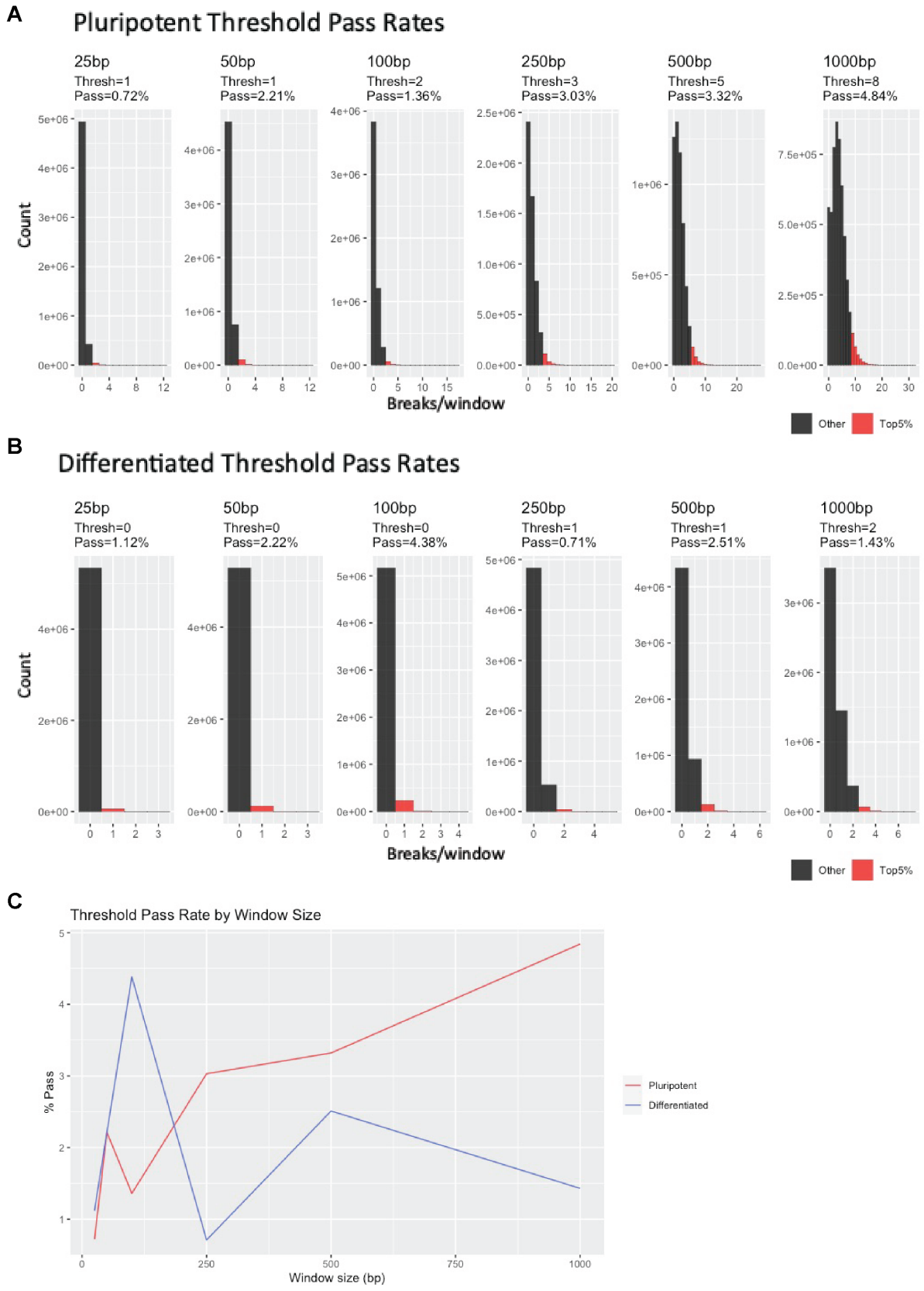


Figure 5.11 Window size affects threshold pass rate

Top: histograms of break frequencies per window of using variable window lengths and a 25bp slide distance in pluripotent (A) and differentiated (B) sample datasets. Windows passing the threshold (top 5%) highlighted in red,

absolute threshold values (break counts) and pass rates annotated above. C) Threshold pass rate by window size in pluripotent (red) and differentiated (blue) sample datasets.

Post-thresholding, overlapping sliding windows are combined into a final hotspot interval, resulting in hotspots of variable length (Figure 5.12 A). I plotted the distribution of hotspot lengths called using variable window sizes and observed a bimodal population in hotspots called at larger window sizes (250bp-1000bp), with a notable second peak at ~$2Windowsize$, which may suggest that, above 250bp, windows are sufficiently large to bridge two distinct hotspots (Figure 5.12 B). On closer inspection, the second peak in hotspot length is at $2Windowsize - d$ where *d* is the slide distance. As a uniform *d* of 25bp was used in this analysis, I reasoned that the lack of bimodal length distribution in lower window sizes may simply be due to lack of resolution. To test this hypothesis, I took windows of 100bp (the largest window size not to yield a bimodal distribution of lengths in the previous analysis) and offset by variable slide distances from 5-100bp. I then called hotspots as previously and plotted the resulting length distributions (Figure 5.13). I observed a bimodal distribution at slide distances below 10bp, with a prominent second peak at $2Windowsize - d$, suggesting bimodal length distribution is also detectable at smaller window sizes, provided the slide distance is low enough. Ultimately, these hotspots will be analysed via DiffBind software which trims hotspots to the region of maximal density across samples (Ross-Innes *et al.*, 2012), thus instances where hotspots are too long (i.e. the second peak in the bimodal distribution) will be truncated in downstream analysis.

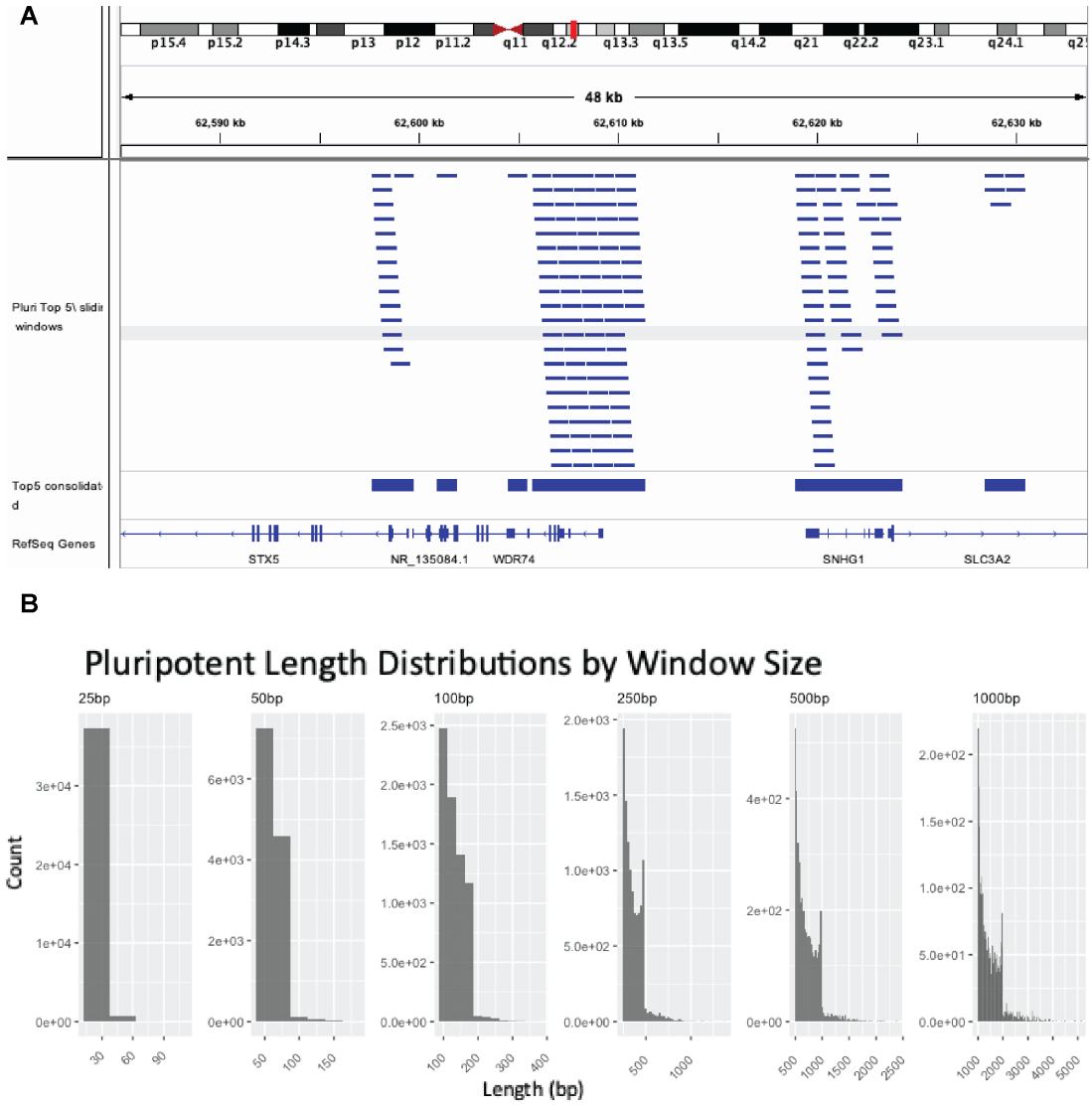


Figure 5.12 Window size affects hotspot length distribution.

A) Representative genome browser track illustrating overlapping sliding windows (upper) combined into hotspots (lower) of variable lengths following thresholding. B) Histograms of combined hotspot length distributions for different window sizes.

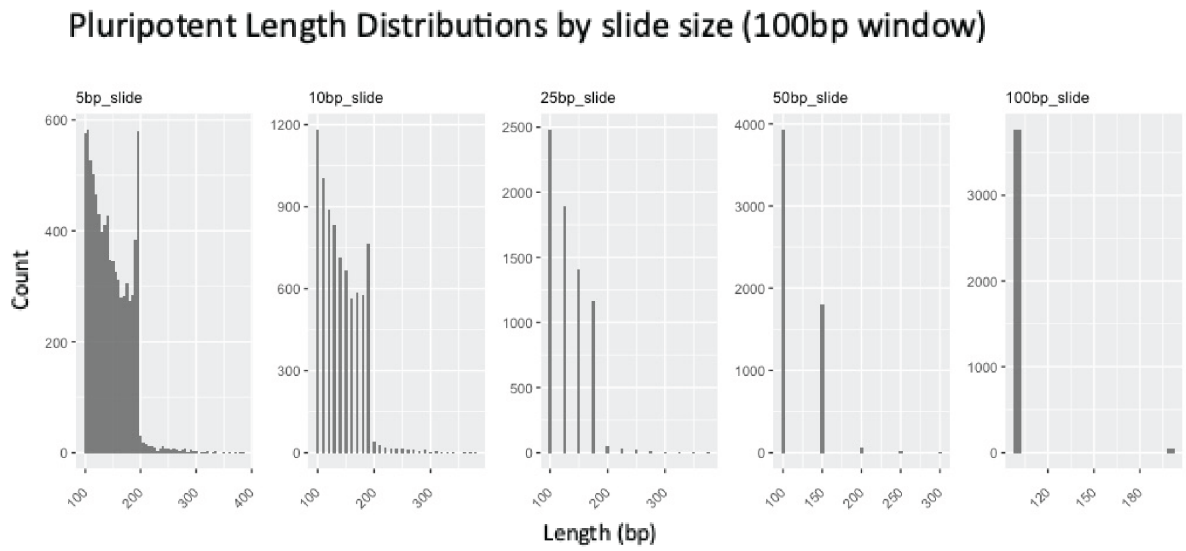


Figure 5.13 length distributions of hotspots called with a 100bp window and variable slide distances.

I next aimed to systematically determine the optimal window size and threshold level for hotspot identification in my data. To this end, I divided the reference genome into overlapping windows as previously but using a wider range of window sizes: 50bp-100Mb. A small slide distance is desirable as it gives greater resolution of hotspots. However, combining large window sizes with small slide distances becomes very computationally expensive. I therefore selected a slide distance of 10% of the total window size for each test window size in the range.

To determine suitability of each window size, I used a sample pluripotent dataset, alongside a randomised dataset whereby individual break coordinates were shuffled at random within chromosomes. I first determined the threshold break value for each window size at threshold levels of 1%, 5% and 25% in both real and randomised datasets. I then plotted the fold-change in real over random threshold value (Figure 5.14 A). All threshold levels appear to have two peaks of enrichment over random, one at a small window size <1kb and a second at ~100kb. This could indicate the presence of both narrow and broad hotspots in the data. Given the previous observed issue of poor threshold pass rate at smaller window sizes (Figure 5.11), I plotted threshold pass rate by window size for the real dataset (Figure 5.14 B). As previously observed, smaller windows typically have lower threshold pass rates. Threshold pass rate approached maximal at around a 10kb window size in all threshold levels tested, indicating that accurate thresholding based on percentiles requires a window size of 10kb or greater (Figure 5.14 B). I finally sought to determine the specificity of hotspot calling at each window size and threshold level i.e. the ability to call genuine regions of enrichment and not random fluctuations in break coverage. To achieve this, I calculated threshold break values in the real

dataset and applied that threshold to both the real and randomised break datasets. I then calculated the fold change in the number of hotspots called in the real dataset versus the random dataset as a readout of specificity (Figure 5.14 C). Broader windows called a greater number of hotspots over random up to 100kb-1Mb at all threshold levels. At 5% and 25% threshold levels, this decreased in windows greater than 1Mb, likely as a result of the window size approaching the chromosome length. Indeed, only chromosomes 1-14 are longer than 100Mb, the largest window size used here.

Based on these data I selected a threshold level of 1% for hotspot identification, which consistently had a higher fold change over random in threshold value than the less stringent threshold levels and had a greater fold change in hotspots called over random at all window sizes. I selected a window size of 100kb for hotspot identification. Whilst 100bp window size showed the greatest fold change in threshold value versus random, at a 1% threshold, this may simply reflect a very stringent effective threshold with a pass rate of ~0.3% (Figure 5.14 B). Moreover, 100bp window size showed relatively poor specificity at a 1% threshold level, calling less than five-fold more breaks in the real versus randomised dataset. I therefore selected a window size of 100kb, based on having: i) the second highest Real/Random threshold value ii) a near 1% threshold pass rate and iii) an infinite enrichment of hotspots called in the real versus random datasets. i.e., no windows in the random dataset passed the threshold calculated on the real dataset (Figure 5.14).

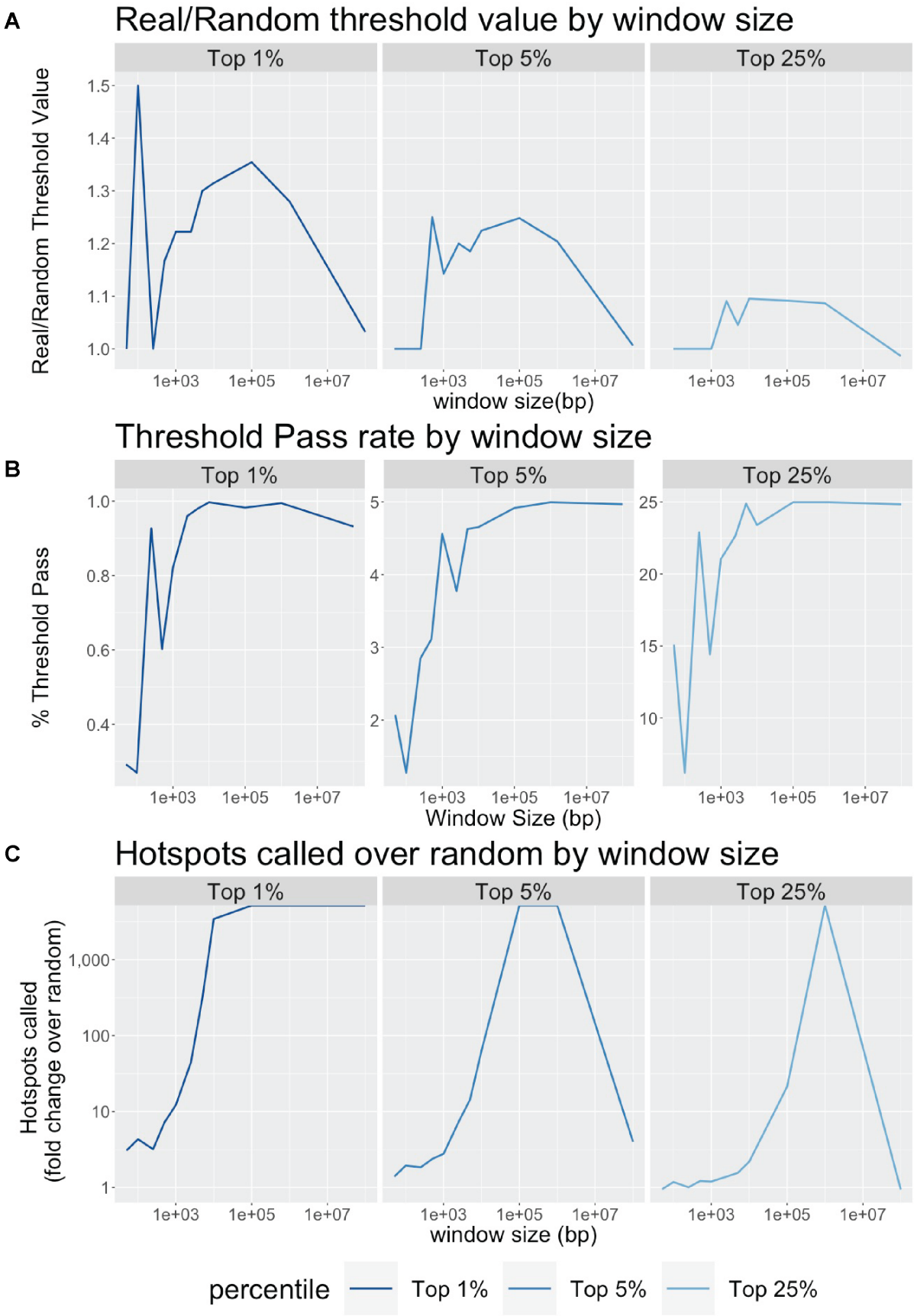


Figure 5.14 Window size optimisation

Real and randomised breaks were counted in sliding windows of sizes of 50bp-100mb each with a slide distance of 10% their length. Threshold levels of 1%, 5% and 25% were set. A) Real/Random break threshold value (breaks per window) by window size. B) Real threshold pass rate by window size. C) Number of hotspots called in Real/Random datasets by window size (thresholds calculated on real datasets).

I called genome-wide hotspots in all samples using a 1% threshold level with a 100kb window and 10kb slide. As an additional quality control step, to ensure that hotspots called were reproducible between replicates of a given sample, I randomised hotspot positions of each sample. I then calculated the number of overlaps in replicates of real hotspots and separately counted the number of overlaps in replicates of random hotspots to calculate fold change in overlap (Figure 5.15 A-B). All samples showed a greater number of overlaps in real versus randomised hotspots, however this effect seemed to be greater in pluripotent than differentiated samples (Figure 5.15 A). This could reflect more random breakage in differentiated cells, or alternatively could be due to a greater degree of heterogeneity amongst the differentiated cells. Interestingly, the fibroblast line CRL2429, which is terminally differentiated, showed the poorest enrichment of overlap amongst biological replicates versus randomised data, suggesting breakage may occur in a more stochastic manner in these cells.

In summary, I have developed, validated, and employed an unbiased pipeline for the identification of genomic DNA damage hotspots from INDUCE-seq data. I next interrogate differences in breakage between cell states.

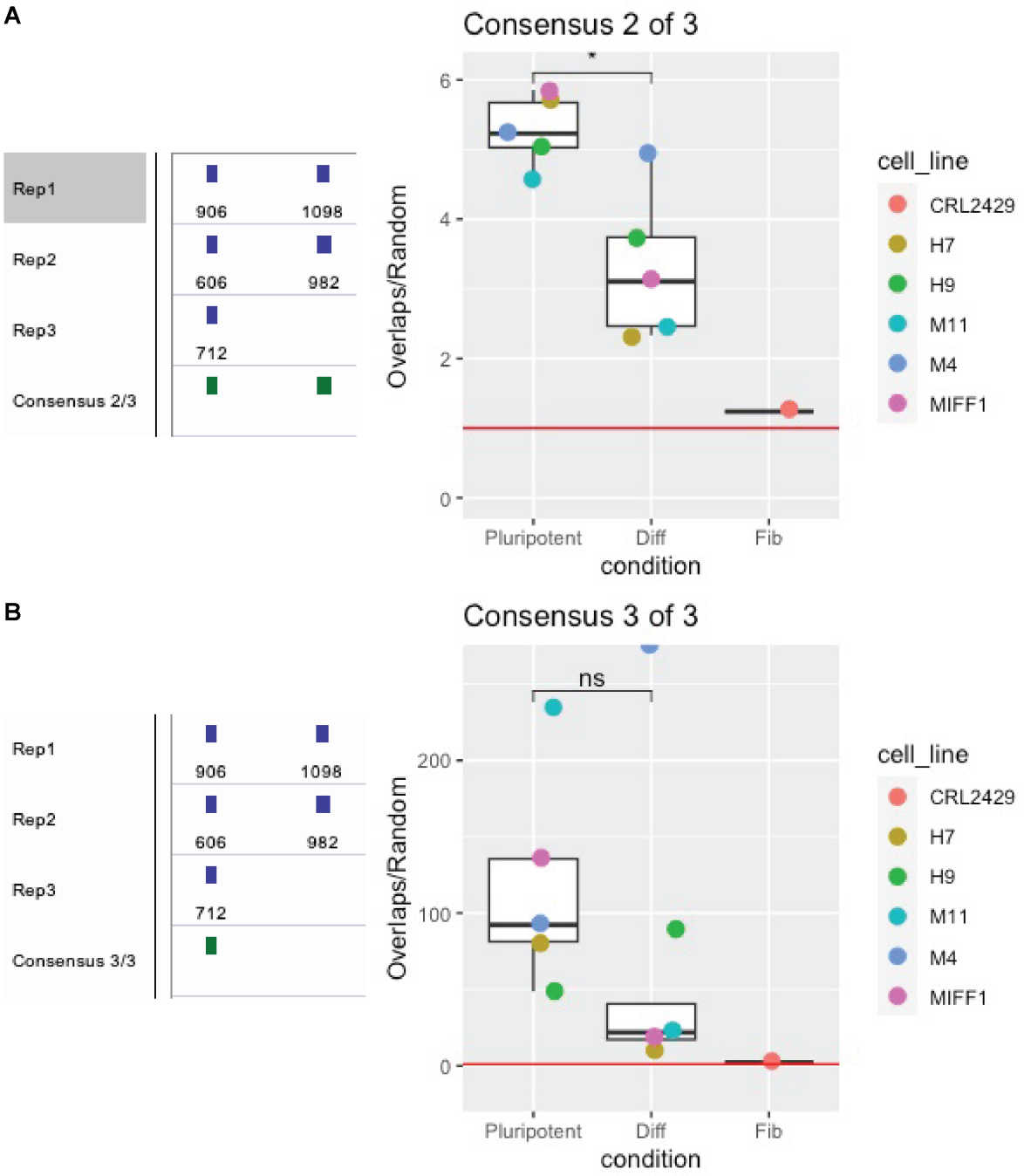


Figure 5.15 Hotspot QC: reproducibility between replicates

For each sample, hotspot overlaps between biological replicates were counted, and normalised against overlap in randomly shuffled hotspots. Left shows schematic representations of consensus (overlapping) regions in 2/3 (A) and 3/3 (B) replicates. Boxplots (right) show number of overlaps/random. Coloured points represent cell lines, y intercept (red) at 1 indicates no enrichment in overlap versus random. (Wilcoxon test, n=5 cell lines).

5.2.3. Identifying and annotating pluripotent-specific DSB hotspots

Having called hotspots in each sample, I next discerned regions of differential DSB enrichment between pluripotent and differentiated cell types to determine: i) which hotspots are unique to pluripotent cells, ii) whether these hotspots correspond to sites of recurrent genomic rearrangements and iii) the likely cause of damage at these sites.

For the purposes of differential DSB enrichment analysis in hotspots, I used DiffBind, an R package which modifies the DESeq2 pipeline (commonly used for differential gene expression analysis (Love *et al.*, 2014)) for differential enrichment of chromatin features (Ross-Innes *et al.*, 2012). Briefly, the software identifies a consensus set of hotspots, which is any hotspot present in two or more samples, trims the consensus hotspots to 100kb at the site of maximal mean breakage across all samples, and calculates differential enrichment of breaks in these regions between different conditions, i.e. pluripotent versus differentiated. Strikingly, when plotting a correlation heatmap of breakage in samples over this consensus hotspot set, samples cluster primarily based on condition (i.e. pluripotent or differentiated), and secondarily based on cell line, demonstrating the presence of a pluripotent DNA damage signature which is common across cell lines (Figure 5.16 A). I carried out differential enrichment analysis of the breakage in consensus hotspots between pluripotent and differentiated samples (Figure 5.16 A). Of the 1555 consensus hotspots analysed, 631 were significantly enriched for breaks in pluripotent samples (pluripotent-specific hotspots), whilst 596 were significantly enriched for breaks in differentiated samples (Figure 5.16 A).

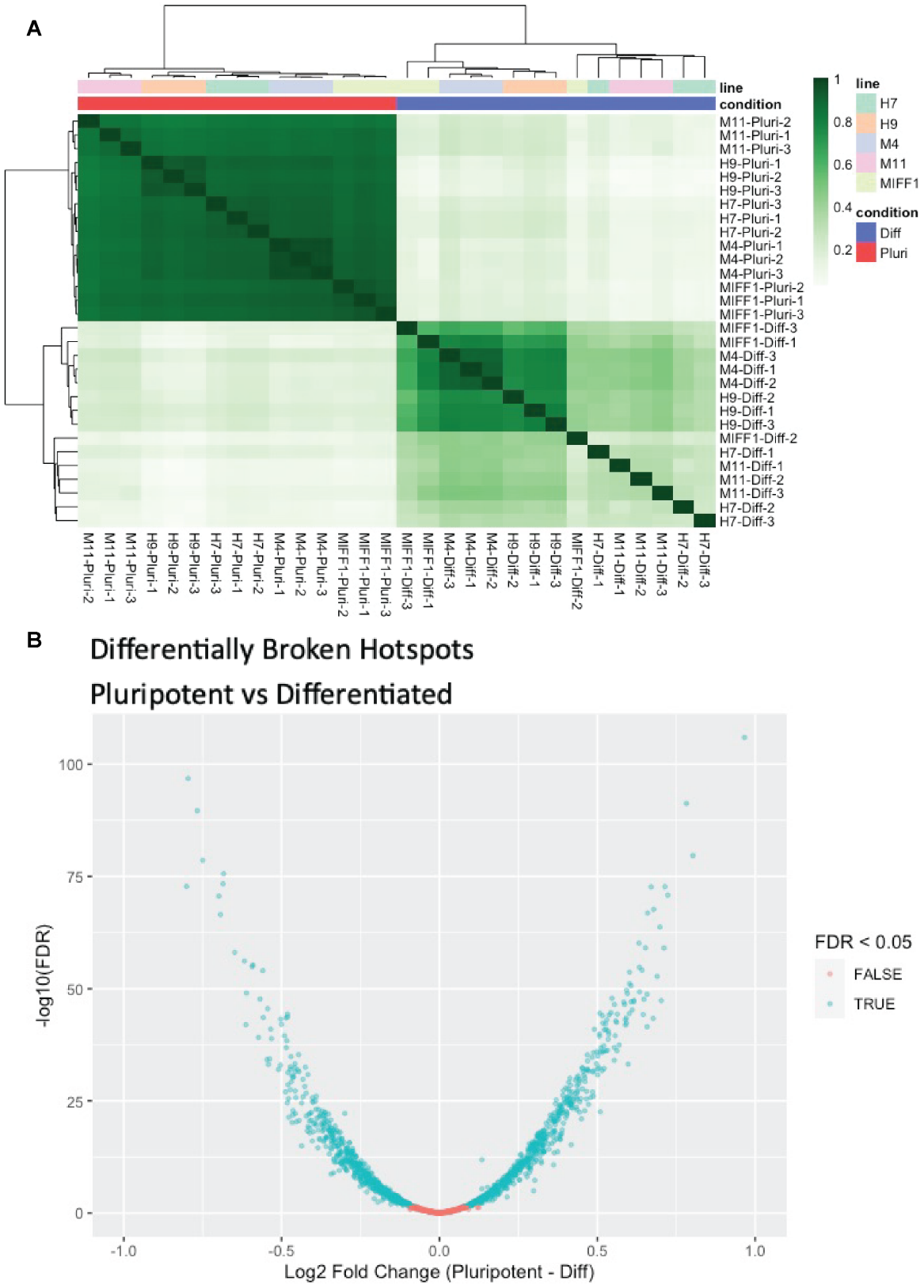


Figure 5.16 Differential Hotspot enrichment between pluripotent and differentiated cell types

A) Correlation heatmap of all pluripotent and differentiated samples, coloured by Pearson's correlation coefficient. Clusters calculated by Euclidean distance. B) Volcano plot of hotspot break enrichment in pluripotent versus differentiated cells. Significantly differentially enriched hotspots in blue (FDR<0.05).

At 100kb, the DSB hotspots identified in this study are large compared to studies attributing damage to the action of topoisomerase enzymes, where hotspot length was typically 200bp-1kb (Dellino *et al.*, 2019; Gothe *et al.*, 2019). The hotspots reported here are more comparable in size to the broad regions of break enrichment identified in neural progenitor cell types (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020). The studies in neural progenitor cell types find an enrichment of hotspots in the bodies of long actively transcribed genes, particularly under conditions of replication stress, proposed to be mediated by interference of transcription and replication (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020). Given I had previously noted a genome-wide association with transcriptional activity and DSB density in hPSC, I assayed what proportion of pluripotent specific DSB hotspots overlapped with long, actively transcribed genes. I subset the top quartile of genes based on length (>56.7kb) and expression in H9 cells (>28 TPM), and quantified overlap with pluripotent DSB hotspots (Figure 5.17 A). 207 of 631 hotspots overlapped with long, highly expressed genes, significantly more than expected (Fisher's test $P=6.21\times 10^{-63}$) (Figure 5.17 B). This significant enrichment of long actively transcribed genes in pluripotent-specific DSB hotspots, is suggestive of replication-stress driven breakage, mediated by interference between transcription and replication.

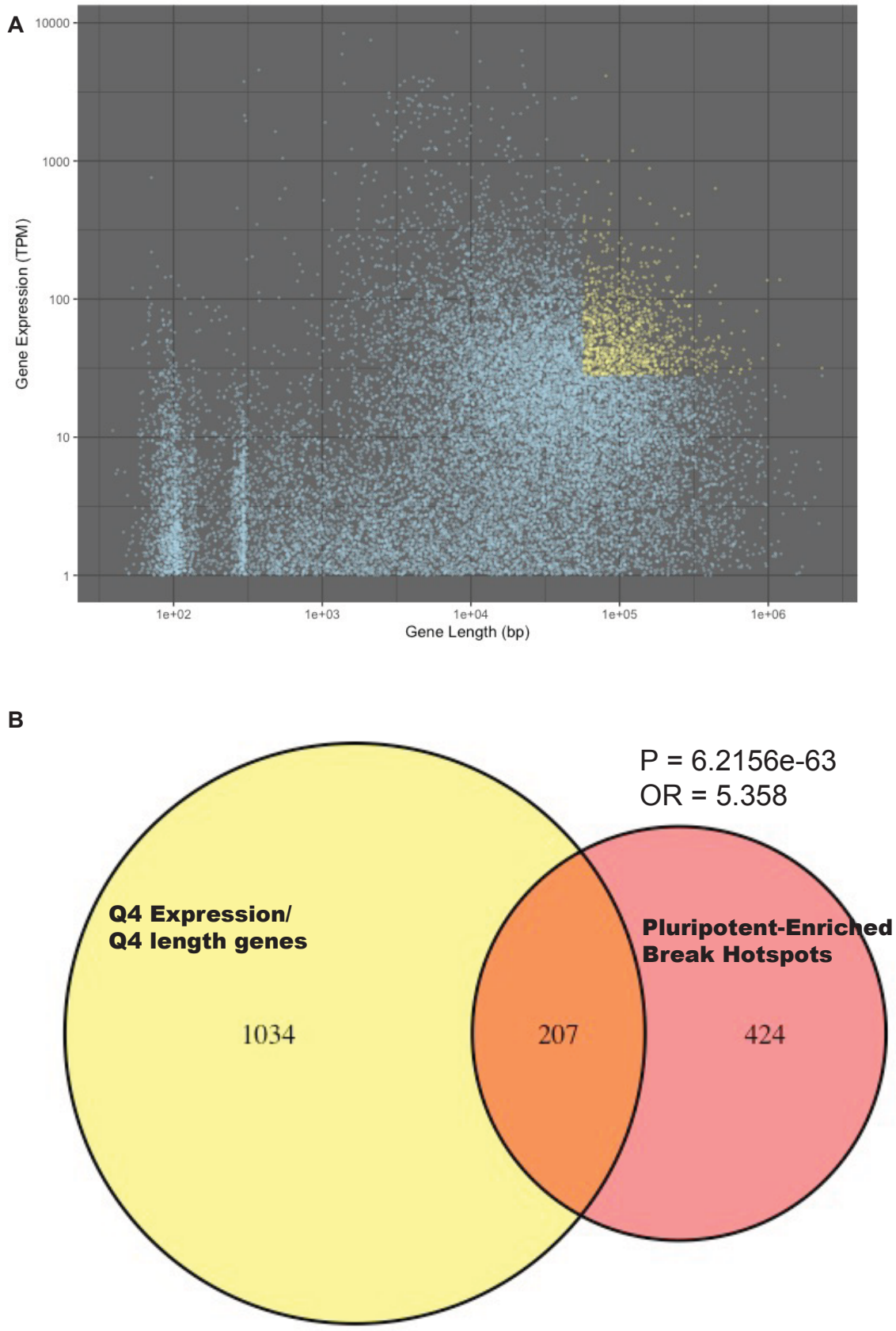


Figure 5.17 Pluripotent hotspot enrichment in long, actively transcribed genes

A) Gene length by Expression level (TPM) in H9 hESCs. Top quartile (Q4) genes based on expression and length highlighted in yellow. B) Q4 expression Q4 length genes and pluripotent-specific DNA damage hotspots. Fisher's test, OR= Odds Ratio and P-value annotated.

I next wanted to determine if any of the identified DSB hotspots corresponded to sites of recurrent chromosomal rearrangements in hPSC. I used a database of >4000 abnormal karyotype reports in hPSC lines, obtained from WiCell (www.wicell.org) (Stavish *et al., manuscript in preparation*), and ranked cytogenetic bands based on the number of times they have been documented as the break point of a translocation in hPSC, normalised to the total length of the cytogenetic band(s). I intersected pluripotent-specific hotspots with cytogenetic bands implicated in translocations and, for each hotspot lying in one of these bands, I plotted the mean hotspot break density against the normalised translocation frequency of cytogenetic bands, revealing no positive correlation between the two parameters (Figure 5.18 A). Pluripotent-specific hotspots have significantly higher break density in pluripotent samples than differentiated samples, however, other hotspots may still have a high pluripotent break density. I therefore plotted all hotspots' break densities in pluripotent cells against cytoband translocation frequency and again, observed no positive correlation (Figure 5.18 B). This is not necessarily unexpected: it is well documented that many of the chromosomal abnormalities recurrently observed in hPSC cultures endow cells with a fitness advantage, allowing them to outcompete wild type counterparts and hence rise above the threshold of detection (Olariu *et al.*, 2010; Avery *et al.*, 2013; Price *et al.*, 2021) i.e. it is known that variant cells are not selected for purely at the point of mutation or indeed DNA damage. Certain hotspots are however high break density and lie within a cytogenetic region recurrently implicated in chromosomal translocations in hPSC. For instance, a pluripotent-specific hotspot on chromosome 1q21.1-21.3, lies within the 7th most translocation-dense region and has a break density around two-fold higher than any other hotspot (Figure 5.18 B). Given this greatly increased break density and high translocation frequency at the 1q21 hotspot, I decided to further analyse this particular region with the aim of identifying and validating a specific cause of genome damage.

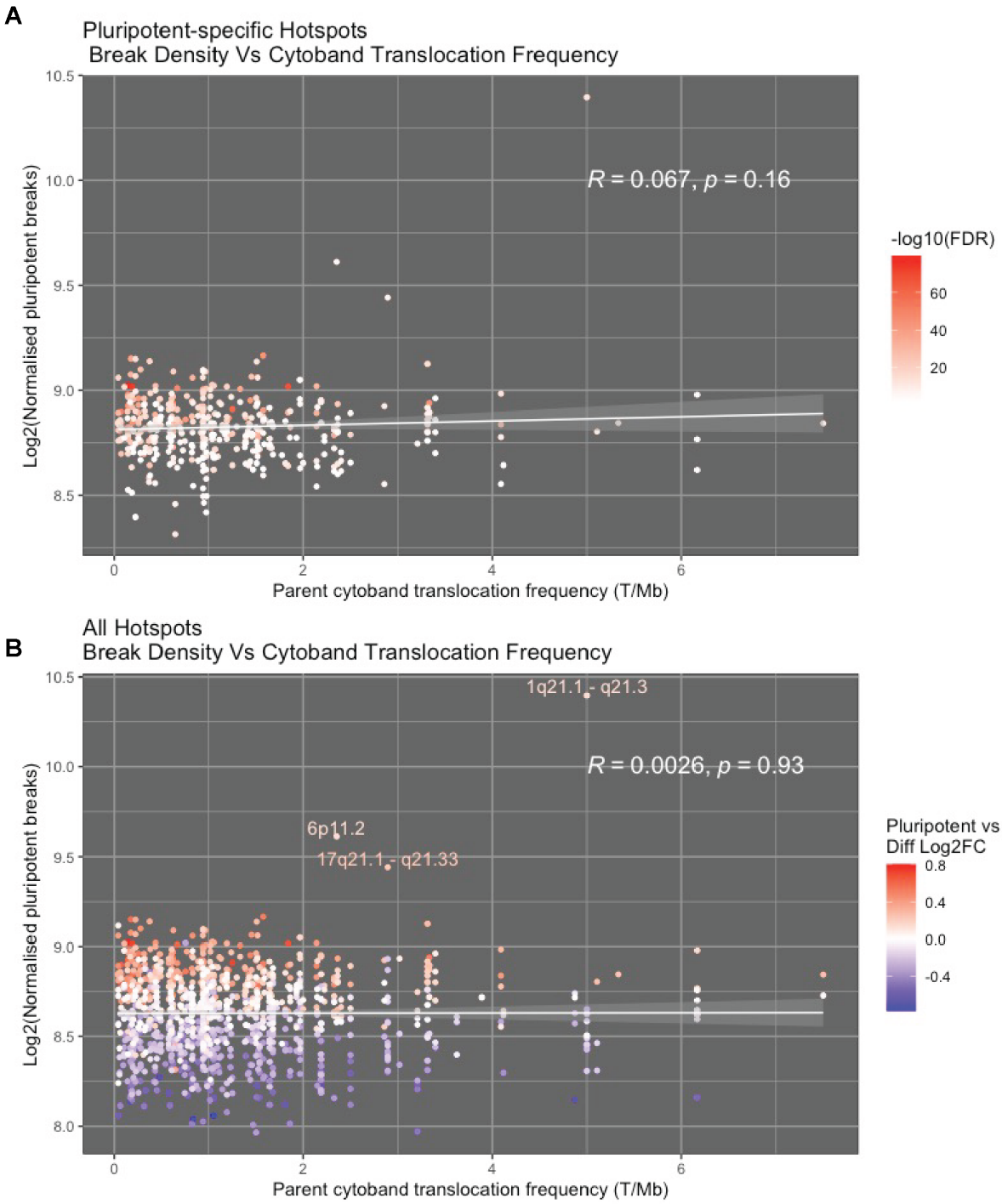


Figure 5.18 Hotspot Break frequency versus cytoband translocation frequency

Pluripotent specific (A) or all hotspots (B) were plotted for mean pluripotent break density (normalised to breaks per million per hotspot) against the frequency of translocations in its parent cytoband band, normalised for cytoband length. Pearson's correlation coefficient annotated.

The 1q21.1 hotspot lies within the common fragile site FRA1F (Figure 5.19). Common fragile sites are genomic regions which manifest as chromosomal breaks when cells are subjected to replication stress (Georgakilas *et al.*, 2014). CFS are classically defined cytogenetically and

are therefore coarsely mapped, indeed FRA1F is defined as the entirety of 1q21 (12.4Mb) (Kumar *et al.*, 2019). The 1q21 hotspot, by contrast at 100kb, makes up a very small fraction of FRA1F, but may represent a more precise mapping of the FRA1F site in hPSC. The 1q21 hotspot overlaps with several transcripts, but notably *PDE4DIP* is Q4 based on length and Q4 based on expression levels in H9 cells, suggesting transcription associated damage by interference with DNA replication. Interestingly, the 1q21 hotspot extends ~20kb upstream of the TSS of *PDE4DIP*, possibly suggesting transcription-mediated replication stress is not a direct result of transcription replication collisions. Le Tallec and colleagues fine-mapped CFS in erythroblasts and found that, whilst CFS largely overlap with long genes, breakage was often extended beyond the gene bodies, suggesting transcription replication collisions were not responsible (LeTallec *et al.*, 2013). Active transcription of DNA has been shown to reposition (Gros *et al.*, 2015b) or even inactivate genic licensed origins prior to firing (Macheret and Halazonetis, 2018). An alternative model for such transcription-dependent breakage, is that transcription over long genes displaces intragenic replication origins, therefore replication of long transcriptional units is dependent on forks travelling large distances from origins flanking the gene body (Ji *et al.*, 2022) . The lack of origins renders such regions highly sensitive to replication stress, as stalled forks are not readily recovered by convergence with neighbouring forks, increasing the likelihood of fork cleavage and DSB formation (Brison *et al.*, 2019).

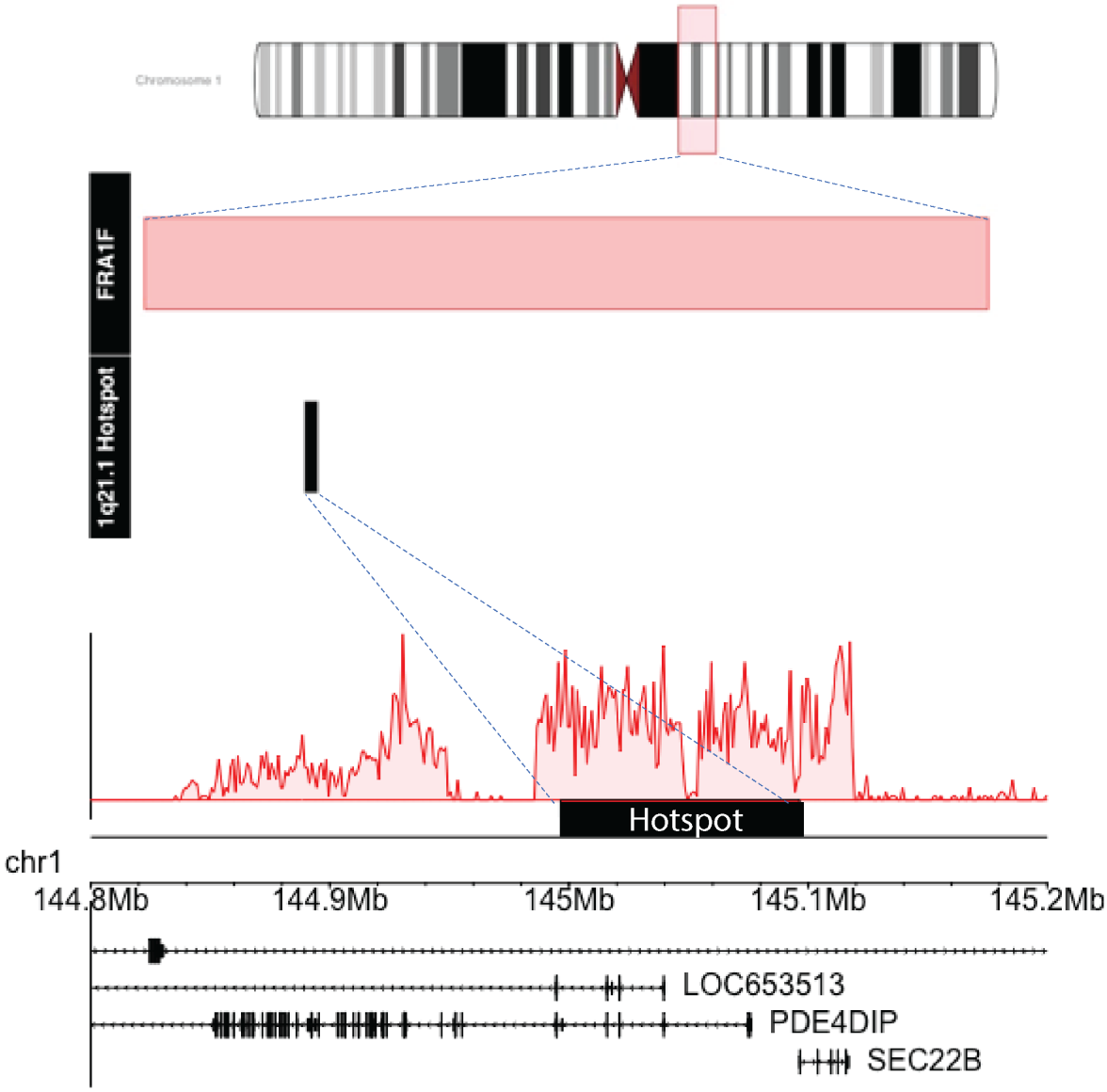


Figure 5.19 1q21 hotspot

Top panel: Chromosome 1 schematic with fragile site FRA1F highlighted in red. Middle panel: Zoom in on FRA1F and 1q21 hotspot. Bottom panel: Zoom in on 1q21 hotspot (black box); red density plot represents break coverage in example H9 dataset, with genes annotated beneath.

The 1q21 hotspot lies within a common fragile site and partly overlaps with a long highly expressed gene. Moreover the hotpot is significantly enriched for breaks in pluripotent cells, which are known to suffer high basal levels of replication stress (Halliwell *et al.*, 2020). I hypothesized replication stress drives breakage in this region, yielding structural variants in hPSC, and next sought to validate this experimentally.

5.2.4. Experimental validation the cause of structural variation at chromosome 1q in hPSC

To determine whether replication stress drives breakage and ultimately formation of structural variants at the 1q21 hotspot in hPSC, I aimed to modulate levels of replication stress and measure the effect on mutation rate. Previous work from our lab identified that hPSC suffer constitutive replication stress which can be partly rescued by the addition of exogenous nucleosides, proposing a model whereby hPSC have low dNTP pools due to a rapid cell cycle, which drives replication stress (Halliwell *et al.*, 2020). I therefore devised an experiment where replication stress would be either alleviated by nucleoside supplementation, or exacerbated by treatment with the DNA polymerase inhibitor aphidicolin (APH) (Baranovskiy *et al.*, 2014), as well as a control condition where hPSC are cultured in standard pluripotent culture medium. Following treatments, I would quantify mutation rate to determine the effect of replication stress.

Exogenous nucleoside concentration has been previously optimised for hPSC culture (Halliwell *et al.*, 2020), however APH concentration has not. I therefore first titrated APH for a 24-hour treatment of hPSC and took a readout of cell numbers, cell cycle phase occupancy and S/G2 phase γh2AX foci as a measure of replication-stress induced DNA damage (Figure 5.20 A-D). Concentrations of 400nM APH and above yielded reduced cell numbers relative to the untreated condition following both 24 hours treatment and 24 hours recovery, which is likely caused by a decrease in proliferation as well as an increase in levels of apoptosis (Figure 5.20 A). Importantly, at 400nM APH, there were greater numbers of cells following 24 hours recovery, than in the same condition following 24 hours treatment, indicating that a population of cells continue to proliferate (Figure 5.20 A). Following 24 hours APH treatment, there was a dose-dependent increase in γh2AX foci per S/G2 phase nucleus, which was significantly higher than the control condition in concentrations greater than 12.5nM (Figure 5.20 B-C). γh2AX foci counts were largely comparable across conditions following 24 hours recovery, indicating that DNA damage was repaired in surviving cells (Figure 5.20 B-C). Approximation of cell cycle phase occupancy from DAPI intensity showed 400nM APH-treated cells appear to accumulate in S-phase following 24 hours treatment indicating replication stress, and subsequently are enriched in G1 following 24 hours recovery, suggesting cells have progressed through the cell cycle (Figure 5.20 D). As 400nM APH induced a high number of γh2AX foci, and was permissive of continued proliferation following recovery, I selected this concentration for induction of replication stress in hPSC.

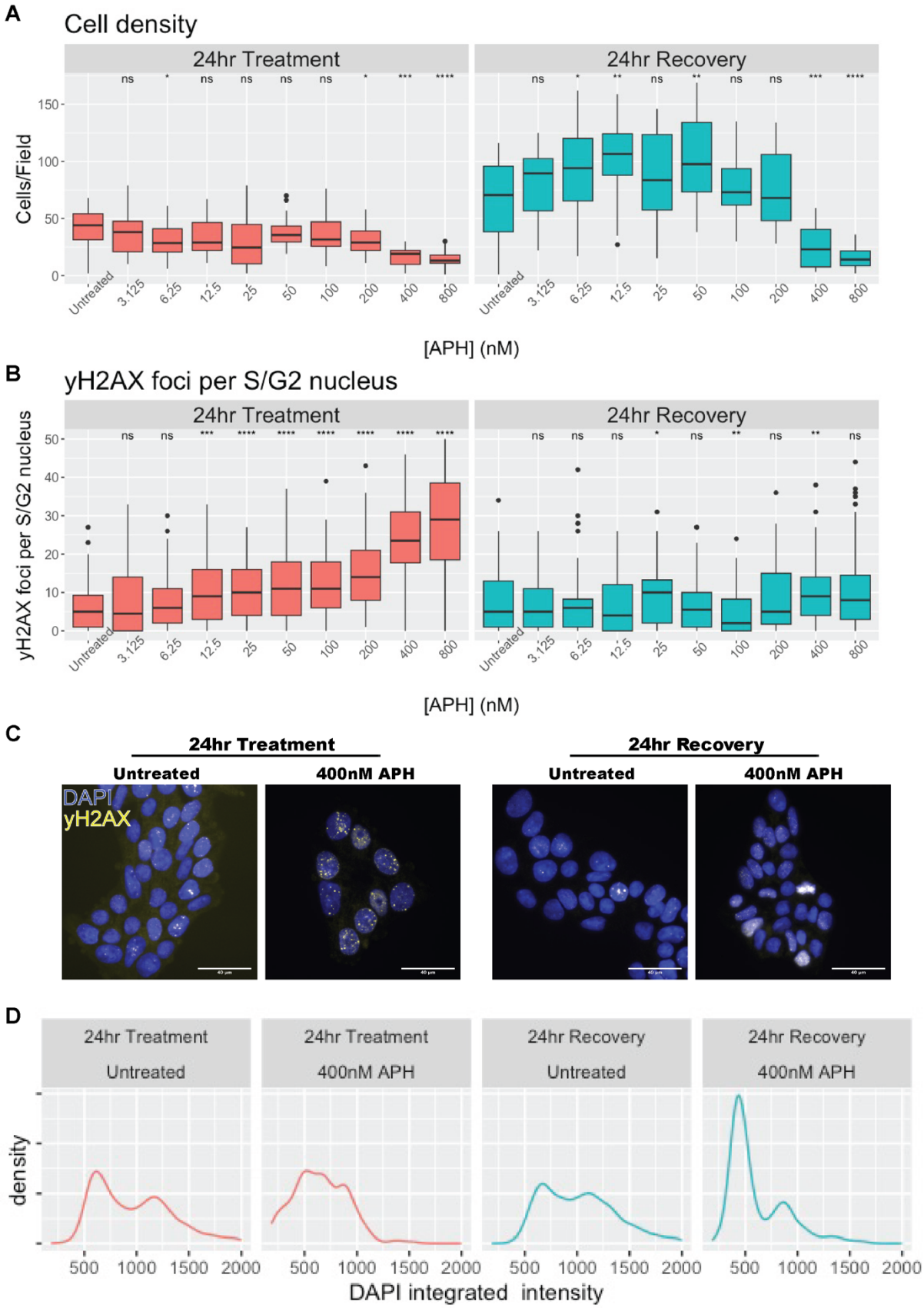


Figure 5.20 APH titration for replication stress induction

Cells were treated with the indicated concentrations of APH for 24 hours, then allowed to recover in the absence of APH for a further 24 hours.

A) Cell numbers per image in each condition (n=30 images). B) γh2AX foci per S/G2 nucleus (n=115 cells per condition) Boxes represent interquartile range, mid-line represents median value. Dots represent outliers. C) Representative γh2AX immunofluorescence images of untreated cells and 400nM APH-treated cells. D) Density plots of DAPI integrated intensity per nucleus as a means of approximating cell cycle distributions, in untreated and 400nM APH treated cells. All data from one experiment only.

To quantify the effect of replication stress on mutation rate in hPSC, I cultured H9 hESC under standard pluripotent conditions as a control, as well as supplementing with exogenous nucleosides for 24 hours (to alleviate replication stress) or treating with 400nM APH for 24 hours (to exacerbate replication stress). Cells were then allowed to recover in standard pluripotent conditions for 24 hours only, to allow repair of DNA damage, without providing opportunity for significant clonal expansion of any newly induced variant cells. Following recovery, cells were single-cell sorted and cloned in 96 well plates. Using a cloning approach eliminates the effects of cell competition within a heterogenous population of cells (Price *et al.*, 2021), as well as simplifying detection of variants which is binary in a clonal population, i.e. a clone is either variant or wild type. I briefly expanded clones for approximately two weeks, following which clones were harvested for DNA extraction and qPCR detection of copy number alterations of chromosome 1q (Figure 5.21 A). I ultimately aimed to determine the frequency of structural variants, specifically mapping to the 1q21 hotspot, however no method for structural variant detection is readily scalable to hundreds of clones. Thus, as structural variants involving the chromosome 1q in hPSC are overwhelmingly associated with a gain of copy (Baker *et al.*, 2016), I reasoned copy number change would serve as a useful, scalable, initial screen of clones.

I screened ~480 clones, of which 423 passed post-amplification quality control (based on Cq values and standard deviation in triplicate wells). Following single cell deposition, any copy number change should be present in 100% of cells of a given clone. I therefore set a threshold of 2.5 for calculated chromosome 1q copy number, above which clones are designated as variant. Of 127 clones from the untreated condition, there were no apparent variants. By contrast there were 2 variants of 125 clones in the nucleoside-treated condition (1.6%), and 6 of 171 in clones of the APH treated condition (3.5%) (Figure 5.21 B). The rarity of spontaneous mutation in hPSC is consistent with previous studies on mutation rate in our lab (Thompson *et al.*, 2020). Whilst APH-treated cells harboured an apparent increase in variant frequency compared with the untreated control cells, the scarcity of variants means the experiment lacks statistical power at this scale. Nonetheless, the frequency of variants in APH-treated cells may be indicative of replication-stress driven mutation at 1q21. Low-pass whole genome sequencing is currently in progress to identify the specific breakpoints of CNVs in clones of interest in both nucleoside and APH treated conditions.

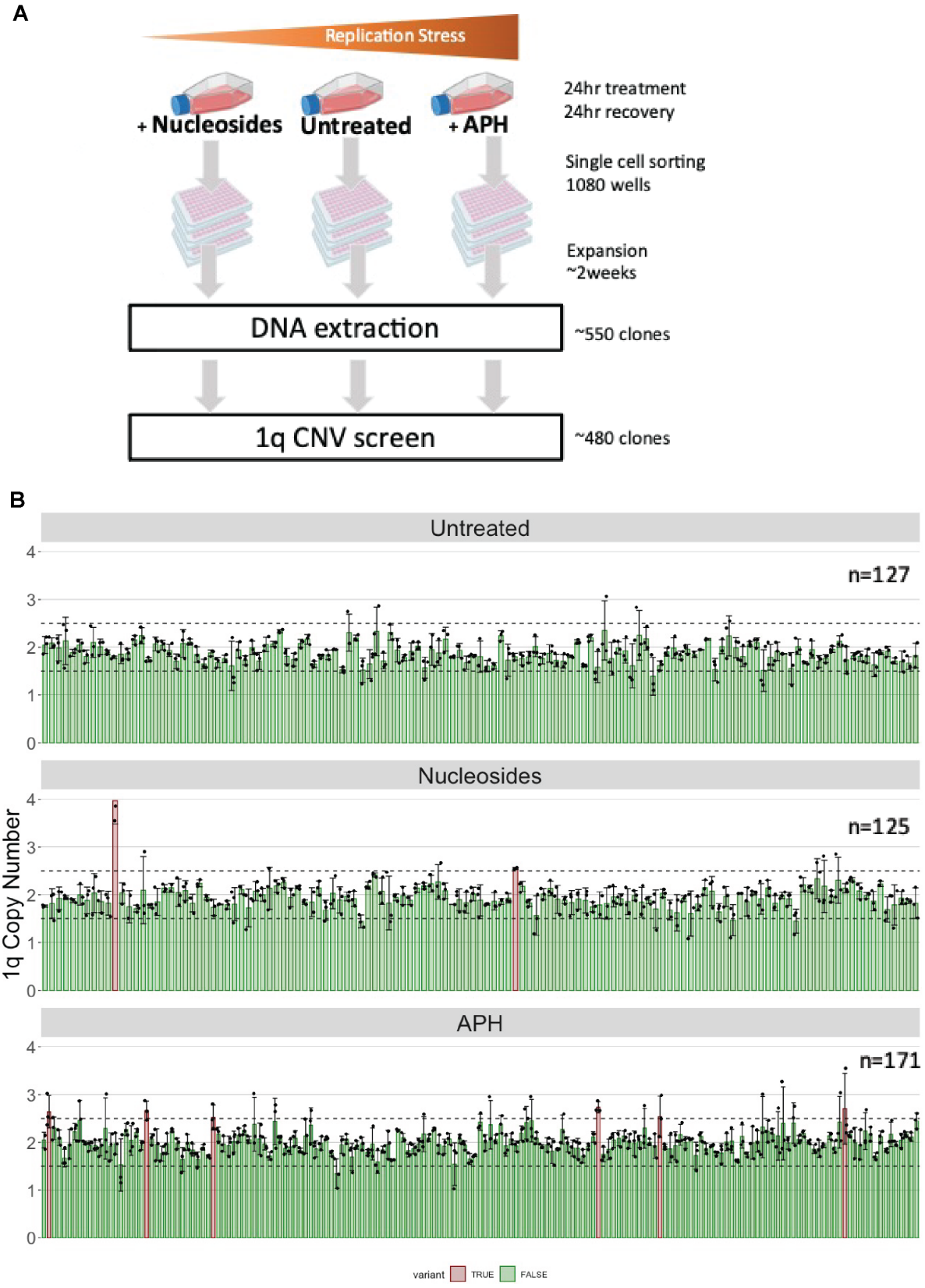


Figure 5.21 CNV frequency in replication stress-modulated clones

A) Schematic of mutation rate experiment. Cells were subject to the indicated conditions for 24 hours, then allowed to recover for 24 hours. Single cells were then deposited into wells of 96 well plates via FACS to generate clonal populations. Clones were expanded in culture for 2 weeks, then harvested for DNA extraction and qPCR screening of MDM4 (1q) copy number. B) Mean calculated MDM4 (1q) copy numbers of clones of each condition. Each bar represents an individual clone, error bars +/- SD, black points indicate calculated copy numbers in individual triplicate wells (n=3 reactions).

5.3. Discussion

In this chapter I have developed an unbiased pipeline for the identification of DSB hotspots in pluripotent and differentiated cells. I have used this pipeline on all samples and carried out differential enrichment analysis to determine regions specifically enriched for breaks in pluripotent cells. Pluripotent-specific hotspots were found to overlap with long highly expressed genes, suggestive of transcription-mediated replication stress. For experimental validation, I focussed on the 1q21 hotspot which exhibits ~two-fold higher break density than any other hotspot genome wide and lies within a recurrent breakpoint for chromosomal translocations in pluripotent cells. By modulating levels of replication stress, followed by single-cell cloning and genetic screening, I found preliminary evidence to suggest that replication stress drives copy number changes on chromosome 1q, suggesting culture conditions directly contribute to the genetic stability and ultimate safety of hPSC.

5.3.1. Unbiased identification of hotspots in pluripotent and differentiated cells

Previous DSB mapping studies have identified hotspots using peak calling algorithms designed for ChIP-seq experiments. Such peak calling algorithms are designed specifically for high-coverage overlapping sequencing reads, typically with high background, in which peaks correspond to regions of known or assumed sizes such as narrow transcription factor binding sites or broad domains of modified histones (Thomas *et al.*, 2017). As modern DSB mapping techniques, such as HTGTS or direct break-labelling, map DSBs to nucleotide or near-nucleotide resolution, DSB mapping data, unlike ChIP-seq data, is comprised of rare, low noise, largely non-overlapping intervals of 1bp in length and hence sub optimal for use with ChIP-seq peak callers (Bouwman and Crosetto, 2018). I therefore developed an unbiased sliding windows-based approach for identification of DSB hotspots which is DSB-centric, and by using quantiles as a means of thresholding, is relatively insensitive to differences in total break numbers between samples.

A critical parameter in peak calling or indeed identification of hotspots, is the scale at which algorithms search for regions of enrichment. In the ChIP-seq field, this is carried out at narrow (200bp-1kb) or broad scale (1kb-100kb) depending on the protein of interest (Starmer and Magnuson, 2016). Crucially, the size of a DNA damage hotspot in hPSC is unknown and likely depends on the mechanism of breakage. Despite DSB hotspots being identified in various cell types, no previous study has systematically determined the scale at which breaks are enriched. By comparing hotspot numbers called in a real INDUCE-seq dataset relative to a randomised dataset over increasing window sizes and variable threshold quantiles, I found a 100kb window coupled with a 1% threshold called ~1% of windows as hotspots in the real dataset but 0% of windows in the randomised dataset, demonstrating good specificity. The scale of hotspots identified in this manner was similar to those called using SICER in HTGTS studies of neural progenitor cell types, attributed to replication stress driven breakage, mediated by transcription (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020). My approach yielded very poor specificity at smaller window sizes <1kb. Narrow peak calling algorithms used in break-labelling studies have identified hotspots <1kb in length in various cell types, attributed to TOP2 action at chromatin loop boundaries and RNAPII pause sites (Dellino *et al.*, 2019; Gothe *et al.*, 2019) or replication stress either at short poly-dA stretches or R-loops (Tubbs *et al.*, 2018; Chakraborty *et al.*, 2020). One explanation for poor specificity at small window sizes in my dataset, is that DSBs are not enriched over short regions. However, given that annotation studies in the previous chapter revealed strong enrichment of DSBs over narrow epigenetic features, a more likely explanation is that the thresholds used in my optimisation were too lenient to detect specific DSB enrichment at small window sizes.

Whilst HTGTS-based studies in neural progenitors have identified enrichment of breaks over broad regions, their method of analysis was biased towards broad hotspot identification (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020). Conversely, break-mapping studies identifying narrow hotspots of, for example, TOP2-driven damage used analyses specifically designed for narrow peak discovery (Dellino *et al.*, 2019; Gothe *et al.*, 2019). It is unlikely that damage mechanisms identified by these different analyses are mutually exclusive in the respective cell types used, and the same is true of hPSC. Future work on narrow hotspot enrichment in hPSC will require a more comprehensive threshold/window size matrix for determination of the optimum combination for hotspot discovery at small scale.

With the systematic approach used here, I identified hotspots at 100kb scale which were reproducible between replicates of a given sample, and, following derivation of a consensus hotspot dataset, correlated well with samples of the same condition, providing reassurance that these were bona fide biological features. A potential shortcoming of the hotspot calling approach used here is that, unlike classical peak calling algorithms, it does not statistically

determine the significance of individual hotspots. However, initial hotspot identification by this method simply serves to identify candidate regions. Subsequent analysis, using DiffBind statistically determines enrichment of DSBs at these regions in pluripotent cells, relative to their differentiated controls. The ultimate result of this analysis was the discovery of 631 hotspots, identified independent of annotation and significantly enriched for breaks in pluripotent cells.

5.3.2. Characterisation of pluripotent-specific DSB hotspots

Given the scale of the hotspots identified in this study, and the enrichment of DSBs in actively transcribed regions identified in the previous chapter, I reasoned that interference between replication and transcription drives fragility at these regions. I found that pluripotent-specific hotspots showed significant overlap with long, highly transcribed genes, consistent with HTGTS studies, in NPCs, all of which also demonstrated a heightened sensitivity of these sites to replication stress (Wei *et al.*, 2016; Tena *et al.*, 2020; Wang *et al.*, 2020). Notably CFS are also broad, replication stress-sensitive genomic regions often overlapping with long genes (Ji *et al.*, 2022).

I next looked to see if DSB hotspots corresponded to sites of recurrent genetic changes in hPSC. For the purposes of this study, I focussed on translocations only, as such structural rearrangements are absolutely dependent on DSB formation (Roukos and Misteli, 2014). I intersected pluripotent-specific hotspots with cytogenetic regions harbouring a documented translocation breakpoint in hPSC, revealing 437/631 hotspots lie in translocated regions. As some cytogenetic regions are more frequently involved in structural rearrangements than others (Baker, Adam J Hirst, *et al.*, 2016), I next looked to see if hotspot break density correlates with translocation frequency but found no significant association between the two parameters. From this analysis I did note a particularly interesting hotspot on 1q21, which has ~two-fold higher break density than any other hotspot. Moreover, the 1q21 locus is frequently implicated in chromosomal translocations in hPSC (Stavish *et al., manuscript in preparation)*. Upon closer inspection, I found the 1q21 hotspot lies within fragile site FRA1F and overlaps with *PDE4DIP,* a highly expressed, long gene. Given colocalization of the hotspot with a fragile site, and a long highly expressed gene, I hypothesized that breakage in this region is driven by transcription-associated replication stress. This hypothesis is bolstered by the previous observation from our group of constitutive replication stress in pluripotent cells relative to differentiated derivatives (Halliwell *et al.*, 2020).

Interestingly, I found the 1q21 DSB hotspot extends ~20kb upstream of *PDE4DIP.* This is unlike the early work of Wei and colleagues, who found break hotspots in neural progenitor

cells to be uniquely found *within* bodies of actively transcribed genes (Wei *et al.*, 2016), subsequently proposed to stem from increased transcription-replication collisions (Wang *et al.*, 2020). Contrastingly, a study on CFS, using molecular fine mapping with FISH probes, found that, whilst CFS were near universally overlapping with long genes, the break sites often extended beyond gene bodies, akin to the 1q21 hotspot identified here (LeTallec *et al.*, 2013). The proposed model for such breakage is that RNAPII transcription over long genes displaces licensed intragenic replication origins prior to firing, resulting in long regions of DNA being replicated by solitary unidirectional replication forks, thereby heightening sensitivity to replication stress (Ji *et al.*, 2022). Such a model would fit with the 1q21 hotspot's extension upstream of *PDE4DIP*. However, proving that transcription replication collisions are not causative would require inhibiting S-phase transcription and demonstrating that it does not affect fragility at the hotspot, as was recently demonstrated at several CFS loci in lymphoblastoid cells (Brison *et al.*, 2019).

Owing to time constraints, my annotation of hotspots focussed on transcription's coincidence with DSBs, as identified in the previous chapter. Other covariates likely influence break frequency. For example, others have found DSBs to correlate with various features, including, open chromatin, early replication and the presence of CTCF binding sites (Mourad *et al.*, 2018; Ballinger *et al.*, 2019; Sun *et al.*, 2023). FishHook, is a bioinformatics package developed for identification of driver mutations in cancer which could be readily modified to cross-correlate DSB hotspots with user-input covariates (Imielinski *et al.*, 2017). By correlating hotspot break frequency with covariates identified in other cell types, it should be possible to identify those which predispose to DSB formation in hPSC, furthermore it may be possible to find hotspots with unexpectedly high breakage based on known covariates, potentially revealing novel mechanisms of DSB formation.

5.3.3. Experimental validation of the 1q21 hotspot DSB cause consequence

I hypothesized that replication stress drives breakage at the 1q21 hotspot. A logical progression from this hypothesis would be to first confirm that modulating replication stress affects DSB frequency at this site. For such validation, I had anticipated treating cells with low-doses of APH and either carrying out a second round of INDUCE-seq to determine the effect on DSBs per million in the 1q21 hotspot, or alternatively to use an orthogonal approach, by means of high throughput break-apart FISH, which makes use of FISH probes flanking a

putative break point, the proximity of which can be used to determine whether a locus is intact or not (Burman *et al.*, 2015). However, owing to long lead times on both sequencing and probe delivery, these experiments were omitted from this body of work.

I instead interrogated whether replication stress drives mutation at the 1q21 hotspot. Of particular interest was the frequency of translocations with a break-point mapping to the 1q21 hotspot. To this end, I briefly modulated replication stress in H9 hESCs and generated clonal populations from each culture condition for screening. As no method is readily available for high-throughput screening of structural variants, I initially screened clones for copy number changes of the *MDM4* gene on 1q, as documented chromosomal translocations involving 1q in hPSC largely involve gain of copy (Baker *et al.*, 2016). I observed a 0% frequency of *MDM4 gains* in untreated cells, versus a 3.5% frequency in APH-treated cells. Whilst this may be indicative of replication stress-driven mutation, the experiment lacks statistical power owing to the rarity of de novo copy gains. This low mutation rate in hPSC lines has been reported by our lab and others (Rouhani *et al.*, 2016; Thompson *et al.*, 2020). Interestingly, I identified two apparent *MDM4* gains (1.6% frequency) in the nucleoside treated condition, which was designed to reduce levels of replication stress (Bester *et al.*, 2011; Halliwell *et al.*, 2020). At the time of conducting this experiment I did not take a parallel readout of replication stress levels, by e.g. the DNA fibre assay (Jackson and Pombo, 1998), and it is therefore possible that the nucleoside treated condition failed to alleviate replication stress in these cells. It is also important to stress that these data are preliminary, and apparent gain of *MDM4* copy does not necessarily represent a chromosomal rearrangement. HPSC suffer a high rate of mitotic errors (Halliwell *et al.*, 2020), which can yield whole chromosome gains. Moreover replication stress is known to cause mis segregation of whole chromosomes (Burrell *et al.*, 2013; Palmerola *et al.*, 2022). Clones of interest, i.e., apparent copy number gains, are therefore being further characterized by low-pass whole genome sequencing to firstly corroborate that clones do harbour copy number changes and are not false positives, and secondly, to determine whether the copy number change has a break point within the 1q21 hotspot, indicating breakage has driven rearrangement.

Future work will first focus on confirming whether break density in this region is affected by replication stress. If replication stress is causative, then a mutation rate experiment could be repeated, on a larger scale. A notable improvement to throughput in clone screening would be to multiplex the qPCR reaction, whereby the internal reference and test loci are amplified in the same reaction, effectively doubling the sample throughput for a given run and reagent quantity.

In summary, I have developed an unbiased method for the identification of DSB hotspots specific to pluripotent cells. Break densities in these hotspots correlate well between cell lines

of a given condition, demonstrating a pluripotent DSB signature. I found many of these hotspots to lie within sites of recurrent genomic rearrangements and found replication stress to be a likely driver of fragility in these regions. Furthermore, I have generated preliminary experimental evidence to suggest replication-stress driven mutation at the most prevalent hotspot on chromosome 1q21.

6. General Discussion

6.1. Summary of results

Human PSC acquire recurrent genetic changes over prolonged *in vitro* culture. Such changes alter the differentiation capacity of hPSC, limit their use in models of toxicology and pose a significant obstacle to the progression of hPSC-based cell therapies to clinic (Halliwell *et al.*, 2020; Andrews *et al.*, 2022). The mutation event yielding such variants, often requires DNA damage in the form of a DSB, followed by erroneous repair. Pluripotent stem cells are known to suffer higher constitutive levels of DNA damage than their differentiated counterparts (Ahuja *et al.*, 2016; Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018; Halliwell *et al.*, 2020), however the molecular basis of this phenotype remains poorly characterised. Until now, no study has systematically mapped the genomic location of DSBs in hPSC, which is important, as knowledge of DSB position can provide valuable information as to possible DSB cause, based on context.

In this thesis, I have generated the first genome-wide map of endogenous DSBs in hPSC lines cultured under pluripotent and differentiated conditions. Using DSB maps, I sought to identify the context of DSBs, to infer putative causes and to identify regions of DSB enrichment, termed “hotspots”, to ultimately gain insights into the origins of genetic variants in hPSC. Through annotation of genetic, epigenetic, and transcriptomic datasets, I have identified transcriptional activity as a major pre-disposing factor for DSB formation. Further, I have identified a recurrent site of damage, specifically enriched in pluripotent cells, which corresponds to a site of recurrent structural rearrangement in hPSC. From the context of this DSB “hotspot”, I propose fragility is driven by transcription-mediated replication stress and have generated preliminary data to suggest replication stress increases mutation frequency in this region.

6.1.1. Characteristics of genome-wide breaks

Annotating DSB maps with genomic features showed that both pluripotent and differentiated cells are enriched for DSBs in genic regions, most notably promoters. Further epigenetic annotation with publicly available sequencing datasets for the H9 cell line, revealed greatest enrichment of DSBs in open, active chromatin at sites of transcription initiation, namely promoters and enhancers. RNA-seq data confirmed a positive association between gene expression level and DSB density. However, interestingly, R-loops, were found to be specifically depleted of DSBs in H9 cells.

Inspection of DSB coverage around the TSS of actively transcribed genes, revealed two distinct classes of DSB-dense promoters, with near-inverse DSB coverage profiles. Cluster 1 promoters harboured high density of DSBs in the regions flanking the TSS, with a low-DSB TSS. Cluster 2 promoters harboured a sharp peak of DSB coverage at the TSS with lower coverage in the flanking regions. Annotating genes of each cluster revealed that genes of cluster 2 had a significantly higher RNAPII pause index than genes of any other clusters. I propose that high DSB density at the TSS of cluster 2 genes is mediated by TOP2-cleavage acting in a physiological manner to release paused RNAPII into productive transcription elongation, as seen in other cell types (Dellino *et al.*, 2019; Sandeep Singh *et al.*, 2020). Interestingly, Cluster 1 genes harboured a conspicuously protected TSS, and this DSB pattern was conserved across hPSC lines. The mechanism of specific TSS protection in these genes remains to be determined.

This work demonstrates that DSBs occur in a non-random fashion in hPSC and further provides the first evidence of transcription-associated damage in this cell type.

6.1.2. Hotspot identification and DSB mechanism validation

I next developed an unbiased pipeline for the identification of DSB hotspots across samples, finding DSBs to be enriched in broad genomic regions of 100kb in length. Unbiased clustering of samples based on DSB density in consensus hotspots revealed that samples cluster primarily based on condition, i.e. *pluripotent* or *differentiated*, ahead of cell line, demonstrating a pluripotent-specific DSB signature. I carried out differential enrichment analysis of DSBs in hotspots between pluripotent and differentiated samples to identify pluripotent-specific hotspots. Pluripotent-specific hotspots significantly overlap with long, actively transcribed genes, suggestive of transcription-associated damage. Whilst there was no positive correlation between cytogenetic bands' translocation frequency and hotspot DSB density, a pluripotent-specific hotspot on chromosome 1q21 was amongst the most frequently implicated in chromosomal translocations in hPSC, with a DSB density around two-fold higher than any other region genome-wide. On close inspection, the 1q21 hotspot was found to lie within common fragile site FRA1F and partially overlapped with the long, actively transcribed gene *PDE4DIP*. Notably however, the hotspot extended more than 20kb upstream of *PDE4DIP,* indicating displacement of replication origins by transcription may underpin DSB formation in the region, akin to the fragility mechanism proposed by others at CFS (LeTallec *et al.*, 2013; Brison *et al.*, 2019). The 1q21 hotspot is significantly enriched for DSBs in pluripotent relative to differentiated cells, mirroring our previous observation of increased replication stress in pluripotent versus differentiated states (Halliwell *et al.*, 2020). As chromosome 1q21 is

frequently implicated in chromosomal translocations, I hypothesized that replication stress drives the DSB formation which precedes translocation at this locus. Chromosomal translocations in hPSC are predominantly accompanied by a gain of copy number (Baker *et al.*, 2016). Thus, to test my hypothesis, I compared the frequency of chromosome 1q copy number change in cells transiently cultured under conditions of variable replication stress. Clones cultured in the control condition had 0% 1q gain frequency, whereas 3.5% of aphidicolin-treated clones harboured an apparent 1q gain, providing the first mechanistic evidence for replication stress' contribution to translocation biogenesis in hPSC. As we have previously shown replication stress in hPSC to be ameliorated by medium supplementation with exogenous nucleosides (Halliwell *et al.*, 2020), it is tempting to speculate that such culture conditions could lower the rate of chromosomal translocations in hPSC, although this remains to be experimentally proven.

6.2. Limitations of the study

At the time of writing, no DSB-mapping technique has been applied to single cells. Hence, whilst γH2AX immunofluorescence carried out in this study would suggest DSB levels are variable amongst individual cells in hPSC cultures, INDUCE-seq (Dobbs *et al.*, 2022) cannot resolve such heterogeneity. Moreover, whilst nominally noise-free, INDUCE-seq, and indeed any other direct break-labelling methodology, are potentially subject to noise arising from either DNA end-resection at HR intermediates, or genome fragmentation in apoptotic cells. These issues are scarcely referred to in the DSB-mapping literature. Indeed, END-seq is the only methodology which acknowledges resected DSBs will align to a region distal to the initial DSB (Canela *et al.*, 2016). As current DSB-mapping methodologies only map population-level DSBs such techniques may be skewed by the presence of anomalous cells, such as those undergoing apoptosis. To the best of my knowledge, no study alludes to noise introduced by apoptotic genome fragmentation. Notably, hPSCs commit to apoptosis more readily than other cell types, particularly following single-cell dissociation (Watanabe *et al.*, 2007). Thus, noise due to apoptotic genome fragmentation is likely a pluripotent-specific issue. Conceivably, this could be remedied by inclusion of pan-caspase inhibitors, such as z-VAD-FMK (Garcia-Calvo *et al.*, 1998), prior to populating plates for sequencing. An alternative approach would entail FACS sorting non-apoptotic cells stained with a live dye for cleaved caspase-3, ahead of seeding cells for sequencing. It is likely that some of the observed INDUCE-seq signal in this study, particularly in pluripotent samples, emanates from apoptotic genome fragmentation. Fragmentation, however, occurs in a pseudo-random inter-nucleosomal fashion (Kawane and Nagata, 2008), and therefore such signal should effectively raise the background level of DSBs rather than introducing artefactual spikes in the data, hence enrichment identified in this

study should remain unaffected. Similarly, resection occurs over variable distances and generally increases as a function of time, following DSB formation (Canela *et al.*, 2016), hence, sites of genuine DSB enrichment should still be detectable over neighbouring regions.

Annotation of gene clusters with inverse break coverages over the promoter region revealed that genes with high break coverage over the TSS had a high RNAPII pause index, reminiscent of recently described TOP2 cleavage in other cell types (Dellino *et al.*, 2019; Sandeep Singh *et al.*, 2020). However, work in this thesis does not demonstrate TOP2 involvement at these sites, and an increased pause index could alternatively yield DSBs via formation of short, genotoxic R-loops (Chen *et al.*, 2017; Castillo-Guzman and Chédin, 2021), or high pause index could be coincidental at sites of active DNA demethylation, where DSB formation is a product of SSBs (Ray *et al.*, 2022; Wang *et al.*, 2022; Wu *et al.*, 2021). TOP2 activity at these TSS could be tested by carrying out INDUCE-seq on cells cultured in the presence and absence of the TOP2 poison etoposide, which inhibits re-ligation of TOP2-cleaved DNA, as used by Gothe and colleagues (Gothe *et al.*, 2019). Owing to its novelty, a more pressing follow-up investigation would be to determine the origin of the protected TSS in cluster 1 genes. I speculate that, as highly expressed genes, TSS of genes in cluster 1 are bound by proteins which facilitate high transcriptional activity, whilst protecting them from R-loop-, topoisomerase-, or oxidative demethylation-mediated damage. Further annotation of published ChIP-seq datasets is required to validate this.

The sliding-window approach to hotspot identification employed in this study, with empirical determination of appropriate window size, eliminates user bias and presumptions as to the size and form of a DSB hotspot. However, unlike certain peak calling algorithms, it fails to account for regional biases in coverage (Zhang *et al.*, 2008; Zang *et al.*, 2009; Starmer and Magnuson, 2016). Such biases are predominantly a product of preferential amplification of GC-rich DNA during PCR-based library preparation (Kebschull and Zador, 2015), hence not an issue in INDUCE-seq (Dobbs *et al.*, 2022). However, replication timing may also affect coverage to a lesser extent, i.e., early-replicating regions may have higher coverage of DSBs, simply by virtue of being higher copy number on a population scale. Normalization to an “input” sample, akin to those used by certain ChIP-seq peak callers, would require further sequencing, with a PCR-free gDNA library (Zhang *et al.*, 2008; Zang *et al.*, 2009; Starmer and Magnuson, 2016). For the purpose of this study, I argue that normalization is not strictly necessary. Differences in read coverage of whole genome sequencing between early and late replicating domains are small (<1.5 fold typically) (Koren *et al.*, 2021). By contrast, there is an effective 4-fold range in DSB densities in pluripotent-specific hotspots alone, hence it is unlikely that hotspots identified in this study are an artefact of replication timing.

I propose replication stress drives DSBs at the prevalent 1q21 hotspot which in turn drives formation of translocations. Whilst a qPCR-based approach shows an apparent increased copy number change frequency at chromosome 1q21 under conditions of increased replication stress, it does not show the structural nature of these gains. Replication stress is known to induce mis-segregation of entire chromosomes as well as DSB formation (Burrell *et al.*, 2013), hence putative translocation-harbouring clones require validation. To this end, we are currently screening variant clones using low-pass whole genome sequencing, the resolution of which is sufficient to identify whether break points map to the 100kb 1q21 hotspot. Owing to time constraints, here I was not able to test the effect of replication stress on DSB frequency in this region. High-throughput break-apart FISH is ideally suited to quantifying DSBs at this scale, and was recently used by Gothe and colleagues to validate TOP2-mediated DSB recurrence at specific loci in human lymphoblasts (Burman *et al.*, 2015; Gothe *et al.*, 2019).

Finally, the extension of the 1q21 hotspot beyond the TSS of *PDE4DIP* suggests replication origin displacement by transcription underpins replication stress-associated DSBs. To mechanistically test the role of transcription on origin position and DSB formation, one could delete promoter elements of *PDE4DIP* to prevent transcription, as has been accomplished for other genes in hESC (Döpper *et al.*, 2020). Cells with and without intact promoter elements for *PDE4DIP* could then be used in sequencing experiments to determine the position of replication origins and DSB density over the 1q21 hotspot, using SNS-seq, which maps nascent DNA synthesis at replication origins, and INDUCE-seq, respectively (Besnard *et al.*, 2012; Dobbs *et al.*, 2022).

6.3. Future directions

While previous studies in the field (Simara *et al.*, 2017; Vallabhaneni *et al.*, 2018), including our own (Halliwell *et al.*, 2020), showed significantly higher levels of genome damage in hPSCs compared to their somatic counterparts, the significance of my work is in revealing, for the first time, the genomic sites of such damage in hPSC. With this information in hand, future work should resolve the specific mechanisms that predispose hPSC to high levels of genome damage and, ultimately, develop strategies to reduce the appearance of DSBs in hPSC.

As my work has identified transcription-mediated genome damage as a likely cause of DSBs in hPSC, it will be interesting to identify whether such damage reflects genome damage experienced by pluripotent cells during early development *in vivo* and how it could be remedied by modification of culture conditions *in vitro*. It is possible that high levels of transcription, inherent to pluripotent cells (Efroni *et al.*, 2008), may not be easily modifiable without changing

the properties of cells. Nonetheless, culturing hPSC in a different state of pluripotency may afford different intrinsic properties and, ultimately, a more stable genome. For example, the current investigation focused on hPSC cultured in the primed pluripotent state, but various culture systems have been developed for maintenance of hPSC in a so-called naïve pluripotent state (Reviewed by Zhou et al., 2023). Moreover, a formative state, between primed and naïve pluripotency, can be captured and maintained in mouse and human systems (Kinoshita *et al.*, 2021). Naïve, formative and primed are all states of pluripotency, however each substate differs in terms of epigenetic landscape, gene expression and cell cycle regulation (Boward *et al.*, 2016; Kinoshita *et al.*, 2021), which has potential implications for DNA damage. Future work will be required to determine which substate is optimal, in terms of genome stability, for long-term expansion of hPSC.

Similarly, this study focused on hPSC grown in an adherent 2D culture system. Recent advances in 3D culture systems, are able to markedly improve the expansion rate of hPSC (Borys *et al.*, 2020; Dang *et al.*, 2021; Manstein *et al.*, 2021; Cuesta-Gomez *et al.*, 2023). Due to the large numbers of cells typically required for therapeutic applications (Serra *et al.*, 2012), expansion of hPSC for cell therapies will likely shift towards 3D culture systems in the coming years. Again, 3D systems are likely to affect cell biology and ultimately genome damage, indeed the most effective 3D culture system to date, was found to induce expression of naïve pluripotency-associated markers, independently of media modification (Cuesta-Gomez *et al.*, 2023). Interestingly, a recent study found higher proportions of naïve hPSC in G1 phase of the cell cycle when compared with their primed counterparts, suggesting increased G1 duration (Dodsworth *et al.*, 2020). An elongated G1 phase may allow more time for repair of DNA lesions prior to S-phase entry and ultimately reduced DNA damage, as proposed in mESC (Ahuja *et al.*, 2016). The increased expansion rate in 3D over 2D systems is proposed to result from decreased rates of cell death (Cuesta-Gomez *et al.*, 2023), which could be a by-product of reduced DNA damage and associated apoptosis.

Beyond the scope of genome stability in hPSC, the system used here, whereby DSBs were mapped in multiple cell lines and conditions, may serve as a useful resource for understanding common causes of genome damage and indeed genome instability, in normal, diploid human cells. Whilst choice of DSB repair pathway affects mutagenicity (Jiang, 2022), it remains largely unknown whether DSBs of different mechanistic origins, at different loci are equally mutagenic. Recurrent translocations were the only mutations specifically addressed in this study, however, coupling these DSB maps with modern high-sensitivity duplex sequencing methods (Kennedy *et al.*, 2014; Cohen *et al.*, 2021; Gydush *et al.*, 2022), may provide valuable information as to differential outcomes of DSB formation at different locations, regarding SNP

and INDEL frequencies. Conceivably, identifying classes of low-mutagenicity DSBs could inform selection of safer chemotherapeutic drugs for the treatment of cancer.

6.4. Concluding remarks

Genetic variants still pose a significant hurdle to the use of hPSC in research and clinical applications. While the majority of work in the field has focused on understanding how variant cells are selected in culture (Avery *et al.*, 2013; Barbaric *et al.*, 2014; Ben-David *et al.*, 2014; Nguyen *et al.*, 2014; Price *et al.*, 2021), my work provides important insights into the mechanisms that lead to variants' generation in hPSC. The extent of genome damage in hPSC and the genome damage hotspots identified through my work provide a platform for further mechanistic investigation of mutational processes in hPSC and a baseline for optimising culture practices that limit both the DNA damage which yields mutation, as well as the selective pressures which allow variants to rise to prevalence in culture. Together, work presented in this thesis should facilitate increased reproducibility of hPSC research and safer translation of hPSC applications in medicine.

References

Abuhashem, A., Garg, V. and Hadjantonakis, A.K. (2022) 'RNA polymerase II pausing in development: orchestrating transcription', *Open Biology*, 12(1). Available at: https://doi.org/10.1098/RSOB.210220.

Ahuja, A.K. *et al.* (2016) 'A short G1 phase imposes constitutive replication stress and fork remodelling in mouse embryonic stem cells', *Nature Communications*, 7. Available at: https://doi.org/10.1038/ncomms10660.

Akerman, I. *et al.* (2020) 'A predictable conserved DNA base composition signature defines human core DNA replication origins', *Nature communications*, 11(1). Available at: https://doi.org/10.1038/S41467-020-18527-0.

Akopian, V. *et al.* (2010) 'Comparison of defined culture systems for feeder cell free propagation of human embryonic stem cells', *In Vitro Cellular and Developmental Biology - Animal*, 46(3–4), pp. 247–258. Available at: https://doi.org/10.1007/s11626-010-9297-z.

Al-Harbi, S. *et al.* (2020) 'An update on the molecular pathogenesis and potential therapeutic targeting of AML with t(8;21)(q22;q22.1);RUNX1-RUNX1T1', *Blood advances*, 4(1), pp. 229–238. Available at: https://doi.org/10.1182/BLOODADVANCES.2019000168.

Allison, T.F. *et al.* (2018) 'Assessment of established techniques to determine developmental and malignant potential of human pluripotent stem cells', *Nature Communications 2018 9:1*, 9(1), pp. 1–15. Available at: https://doi.org/10.1038/s41467-018-04011-3.

Amir, H. *et al.* (2017) 'Spontaneous Single-Copy Loss of TP53 in Human Embryonic Stem Cells Markedly Increases Cell Proliferation and Survival', *Stem Cells*, 35(4), pp. 872–885. Available at: https://doi.org/10.1002/STEM.2550.

Amit, M. *et al.* (2004) 'Feeder Layer- and Serum-Free Culture of Human Embryonic Stem Cells1', *Biology of Reproduction*, 70(3), pp. 837–845. Available at: https://doi.org/10.1095/biolreprod.103.021147.

Amps, K. *et al.* (2011) 'Screening ethnically diverse human embryonic stem cells

identifies a chromosome 20 minimal amplicon conferring growth advantage', *Nature Biotechnology*, 29(12), pp. 1132–1144. Available at: https://doi.org/10.1038/nbt.2051.

An, R. *et al.* (2014) 'Non-Enzymatic Depurination of Nucleic Acids: Factors and Mechanisms', *PLoS ONE*, 9(12). Available at: https://doi.org/10.1371/JOURNAL.PONE.0115950.

Anand, R.P. *et al.* (2012) 'Overcoming natural replication barriers: Differential helicase requirements', *Nucleic Acids Research*, 40(3), pp. 1091–1105. Available at: https://doi.org/10.1093/nar/gkr836.

Andrews, P.W. *et al.* (1985) 'A pluripotent human stem-cell clone isolated from the TERA-2 teratocarcinoma line lacks antigens SSEA-3 and SSEA-4 in vitro, but expresses these antigens when grown as a xenograft tumor', *Differentiation*, 29(2), pp. 127–135. Available at: https://doi.org/10.1111/j.1432-0436.1985.tb00305.x.

Andrews, P.W. *et al.* (2005) 'Embryonic stem (ES) cells and embryonal carcinoma (EC) cells: opposite sides of the same coin', *Biochemical Society Transactions*, 33(6), pp. 1526–1530. Available at: https://doi.org/10.1042/bst0331526.

Andrews, P.W. *et al.* (2022) 'The consequences of recurrent genetic and epigenetic variants in human pluripotent stem cells', *Cell Stem Cell*, 29(12), pp. 1624–1636. Available at: https://doi.org/10.1016/J.STEM.2022.11.006.

Avery, S. *et al.* (2013) 'BCL-XL Mediates the Strong Selective Advantage of a 20q11.21 Amplification Commonly Found in Human Embryonic Stem Cell Cultures', *Stem Cell Reports*, 1(5), pp. 379–386. Available at: https://doi.org/10.1016/J.STEMCR.2013.10.005.

Avior, Y. *et al.* (2021) 'Cancer-Related Mutations Identified in Primed Human Pluripotent Stem Cells', *Cell Stem Cell*, 28(1), pp. 10–11. Available at: https://doi.org/10.1016/j.stem.2020.11.013.

Avior, Y., Eggan, K. and Benvenisty, N. (2019) 'Cancer-Related Mutations Identified in Primed and Naive Human Pluripotent Stem Cells', *Cell Stem Cell*, 25(4), pp. 456–461. Available at: https://doi.org/10.1016/j.stem.2019.09.001.

Baker, C.L. and Pera, M.F. (2018) 'Capturing Totipotent Stem Cells', *Cell stem cell*, 22(1), pp. 25–34. Available at: https://doi.org/10.1016/J.STEM.2017.12.011.

Baker, D., Hirst, Adam J, *et al.* (2016) 'Detecting Genetic Mosaicism in Cultures of

Human Pluripotent Stem Cells.', *Stem cell reports*, 7(5), pp. 998–1012. Available at: https://doi.org/10.1016/j.stemcr.2016.10.003.

Baker, D., Hirst, Adam J., *et al.* (2016) 'Detecting Genetic Mosaicism in Cultures of Human Pluripotent Stem Cells', *Stem Cell Reports*, 7(5), pp. 998–1012. Available at: https://doi.org/10.1016/j.stemcr.2016.10.003.

Baker, D.E.C. *et al.* (2007) 'Adaptation to culture of human embryonic stem cells and oncogenesis in vivo', *Nature Biotechnology*, pp. 207–215. Available at: https://doi.org/10.1038/nbt1285.

Ballarino, R. *et al.* (2022) 'An atlas of endogenous DNA double-strand breaks arising during human neural cell fate determination', *Scientific Data 2022 9:1*, 9(1), pp. 1–19. Available at: https://doi.org/10.1038/s41597-022-01508-x.

Ballinger, T.J. *et al.* (2019) 'Modeling double strand break susceptibility to interrogate structural variation in cancer', *Genome Biology*, 20(1), pp. 1–15. Available at: https://doi.org/10.1186/S13059-019-1635-1/FIGURES/6.

Bangalore, M.P. *et al.* (2017) 'Genotoxic Effects of Culture Media on Human Pluripotent Stem Cells', *Scientific Reports*, 7. Available at: https://doi.org/10.1038/srep42222.

Bar, S. *et al.* (2017) 'Large-Scale Analysis of Loss of Imprinting in Human Pluripotent Stem Cells', *Cell reports*, 19(5), pp. 957–968. Available at: https://doi.org/10.1016/J.CELREP.2017.04.020.

Bar, S. *et al.* (2019) 'Global Characterization of X Chromosome Inactivation in Human Pluripotent Stem Cells', *Cell Reports*, 27(1), pp. 20-29.e3. Available at: https://doi.org/10.1016/j.celrep.2019.03.019.

Baranello, L. *et al.* (2014) 'DNA Break Mapping Reveals Topoisomerase II Activity Genome-Wide', *International Journal of Molecular Sciences 2014, Vol. 15, Pages 13111-13122*, 15(7), pp. 13111–13122. Available at: https://doi.org/10.3390/IJMS150713111.

Baranello, L. *et al.* (2016) 'RNA Polymerase II Regulates Topoisomerase 1 Activity to Favor Efficient Transcription', *Cell*, 165(2), pp. 357–371. Available at: https://doi.org/10.1016/j.cell.2016.02.036.

Baranovskiy, A.G. *et al.* (2014) 'Structural basis for inhibition of DNA replication by

aphidicolin', *Nucleic Acids Research*, 42(22), p. 14013. Available at: https://doi.org/10.1093/NAR/GKU1209.

Barbaric, I. *et al.* (2014) 'Time-lapse analysis of human embryonic stem cells reveals multiple bottlenecks restricting colony formation and their relief upon culture adaptation', *Stem Cell Reports*, 3(1), pp. 142–155. Available at: https://doi.org/10.1016/j.stemcr.2014.05.006.

Barlow, J.H. *et al.* (2013) 'Identification of early replicating fragile sites that contribute to genome instability', *Cell*, 152(3), pp. 620–632. Available at: https://doi.org/10.1016/j.cell.2013.01.006.

Bárta, T. *et al.* (2010) 'Human embryonic stem cells are capable of executing G1/S checkpoint activation', *Stem Cells*, 28(7), pp. 1143–1152. Available at: https://doi.org/10.1002/stem.451.

Basu, U. *et al.* (2011) 'The RNA exosome targets the AID cytidine deaminase to both strands of transcribed duplex DNA substrates', *Cell*, 144(3), pp. 353–363. Available at: https://doi.org/10.1016/j.cell.2011.01.001.

Batlle, E. and Clevers, H. (2017) 'Cancer stem cells revisited', *Nature Medicine*. Nature Publishing Group, pp. 1124–1134. Available at: https://doi.org/10.1038/nm.4409.

Becker, K.A. *et al.* (2006) 'Self-renewal of human embryonic stem cells is supported by a shortened G1 cell cycle phase', *Journal of Cellular Physiology*, 209(3), pp. 883–893. Available at: https://doi.org/10.1002/jcp.20776.

Bell, D. *et al.* (2011) 'Integrated genomic analyses of ovarian carcinoma', *Nature*, 474(7353), pp. 609–615. Available at: https://doi.org/10.1038/nature10166.

Ben-David, U. *et al.* (2014) 'Aneuploidy induces profound changes in gene expression, proliferation and tumorigenicity of human pluripotent stem cells', *Nature Communications*, 5. Available at: https://doi.org/10.1038/ncomms5825.

Bergström, R. *et al.* (2011) 'Xeno-free culture of human pluripotent stem cells', *Methods in Molecular Biology*, 767, pp. 125–136. Available at: https://doi.org/10.1007/978-1-61779-201-4_9.

Besnard, E. *et al.* (2012) 'Unraveling cell type-specific and reprogrammable human replication origin signatures associated with G-quadruplex consensus motifs', *Nature Structural and Molecular Biology*, 19(8), pp. 837–844. Available at:

https://doi.org/10.1038/nsmb.2339.

Bester, A.C. *et al.* (2011) 'Nucleotide deficiency promotes genomic instability in early stages of cancer development', *Cell*, 145(3), pp. 435–446. Available at: https://doi.org/10.1016/j.cell.2011.03.044.

Bethune, G. *et al.* (2010) 'Epidermal growth factor receptor (EGFR) in lung cancer: An overview and update', *Journal of Thoracic Disease*. AME Publications, pp. 48–51. Available at: www.jthoracdis.com (Accessed: 8 August 2023).

Bétous, R. *et al.* (2012) 'SMARCAL1 catalyzes fork regression and holliday junction migration to maintain genome stability during DNA replication', *Genes and Development*, 26(2), pp. 151–162. Available at: https://doi.org/10.1101/gad.178459.111.

Bétous, R. *et al.* (2013) 'Substrate-Selective Repair and Restart of Replication Forks by DNA Translocases', *Cell Reports*, 3(6), pp. 1958–1969. Available at: https://doi.org/10.1016/j.celrep.2013.05.002.

Biernacka, A. *et al.* (2018) 'i-BLESS is an ultra-sensitive method for detection of DNA double-strand breaks', *Communications Biology*, 1(1). Available at: https://doi.org/10.1038/s42003-018-0165-9.

Blin, M. *et al.* (2019) 'Transcription-dependent regulation of replication dynamics modulates genome stability', *Nature structural & molecular biology*, 26(1), pp. 58–66. Available at: https://doi.org/10.1038/S41594-018-0170-1.

Böhly, N. *et al.* (2022) 'Increased replication origin firing links replication stress to whole chromosomal instability in human cancer', *Cell Reports*, 41(11). Available at: https://doi.org/10.1016/j.celrep.2022.111836.

Boregowda, S. V. *et al.* (2012) 'Atmospheric oxygen inhibits growth and differentiation of marrow-derived mouse mesenchymal stem cells via a p53-dependent mechanism: implications for long-term culture expansion', *Stem cells (Dayton, Ohio)*, 30(5), pp. 975–987. Available at: https://doi.org/10.1002/STEM.1069.

Boros-Oláh, B. *et al.* (2019) 'Drugging the R-loop interactome: RNA-DNA hybrid binding proteins as targets for cancer therapy', *DNA repair*, 84. Available at: https://doi.org/10.1016/J.DNAREP.2019.102642.

Boroviak, T. *et al.* (2014) 'The ability of inner-cell-mass cells to self-renew as

embryonic stem cells is acquired following epiblast specification', *Nature cell biology*, 16(6), pp. 513–525. Available at: https://doi.org/10.1038/NCB2965.

Borys, B.S. *et al.* (2020) 'Optimized serial expansion of human induced pluripotent stem cells using low-density inoculation to generate clinically relevant quantities in vertical-wheel bioreactors', *Stem Cells Translational Medicine*, 9(9), pp. 1036–1052. Available at: https://doi.org/10.1002/sctm.19-0406.

Botchan, M. and Berger, J. (2010) 'DNA Replication: Making Two Forks from One Prereplication Complex', *Molecular Cell*. Cell Press, pp. 860–861. Available at: https://doi.org/10.1016/j.molcel.2010.12.014.

Bouwman, B.A.M. and Crosetto, N. (2018) 'Endogenous DNA Double-Strand Breaks during DNA Transactions: Emerging Insights and Methods for Genome-Wide Profiling', *Genes*, 9(12). Available at: https://doi.org/10.3390/GENES9120632.

Boward, B., Wu, T. and Dalton, S. (2016) 'Concise Review: Control of Cell Fate Through Cell Cycle and Pluripotency Networks', *Stem Cells*. Wiley-Blackwell, pp. 1427–1436. Available at: https://doi.org/10.1002/stem.2345.

Brickner, J.R., Garzon, J.L. and Cimprich, K.A. (2022) 'Walking a tightrope: The complex balancing act of R-loops in genome stability', *Molecular Cell*. Cell Press, pp. 2267–2297. Available at: https://doi.org/10.1016/j.molcel.2022.04.014.

Brinster, R.L. (1974) 'The effect of cells transferred into the mouse blastocyst on subsequent development', *The Journal of experimental medicine*, 140(4), pp. 1049–1056. Available at: https://doi.org/10.1084/JEM.140.4.1049.

Brinton, B.T., Caddle, M.S. and Heintz, N.H. (1991) 'Position and orientation-dependent effects of a eukaryotic Z-triplex DNA motif on episomal DNA replication in COS-7 cells.', *Journal of Biological Chemistry*, 266(8), pp. 5153–5161. Available at: https://doi.org/10.1016/S0021-9258(19)67768-9.

Brison, O. *et al.* (2019) 'Transcription-mediated organization of the replication initiation program across large genes sets common fragile sites genome-wide', *Nature communications*, 10(1). Available at: https://doi.org/10.1038/S41467-019-13674-5.

Brons, I.G.M. *et al.* (2007) 'Derivation of pluripotent epiblast stem cells from mammalian embryos', *Nature 2007 448:7150*, 448(7150), pp. 191–195. Available at: https://doi.org/10.1038/nature05950.

Bunch, H. *et al.* (2015) 'Transcriptional elongation requires DNA break-induced signalling', *Nature Communications 2015 6:1*, 6(1), pp. 1–12. Available at: https://doi.org/10.1038/ncomms10191.

Burma, S. *et al.* (2001) 'ATM Phosphorylates Histone H2AX in Response to DNA Double-strand Breaks', *Journal of Biological Chemistry*, 276(45), pp. 42462–42467. Available at: https://doi.org/10.1074/jbc.C100466200.

Burman, B., Misteli, T. and Pegoraro, G. (2015) 'Quantitative detection of rare interphase chromosome breaks and translocations by high-throughput imaging', *Genome Biology*, 16(1), pp. 1–14. Available at: https://doi.org/10.1186/S13059-015-0718-X/FIGURES/6.

Burrell, R.A. *et al.* (2013) 'Replication stress links structural and numerical cancer chromosomal instability', *Nature*, 494(7438), pp. 492–496. Available at: https://doi.org/10.1038/nature11935.

Byun, T.S. *et al.* (2005) 'Functional uncoupling of MCM helicase and DNA polymerase activities activates the ATR-dependent checkpoint', *Genes and Development*, 19(9), pp. 1040–1052. Available at: https://doi.org/10.1101/gad.1301205.

Caldecott, K.W. (2008) 'Single-strand break repair and genetic disease', *Nature Reviews Genetics 2008 9:8*, 9(8), pp. 619–631. Available at: https://doi.org/10.1038/nrg2380.

Calvanese, V. *et al.* (2008) 'Cancer Genes Hypermethylated in Human Embryonic Stem Cells', *PLOS ONE*, 3(9), p. e3294. Available at: https://doi.org/10.1371/JOURNAL.PONE.0003294.

Canela, A. *et al.* (2016) 'DNA Breaks and End Resection Measured Genome-wide by End Sequencing', *Molecular Cell*, 63(5), pp. 898–911. Available at: https://doi.org/10.1016/J.MOLCEL.2016.06.034.

Canela, A. *et al.* (2017) 'Genome Organization Drives Chromosome Fragility', *Cell*, 170(3), pp. 507-521.e18. Available at: https://doi.org/10.1016/j.cell.2017.06.034.

Canela, A. *et al.* (2019) 'Topoisomerase II-Induced Chromosome Breakage and Translocation Is Determined by Chromosome Architecture and Transcriptional Activity', *Molecular cell*, 75(2), pp. 252-266.e8. Available at: https://doi.org/10.1016/J.MOLCEL.2019.04.030.

Cannan, W.J. and Pederson, D.S. (2016) ‘Mechanisms and Consequences of Double-strand DNA Break Formation in Chromatin’, *Journal of cellular physiology*, 231(1), p. 3. Available at: https://doi.org/10.1002/JCP.25048.

Castedo, S. (1993) ‘Cytogenetics of testicular germ cell tumors’, *European Journal of Cancer*, 29, p. S30. Available at: https://doi.org/10.1016/0959-8049(93)90755-5.

Castillo-Guzman, D. and Chédin, F. (2021) ‘Defining R-loop classes and their contributions to genome instability’, *DNA repair*, 106. Available at: https://doi.org/10.1016/J.DNAREP.2021.103182.

Cayrou, C. *et al.* (2011) ‘Genome-scale analysis of metazoan replication origins reveals their organization in specific but flexible sites defined by conserved features’, *Genome Research*, 21(9), pp. 1438–1449. Available at: https://doi.org/10.1101/gr.121830.111.

Ceccaldi, R., Rondinelli, B. and D’Andrea, A.D. (2016) ‘Repair Pathway Choices and Consequences at the Double-Strand Break’, *Trends in Cell Biology*. Elsevier Ltd, pp. 52–64. Available at: https://doi.org/10.1016/j.tcb.2015.07.009.

Chakraborty, A. *et al.* (2020) ‘Replication Stress Induces Global Chromosome Breakage in the Fragile X Genome’, *Cell Reports*, 32(12), p. 108179. Available at: https://doi.org/10.1016/j.celrep.2020.108179.

Chappidi, N. *et al.* (2020) ‘Fork Cleavage-Religation Cycle and Active Transcription Mediate Replication Restart after Fork Stalling at Co-transcriptional R-Loops’, *Molecular Cell*, 77(3), pp. 528-541.e8. Available at: https://doi.org/10.1016/j.molcel.2019.10.026.

Chatterjee, N. and Walker, G.C. (2017) ‘Mechanisms of DNA damage, repair, and mutagenesis’, *Environmental and Molecular Mutagenesis*, 58(5), pp. 235–263. Available at: https://doi.org/10.1002/em.22087.

Chen, G. *et al.* (2010) ‘Actin-myosin contractility is responsible for the reduced viability of dissociated human embryonic stem cells’, *Cell Stem Cell*, 7(2), pp. 240–248. Available at: https://doi.org/10.1016/j.stem.2010.06.017.

Chen, G. *et al.* (2011) ‘Chemically defined conditions for human iPSC derivation and culture’, *Nature Methods*, 8(5), pp. 424–429. Available at: https://doi.org/10.1038/nmeth.1593.

Chen, J. (2016) ‘The cell-cycle arrest and apoptotic functions of p53 in tumor initiation and progression’, *Cold Spring Harbor Perspectives in Medicine*, 6(3), p. a026104. Available at: https://doi.org/10.1101/cshperspect.a026104.

Chen, L. *et al.* (2017) ‘R-ChIP Using Inactive RNase H Reveals Dynamic Coupling of R-loops with Transcriptional Pausing at Gene Promoters’, *Molecular Cell*, 68(4), pp. 745-757.e5. Available at: https://doi.org/10.1016/j.molcel.2017.10.008.

Chen, X. *et al.* (2022) ‘Mutant p53 in cancer: from molecular mechanism to therapeutic modulation’, *Cell Death and Disease*. Springer Nature, pp. 1–14. Available at: https://doi.org/10.1038/s41419-022-05408-1.

Chen, Y. *et al.* (2021) ‘A versatile polypharmacology platform promotes cytoprotection and viability of human pluripotent and differentiated cells’, *Nature Methods 2021 18:5*, 18(5), pp. 528–541. Available at: https://doi.org/10.1038/s41592-021-01126-2.

Chen, Y.H. *et al.* (2019) ‘Transcription shapes DNA replication initiation and termination in human cells’, *Nature structural & molecular biology*, 26(1), p. 67. Available at: https://doi.org/10.1038/S41594-018-0171-0.

Chiappetta, G. *et al.* (2015) ‘PATZ1 acts as a tumor suppressor in thyroid cancer via targeting p53-dependent genes involved in EMT and cell migration’, *Oncotarget*, 6(7), pp. 5310–5323. Available at: https://doi.org/10.18632/oncotarget.2776.

Chiarle, R. *et al.* (2011) ‘Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells’, *Cell*, 147(1), pp. 107–119. Available at: https://doi.org/10.1016/J.CELL.2011.07.049.

Church, D.M. *et al.* (2011) ‘Modernizing reference genome assemblies’, *PLoS biology*, 9(7). Available at: https://doi.org/10.1371/JOURNAL.PBIO.1001091.

Cimprich, K.A. and Cortez, D. (2008) ‘ATR: An essential regulator of genome integrity’, *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 616–627. Available at: https://doi.org/10.1038/nrm2450.

Cisneros-Aguirre, M., Ping, X. and Stark, J.M. (2022) ‘To indel or not to indel: Factors influencing mutagenesis during chromosomal break end joining’, *DNA Repair*, 118, p. 103380. Available at: https://doi.org/10.1016/J.DNAREP.2022.103380.

Cohen, J.D. *et al.* (2021) ‘Detection of low-frequency DNA variants by targeted sequencing of the Watson and Crick strands’, *Nature Biotechnology*, 39(10), pp.

1220–1227. Available at: https://doi.org/10.1038/s41587-021-00900-z.

Courtot, A.M. *et al.* (2014) 'Morphological analysis of human induced pluripotent stem cells during induced differentiation and reverse programming', *BioResearch Open Access*, 3(5), pp. 206–216. Available at: https://doi.org/10.1089/biores.2014.0028.

Cramer, P. (2019) 'Organization and regulation of gene transcription', *Nature 2019 573:7772*, 573(7772), pp. 45–54. Available at: https://doi.org/10.1038/s41586-019-1517-4.

Cropp, C.S. *et al.* (1990) 'Loss of heterozygosity on chromosomes 17 and 18 in breast carcinoma: Two additional regions identified', *Proceedings of the National Academy of Sciences of the United States of America*, 87(19), pp. 7737–7741. Available at: https://doi.org/10.1073/pnas.87.19.7737.

Crosetto, N. *et al.* (2013) 'Nucleotide-resolution DNA double-strand breaks mapping by next-generation sequencing', *Nature methods*, 10(4), p. 361. Available at: https://doi.org/10.1038/NMETH.2408.

Cuesta-Gomez, N. *et al.* (2023) 'Suspension culture improves iPSC expansion and pluripotency phenotype', *Stem Cell Research and Therapy*, 14(1), p. 154. Available at: https://doi.org/10.1186/s13287-023-03382-9.

Danecek, P. *et al.* (2021) 'Twelve years of SAMtools and BCFtools', *GigaScience*, 10(2). Available at: https://doi.org/10.1093/GIGASCIENCE/GIAB008.

Dang, T. *et al.* (2021) 'Computational fluid dynamic characterization of vertical-wheel bioreactors used for effective scale-up of human induced pluripotent stem cell aggregate culture', *The Canadian Journal of Chemical Engineering*, 99(11), pp. 2536–2553. Available at: https://doi.org/10.1002/cjce.24253.

Deaton, A.M. and Bird, A. (2011) 'CpG islands and the regulation of transcription', *Genes & Development*, 25(10), pp. 1010–1022. Available at: https://doi.org/10.1101/GAD.2037511.

Dellino, G.I. *et al.* (2019) 'Release of paused RNA polymerase II at specific loci favors DNA double-strand-break formation and promotes cancer translocations', *Nature Genetics 2019 51:6*, 51(6), pp. 1011–1023. Available at: https://doi.org/10.1038/s41588-019-0421-z.

Desmarais, J.A. *et al.* (2012) 'Human Embryonic Stem Cells Fail to Activate CHK1 and

Commit to Apoptosis in Response to DNA Replication Stress', *STEM CELLS*, 30(7), pp. 1385–1393. Available at: https://doi.org/10.1002/stem.1117.

Desmarais, J.A. *et al.* (2016) 'Apoptosis and failure of checkpoint kinase 1 activation in human induced pluripotent stem cells under replication stress', *Stem Cell Research and Therapy*, 7(1), pp. 1–7. Available at: https://doi.org/10.1186/S13287-016-0279-2/FIGURES/3.

Dewar, J.M., Budzowska, M. and Walter, J.C. (2015) 'The mechanism of DNA replication termination in vertebrates', *Nature*, 525(7569), pp. 345–350. Available at: https://doi.org/10.1038/nature14887.

Dobbs, F.M. *et al.* (2022) 'Precision digital mapping of endogenous and induced genomic DNA breaks by INDUCE-seq', *Nature communications*, 13(1), p. 3989. Available at: https://doi.org/10.1038/S41467-022-31702-9.

Dodsworth, B.T. *et al.* (2020) 'Rates of homology directed repair of CRISPR-Cas9 induced double strand breaks are lower in naïve compared to primed human pluripotent stem cells', *Stem Cell Research*, 46, p. 101852. Available at: https://doi.org/10.1016/j.scr.2020.101852.

Döpper, H. *et al.* (2020) 'Biallelic and monoallelic deletion of the RB1 promoter in six isogenic clonal H9 hESC lines', *Stem Cell Research*, 45, p. 101779. Available at: https://doi.org/10.1016/j.scr.2020.101779.

Draper, J.S. *et al.* (2004) 'Recurrent gain of chromosomes 17q and 12 in cultured human embryonic stem cells', *Nature Biotechnology*, 22(1), pp. 53–54. Available at: https://doi.org/10.1038/nbt922.

Durinck, S. *et al.* (2009) 'Mapping identifiers for the integration of genomic datasets with the R/ Bioconductor package biomaRt', *Nature Protocols*, 4(8), pp. 1184–1191. Available at: https://doi.org/10.1038/nprot.2009.97.

Durkin, S.G. and Glover, T.W. (2007) 'Chromosome fragile sites', *Annual review of genetics*, 41, pp. 169–192. Available at: https://doi.org/10.1146/ANNUREV.GENET.41.042007.165900.

Edgington, E.S. (1980) 'Validity of Randomization Tests for one-Subject Experiments', *http://dx.doi.org/10.3102/10769986005003235*, 5(3), pp. 235–251. Available at: https://doi.org/10.3102/10769986005003235.

Efroni, S. *et al.* (2008) ‘Global Transcription in Pluripotent Embryonic Stem Cells’, *Cell Stem Cell*, 2(5), pp. 437–447. Available at: https://doi.org/10.1016/j.stem.2008.03.021.

Enver, T. *et al.* (2009) ‘Stem cell states, fates, and the rules of attraction’, *Cell stem cell*, 4(5), pp. 387–397. Available at: https://doi.org/10.1016/J.STEM.2009.04.011.

Ernst, J. and Kellis, M. (2010) ‘Discovery and characterization of chromatin states for systematic annotation of the human genome’, *Nature Biotechnology 2010 28:8*, 28(8), pp. 817–825. Available at: https://doi.org/10.1038/nbt.1662.

Evans, M.J. and Kaufman, M.H. (1981) ‘Establishment in culture of pluripotential cells from mouse embryos’, *Nature 1981 292:5819*, 292(5819), pp. 154–156. Available at: https://doi.org/10.1038/292154a0.

Fazeli, A. *et al.* (2011) ‘Altered patterns of differentiation in karyotypically abnormal human embryonic stem cells’, *The International Journal of Developmental Biology*, 55(2), pp. 175–180. Available at: https://doi.org/10.1387/ijdb.103177af.

Fernández-Cid, A. *et al.* (2013) ‘An ORC/Cdc6/MCM2-7 Complex Is Formed in a Multistep Reaction to Serve as a Platform for MCM Double-Hexamer Assembly’, *Molecular Cell*, 50(4), pp. 577–588. Available at: https://doi.org/10.1016/j.molcel.2013.03.026.

Finch, B.W. and Ephrussi, B. (1967) ‘RETENTION OF MULTIPLE DEVELOPMENTAL POTENTIALITIES BY CELLS OF A MOUSE TESTICULAR TERATOCARCINOMA DURING PROLONGED CULTURE in vitro AND THEIR EXTINCTION UPON HYBRIDIZATION WITH CELLS OF PERMANENT LINES’, *Proceedings of the National Academy of Sciences of the United States of America*, 57(3), pp. 615–621. Available at: https://doi.org/10.1073/PNAS.57.3.615.

Flamier, A., Singh, S. and Rasmussen, T.P. (2017) ‘A standardized human embryoid body platform for the detection and analysis of teratogens’, *PloS one*, 12(2). Available at: https://doi.org/10.1371/JOURNAL.PONE.0171101.

Fletcher, C.E. *et al.* (2022) ‘A non-coding RNA balancing act: miR-346-induced DNA damage is limited by the long non-coding RNA NORAD in prostate cancer’, *Molecular cancer*, 21(1). Available at: https://doi.org/10.1186/S12943-022-01540-W.

Forment, J. V. and Jackson, S.P. (2015) ‘A flow-cytometry-based method to simplify

the analysis and quantification of protein association to chromatin in mammalian cells', *Nature protocols*, 10(9), p. 1297. Available at: https://doi.org/10.1038/NPROT.2015.066.

Fornes, O. *et al.* (2020) 'JASPAR 2020: update of the open-access database of transcription factor binding profiles', *Nucleic Acids Research*, 48(D1), pp. D87–D92. Available at: https://doi.org/10.1093/NAR/GKZ1001.

Forristal, C.E. *et al.* (2010) 'Hypoxia inducible factors regulate pluripotency and proliferation in human embryonic stem cells cultured at reduced oxygen tensions', *REPRODUCTION*, 139(1), pp. 85–97. Available at: https://doi.org/10.1530/REP-09-0300.

Forsyth, N.R. *et al.* (2006) 'Physiologic oxygen enhances human embryonic stem cell clonal recovery and reduces chromosomal abnormalities', *Cloning and Stem Cells*, 8(1), pp. 16–23. Available at: https://doi.org/10.1089/clo.2006.8.16.

Frankish, A. *et al.* (2019) 'GENCODE reference annotation for the human and mouse genomes', *Nucleic Acids Research*, 47(D1), pp. D766–D773. Available at: https://doi.org/10.1093/nar/gky955.

Fusaki, N. *et al.* (2009) 'Efficient induction of transgene-free human pluripotent stem cells using a vector based on Sendai virus, an RNA virus that does not integrate into the host genome', *Proceedings of the Japan Academy. Series B, Physical and biological sciences*, 85(8), pp. 348–362. Available at: https://doi.org/10.2183/PJAB.85.348.

Fynes, K. *et al.* (2014) 'The Differential Effects of 2% Oxygen Preconditioning on the Subsequent Differentiation of Mouse and Human Pluripotent Stem Cells', *https://home.liebertpub.com/scd*, 23(16), pp. 1910–1922. Available at: https://doi.org/10.1089/SCD.2013.0504.

Gaillard, H., García-Muse, T. and Aguilera, A. (2015) 'Replication stress and cancer', *Nature reviews. Cancer*, 15(5), pp. 276–280. Available at: https://doi.org/10.1038/NRC3916.

Garber, K. (2015) 'RIKEN suspends first clinical trial involving induced pluripotent stem cells', *Nature biotechnology*, 33(9), pp. 890–891. Available at: https://doi.org/10.1038/nbt0915-890.

Garcia-Calvo, M. *et al.* (1998) 'Inhibition of human caspases by peptide-based and macromolecular inhibitors', *Journal of Biological Chemistry*, 273(49), pp. 32608–32613. Available at: https://doi.org/10.1074/jbc.273.49.32608.

García-Sanz, P. *et al.* (2017) 'Chromatin remodelling and DNA repair genes are frequently mutated in endometrioid endometrial carcinoma', *International Journal of Cancer*, 140(7), pp. 1551–1563. Available at: https://doi.org/10.1002/ijc.30573.

Garg, P. and Burgers, P.M.J. (2005) 'DNA Polymerases that Propagate the Eukaryotic DNA Replication Fork', *Critical Reviews in Biochemistry and Molecular Biology*, 40(2), pp. 115–128. Available at: https://doi.org/10.1080/10409230590935433.

Gaspar-Maia, A. *et al.* (2009) 'Chd1 regulates open chromatin and pluripotency of embryonic stem cells', *Nature*, 460(7257), pp. 863–868. Available at: https://doi.org/10.1038/nature08212.

Gearing, L.J. *et al.* (2019) 'CiiiDER: A tool for predicting and analysing transcription factor binding sites', *PLoS ONE*, 14(9). Available at: https://doi.org/10.1371/JOURNAL.PONE.0215495.

Georgakilas, A.G. *et al.* (2014) 'Are common fragile sites merely structural domains or highly organized "functional" units susceptible to oncogenic stress?', *Cellular and Molecular Life Sciences*, 71(23), p. 4519. Available at: https://doi.org/10.1007/S00018-014-1717-X.

Georgoulis, A. *et al.* (2017) 'Genome Instability and γH2AX', *International Journal of Molecular Sciences*, 18(9). Available at: https://doi.org/10.3390/IJMS18091979.

Ginno, P.A. *et al.* (2012) 'R-Loop Formation Is a Distinctive Characteristic of Unmethylated Human CpG Island Promoters', *Molecular Cell*, 45(6), pp. 814–825. Available at: https://doi.org/10.1016/j.molcel.2012.01.017.

Glover, T.W. *et al.* (1984) 'DNA polymerase alpha inhibition by aphidicolin induces gaps and breaks at common fragile sites in human chromosomes', *Human genetics*, 67(2), pp. 136–142. Available at: https://doi.org/10.1007/BF00272988.

Golob, J.L. *et al.* (2008) 'Chromatin remodeling during mouse and human embryonic stem cell differentiation', *Developmental Dynamics*, 237(5), pp. 1389–1398. Available at: https://doi.org/10.1002/dvdy.21545.

Gómez-Cabello, D. *et al.* (2022) 'CtIP-dependent nascent RNA expression flanking

DNA breaks guides the choice of DNA repair pathway', *Nature communications*, 13(1), p. 5303. Available at: https://doi.org/10.1038/S41467-022-33027-Z.

Gordon, D.J., Resio, B. and Pellman, D. (2012) 'Causes and consequences of aneuploidy in cancer', *Nature Reviews Genetics*. Nature Publishing Group, pp. 189–203. Available at: https://doi.org/10.1038/nrg3123.

Gore, A. *et al.* (2011) 'Somatic coding mutations in human induced pluripotent stem cells', *Nature*, 471(7336), pp. 63–67. Available at: https://doi.org/10.1038/nature09805.

Gothe, H.J. *et al.* (2019) 'Spatial Chromosome Folding and Active Transcription Drive DNA Fragility and Formation of Oncogenic MLL Translocations', *Molecular Cell*, 75(2), pp. 267-283.e12. Available at: https://doi.org/10.1016/J.MOLCEL.2019.05.015.

Gottweis, H. (2002) 'Stem Cell Policies in the United States and in Germany', *Policy Studies Journal*, 30(4), pp. 444–469. Available at: https://doi.org/10.1111/J.1541-0072.2002.TB02158.X.

Gros, J. *et al.* (2015a) 'Post-licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA', *Molecular Cell*, 60(5), pp. 797–807. Available at: https://doi.org/10.1016/j.molcel.2015.10.022.

Gros, J. *et al.* (2015b) 'Post-licensing Specification of Eukaryotic Replication Origins by Facilitated Mcm2-7 Sliding along DNA', *Molecular cell*, 60(5), pp. 797–807. Available at: https://doi.org/10.1016/J.MOLCEL.2015.10.022.

Grossmann, V. *et al.* (2011) 'Whole-exome sequencing identifies somatic mutations of BCOR in acute myeloid leukemia with normal karyotype', *Blood*, 118(23), pp. 6153–6163. Available at: https://doi.org/10.1182/blood-2011-07-365320.

Grzelak, A., Rychlik, B. and Bartosz, G. (2001) 'Light-dependent generation of reactive oxygen species in cell culture media', *Free Radical Biology and Medicine*, 30(12), pp. 1418–1425. Available at: https://doi.org/10.1016/S0891-5849(01)00545-7.

Guan, J. *et al.* (2022) 'Chemical reprogramming of human somatic cells to pluripotent stem cells', *Nature 2022 605:7909*, 605(7909), pp. 325–331. Available at: https://doi.org/10.1038/s41586-022-04593-5.

Guled, M. *et al.* (2008) 'Array comparative genomic hybridization analysis of olfactory neuroblastoma', *Modern Pathology*, 21(6), pp. 770–778. Available at:

https://doi.org/10.1038/modpathol.2008.57.

Guo, C.W. *et al.* (2013) 'Culture under low physiological oxygen conditions improves the stemness and quality of induced pluripotent stem cells', *Journal of Cellular Physiology*, 228(11), pp. 2159–2166. Available at: https://doi.org/10.1002/JCP.24389.

Gupta, A. *et al.* (2014) 'Role of 53BP1 in the regulation of DNA double-strand break repair pathway choice', *Radiation Research*. NIH Public Access, pp. 1–8. Available at: https://doi.org/10.1667/RR13572.1.

Gydush, G. *et al.* (2022) 'Massively parallel enrichment of low-frequency alleles enables duplex sequencing at low depth', *Nature Biomedical Engineering*, 6(3), pp. 257–266. Available at: https://doi.org/10.1038/s41551-022-00855-9.

Hahn, S.A. *et al.* (1996) 'DPC4, a candidate tumor suppressor gene at human chromosome 18q21.1', *Science*, 271(5247), pp. 350–353. Available at: https://doi.org/10.1126/science.271.5247.350.

Halliwell, J., Barbaric, I. and Andrews, P.W. (2020) 'Acquired genetic changes in human pluripotent stem cells: origins and consequences', *Nature Reviews Molecular Cell Biology 2020 21:12*, 21(12), pp. 715–728. Available at: https://doi.org/10.1038/s41580-020-00292-z.

Halliwell, J.A. *et al.* (2020) 'Nucleosides Rescue Replication-Mediated Genome Instability of Human Pluripotent Stem Cells', *Stem Cell Reports* [Preprint]. Available at: https://doi.org/10.1016/j.stemcr.2020.04.004.

Halliwell, J.A. *et al.* (2021) 'Nanopore Sequencing Indicates That Tandem Amplification of Chromosome 20q11.21 in Human Pluripotent Stem Cells Is Driven by Break-Induced Replication', *Stem Cells and Development*, 30(11), pp. 578–586. Available at: https://doi.org/10.1089/scd.2021.0013.

Hamperl, S. *et al.* (2017) 'Transcription-replication conflict orientation modulates R-loop levels and activates distinct DNA damage responses', *Cell*, 170(4), p. 774. Available at: https://doi.org/10.1016/J.CELL.2017.07.043.

Hamperl, S. and Cimprich, K.A. (2016) 'Conflict Resolution in the Genome: How Transcription and Replication Make It Work', *Cell*. Cell Press, pp. 1455–1467. Available at: https://doi.org/10.1016/j.cell.2016.09.053.

Hanada, K. *et al.* (2007) 'The structure-specific endonuclease Mus81 contributes to

replication restart by generating double-strand DNA breaks', *Nature Structural and Molecular Biology*, 14(11), pp. 1096–1104. Available at: https://doi.org/10.1038/nsmb1313.

Harris, C.P. *et al.* (2003) 'Comprehensive molecular cytogenetic characterization of cervical cancer cell lines', *Genes, Chromosomes and Cancer*, 36(3), pp. 233–241. Available at: https://doi.org/10.1002/gcc.10158.

Hartono, S.R. *et al.* (2018) 'The Affinity of the S9.6 Antibody for Double-Stranded RNAs Impacts the Accurate Mapping of R-Loops in Fission Yeast', *Journal of Molecular Biology*, 430(3), pp. 272–284. Available at: https://doi.org/10.1016/j.jmb.2017.12.016.

Hassold, T. and Hunt, P. (2001) 'To err (meiotically) is human: The genesis of human aneuploidy', *Nature Reviews Genetics*. Nature Publishing Group, pp. 280–291. Available at: https://doi.org/10.1038/35066065.

Hastings, P.J. *et al.* (2009) 'Mechanisms of change in gene copy number', *Nature reviews. Genetics*, 10(8), pp. 551–564. Available at: https://doi.org/10.1038/NRG2593.

Hazan, I. *et al.* (2019) 'Activation of Oncogenic Super-Enhancers Is Coupled with DNA Repair by RAD51', *Cell Reports*, 29(3), pp. 560-572.e4. Available at: https://doi.org/10.1016/J.CELREP.2019.09.001/ATTACHMENT/C25EA420-AB8E-41F9-B4B6-767032C6233D/MMC1.PDF.

He, Y.F. *et al.* (2011) 'Tet-mediated formation of 5-carboxylcytosine and its excision by TDG in mammalian DNA', *Science*, 333(6047), pp. 1303–1307. Available at: https://doi.org/10.1126/SCIENCE.1210944/SUPPL_FILE/HE.SOM.PDF.

Helmrich, A., Ballarino, M. and Tora, L. (2011) 'Collisions between Replication and Transcription Complexes Cause Common Fragile Site Instability at the Longest Human Genes', *Molecular Cell*, 44(6), pp. 966–977. Available at: https://doi.org/10.1016/j.molcel.2011.10.013.

Herszfeld, D. *et al.* (2006) 'CD30 is a survival factor and a biomarker for transformed human pluripotent stem cells', *Nature Biotechnology*, 24(3), pp. 351–357. Available at: https://doi.org/10.1038/nbt1197.

Hidmi, O.I. and Aqeilan, R.I. (2022) 'R-loops and Topoisomerase 1 facilitate formation

of transcriptional DSBs at gene bodies of hypertranscribed genes', *bioRxiv*, p. 2022.12.12.520103. Available at: https://doi.org/10.1101/2022.12.12.520103.

Hoffman, E.A. *et al.* (2015) 'Break-seq reveals hydroxyurea-induced chromosome fragility as a result of unscheduled conflict between DNA replication and transcription', *Genome research*, 25(3), pp. 402–412. Available at: https://doi.org/10.1101/GR.180497.114.

Hong, Y.J. and Do, J.T. (2019) 'Neural Lineage Differentiation From Pluripotent Stem Cells to Mimic Human Brain Tissues', *Frontiers in bioengineering and biotechnology*, 7. Available at: https://doi.org/10.3389/FBIOE.2019.00400.

Howe, F.S. *et al.* (2017) 'Is H3K4me3 instructive for transcription activation?', *BioEssays : news and reviews in molecular, cellular and developmental biology*, 39(1), pp. 1–12. Available at: https://doi.org/10.1002/BIES.201600095.

hPSCreg (2023) *Clinical Studies · hPSCreg*, *hPSCreg.eu*. Available at: https://hpscreg.eu/clinical_studies (Accessed: 4 August 2023).

Hu, N. *et al.* (2022) 'RUNX1T1 function in cell fate', *Stem Cell Research and Therapy*, 13(1), pp. 1–10. Available at: https://doi.org/10.1186/S13287-022-03074-W/FIGURES/4.

Huang, W. *et al.* (2012) 'ART: a next-generation sequencing read simulator', *Bioinformatics*, 28(4), p. 593. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTR708.

Hussain, S.P., Hofseth, L.J. and Harris, C.C. (2003) 'Radical causes of cancer', *Nature Reviews Cancer 2003 3:4*, 3(4), pp. 276–285. Available at: https://doi.org/10.1038/nrc1046.

Hutson, A.D. and Yu, H. (2021) 'A robust permutation test for the concordance correlation coefficient', *Pharmaceutical statistics*, 20(4), pp. 696–709. Available at: https://doi.org/10.1002/PST.2101.

Iacovoni, J.S. *et al.* (2010) 'High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome', *The EMBO Journal*, 29(8), p. 1446. Available at: https://doi.org/10.1038/EMBOJ.2010.38.

Ilic, D. and Ogilvie, C. (2022) 'Pluripotent Stem Cells in Clinical Setting-New Developments and Overview of Current Status', *Stem cells (Dayton, Ohio)*, 40(9), pp.

791–801. Available at: https://doi.org/10.1093/STMCLS/SXAC040.

Imielinski, M., Guo, G. and Meyerson, M. (2017) 'Insertions and Deletions Target Lineage-Defining Genes in Human Cancers', *Cell*, 168(3), pp. 460-472.e14. Available at: https://doi.org/10.1016/J.CELL.2016.12.025.

Ismail, I.H., Wadhra, T.I. and Hammarsten, O. (2007) 'An optimized method for detecting gamma-H2AX in blood cells reveals a significant interindividual variation in the gamma-H2AX response among humans', *Nucleic Acids Research*, 35(5), p. e36. Available at: https://doi.org/10.1093/nar/gkl1169.

ISSCR (2023) *Standards Document — International Society for Stem Cell Research*, *ISSCR*. Available at: https://www.isscr.org/standards-document (Accessed: 3 August 2023).

Jackson, D.A. and Pombo, A. (1998) 'Replicon clusters are stable units of chromosome structure: Evidence that nuclear organization contributes to the efficient activation and propagation of S phase in human cells', *Journal of Cell Biology*, 140(6), pp. 1285–1295. Available at: https://doi.org/10.1083/jcb.140.6.1285.

Jacobs, K. *et al.* (2016) 'Higher-Density Culture in Human Embryonic Stem Cells Results in DNA Damage and Genome Instability', *Stem Cell Reports*, 6(3), pp. 330–341. Available at: https://doi.org/10.1016/j.stemcr.2016.01.015.

Jagannathan, L., Cuddapah, S. and Costa, M. (2016) 'Oxidative Stress Under Ambient and Physiological Oxygen Tension in Tissue Culture', *Current Pharmacology Reports*, 2(2), pp. 64–72. Available at: https://doi.org/10.1007/S40495-016-0050-5/FULLTEXT.HTML.

Jauniaux, E. *et al.* (1999) 'In-vivo measurement of intrauterine gases and acid-base values early in human pregnancy', *Human Reproduction*, 14(11), pp. 2901–2904. Available at: https://doi.org/10.1093/humrep/14.11.2901.

Jhuang, Y.L. *et al.* (2022) 'SIN3-HDAC complex-associated factor, a chromatin remodelling gene located in the 12p amplicon, is a potential germ cell tumour-specific oncogene', *The Journal of Pathology*, 258(4), pp. 353–365. Available at: https://doi.org/10.1002/PATH.6007.

Ji, F. *et al.* (2022) 'New Era of Mapping and Understanding Common Fragile Sites: An Updated Review on Origin of Chromosome Fragility', *Frontiers in Genetics*, 13, p.

1182. Available at: https://doi.org/10.3389/FGENE.2022.906957/BIBTEX.

Ji, J. *et al.* (2012) 'Elevated Coding Mutation Rate During the Reprogramming of Human Somatic Cells into Induced Pluripotent Stem Cells', *Stem Cells*, 30(3), pp. 435–440. Available at: https://doi.org/10.1002/STEM.1011.

Ji, J. *et al.* (2014) 'Antioxidant supplementation reduces genomic aberrations in human induced pluripotent stem cells', *Stem Cell Reports*, 2(1), pp. 44–51. Available at: https://doi.org/10.1016/j.stemcr.2013.11.004.

Ji, J. *et al.* (2019) 'Expression pattern of CDK12 protein in gastric cancer and its positive correlation with CD8+ cell density and CCL12 expression', *International journal of medical sciences*, 16(8), pp. 1142–1148. Available at: https://doi.org/10.7150/IJMS.34541.

Jia, F. *et al.* (2010) 'A Nonviral Minicircle Vector for Deriving Human iPS Cells', *Nature methods*, 7(3), p. 197. Available at: https://doi.org/10.1038/NMETH.1426.

Jiang, Y. (2022) 'Contribution of Microhomology to Genome Instability: Connection between DNA Repair and Replication Stress', *International Journal of Molecular Sciences*. MDPI, p. 12937. Available at: https://doi.org/10.3390/ijms232112937.

Jinek, M. *et al.* (2012) 'A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity', *Science (New York, N.Y.)*, 337(6096), pp. 816–821. Available at: https://doi.org/10.1126/SCIENCE.1225829.

Johansson, E. and MacNeill, S.A. (2010) 'The eukaryotic replicative DNA polymerases take shape', *Trends in Biochemical Sciences*. Elsevier Current Trends, pp. 339–347. Available at: https://doi.org/10.1016/j.tibs.2010.01.004.

Jones-Villeneuve, E.M.V. *et al.* (1982) 'Retinoic acid induces embryonal carcinoma cells to differentiate into neurons and glial cells', *Journal of Cell Biology*, 94(2), pp. 253–262. Available at: https://doi.org/10.1083/jcb.94.2.253.

Jones, R.M. *et al.* (2013) 'Increased replication initiation and conflicts with transcription underlie Cyclin E-induced replication stress', *Oncogene*, 32(32), pp. 3744–3753. Available at: https://doi.org/10.1038/onc.2012.387.

Jonkers, I., Kwak, H. and Lis, J.T. (2014) 'Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons', *eLife*, 2014(3). Available at: https://doi.org/10.7554/ELIFE.02407.

Ju, B.G. *et al.* (2006) 'A topoisomerase IIβ-mediated dsDNA break required for regulated transcription', *Science*, 312(5781), pp. 1798–1802. Available at: https://doi.org/10.1126/SCIENCE.1127196/SUPPL_FILE/JU.SOM.REV2.PDF.

Jurga, M. *et al.* (2021) 'USP11 controls R-loops by regulating senataxin proteostasis', *Nature Communications 2021 12:1*, 12(1), pp. 1–18. Available at: https://doi.org/10.1038/s41467-021-25459-w.

Kaiser, J. (2011) 'Embryonic stem cells: Researchers mull impact of Geron's sudden exit from field', *Science*, 334(6059), p. 1043. Available at: https://doi.org/10.1126/SCIENCE.334.6059.1043/ASSET/581E1163-0C3F-44FE-ACE9-AEBA8968EE96/ASSETS/GRAPHIC/334_1043_F1.GIF.

KATAOKA, Y. *et al.* (2006) 'Flow Cytometric Analysis of Phosphorylated Histone H2AX Following Exposure to Ionizing Radiation in Human Microvascular Endothelial Cells', *Journal of Radiation Research*, 47(3/4), pp. 245–257. Available at: https://doi.org/10.1269/jrr.0628.

Kawane, K. and Nagata, S. (2008) 'Nucleases in programmed cell death', *Methods in enzymology*, 442, pp. 271–287. Available at: https://doi.org/10.1016/S0076-6879(08)01414-6.

Kebschull, J.M. and Zador, A.M. (2015) 'Sources of PCR-induced distortions in high-throughput sequencing data sets', *Nucleic Acids Research*, 43(21), pp. e143–e143. Available at: https://doi.org/10.1093/NAR/GKV717.

Kennedy, S.R. *et al.* (2014) 'Detecting ultralow-frequency mutations by Duplex Sequencing', *Nature Protocols 2014 9:11*, 9(11), pp. 2586–2606. Available at: https://doi.org/10.1038/nprot.2014.170.

Kent, L.N. and Leone, G. (2019) 'The broken cycle: E2F dysfunction in cancer', *Nature Reviews Cancer*. Nature Publishing Group, pp. 326–338. Available at: https://doi.org/10.1038/s41568-019-0143-7.

Kilpinen, H. *et al.* (2017) 'Common genetic variation drives molecular heterogeneity in human iPSCs', *Nature 2017 546:7658*, 546(7658), pp. 370–375. Available at: https://doi.org/10.1038/nature22403.

Kim, D. *et al.* (2009) 'Generation of Human Induced Pluripotent Stem Cells by Direct Delivery of Reprogramming Proteins', *Cell Stem Cell*, 4(6), pp. 472–476. Available at:

https://doi.org/10.1016/j.stem.2009.05.005.

Kim, H.S. *et al.* (2017) 'Endonuclease EEPD1 Is a gatekeeper for repair of stressed replication forks', *Journal of Biological Chemistry*, 292(7), pp. 2795–2804. Available at: https://doi.org/10.1074/jbc.M116.758235.

Kim, K. *et al.* (2010) 'Epigenetic memory in induced pluripotent stem cells', *Nature*, 467(7313), pp. 285–290. Available at: https://doi.org/10.1038/nature09342.

Kim, S.H. *et al.* (2011) 'Usefulness of a break-apart FISH assay in the diagnosis of Xp11.2 translocation renal cell carcinoma', *Virchows Archiv : an international journal of pathology*, 459(3), pp. 299–306. Available at: https://doi.org/10.1007/S00428-011-1127-5.

Kimura, H. (2013) 'Histone modifications for human epigenome analysis', *Journal of human genetics*, 58(7), pp. 439–445. Available at: https://doi.org/10.1038/JHG.2013.66.

Kinoshita, M. *et al.* (2021) 'Capture of Mouse and Human Stem Cells with Features of Formative Pluripotency', *Cell Stem Cell*, 28(3), pp. 453-471.e8. Available at: https://doi.org/10.1016/j.stem.2020.11.005.

Kitsera, N. *et al.* (2019) 'Nucleotide excision repair of abasic DNA lesions', *Nucleic acids research*, 47(16), pp. 8537–8547. Available at: https://doi.org/10.1093/NAR/GKZ558.

Klein, I.A. *et al.* (2011) 'Translocation-capture sequencing reveals the extent and nature of chromosomal rearrangements in B lymphocytes', *Cell*, 147(1), pp. 95–106. Available at: https://doi.org/10.1016/j.cell.2011.07.048.

Kleinsmith, L.J. and Plerce, G.B. (1964) 'MULTIPOTENTIALITY OF SINGLE EMBRYONAL CARCINOMA CELLS.', *Cancer research* [Preprint].

Koga, K., Wang, B. and Kaneko, S. (2020) 'Current status and future perspectives of HLA-edited induced pluripotent stem cells', *Inflammation and Regeneration*, 40(1), pp. 1–6. Available at: https://doi.org/10.1186/S41232-020-00132-9/FIGURES/1.

Kojima, Y. *et al.* (2014) 'The transcriptional and functional properties of mouse epiblast stem cells resemble the anterior primitive streak', *Cell Stem Cell*, 14(1), pp. 107–120. Available at: https://doi.org/10.1016/j.stem.2013.09.014.

Kokkola, A. *et al.* (1997) '17q12-21 amplicon, a novel recurrent genetic change in intestinal type of gastric carcinoma: A comparative genomic hybridization study', *Genes Chromosomes and Cancer*, 20(1), pp. 38–43. Available at: https://doi.org/10.1002/(SICI)1098-2264(199709)20:1<38::AID-GCC6>3.0.CO;2-A.

König, F., Schubert, T. and Längst, G. (2017) 'The monoclonal S9.6 antibody exhibits highly variable binding affinities towards different R-loop sequences', *PLoS ONE*, 12(6). Available at: https://doi.org/10.1371/journal.pone.0178875.

Konki, M. *et al.* (2016) 'Epigenetic Silencing of the Key Antioxidant Enzyme Catalase in Karyotypically Abnormal Human Pluripotent Stem Cells', *Scientific Reports 2016 6:1*, 6(1), pp. 1–8. Available at: https://doi.org/10.1038/srep22190.

Koren, A., Massey, D.J. and Bracci, A.N. (2021) 'TIGER: inferring DNA replication timing from whole-genome sequence data', *Bioinformatics*. Edited by P. Robinson, 37(22), pp. 4001–4005. Available at: https://doi.org/10.1093/bioinformatics/btab166.

Kotsantis, P. *et al.* (2016) 'Increased global transcription activity as a mechanism of replication stress in cancer', *Nature Communications*, 7. Available at: https://doi.org/10.1038/ncomms13087.

Kouzine, F. *et al.* (2013) 'Transcription-dependent dynamic supercoiling is a short-range genomic force', *Nature Structural and Molecular Biology*, 20(3), pp. 396–403. Available at: https://doi.org/10.1038/nsmb.2517.

Krasilnikova, M.M. and Mirkin, S.M. (2004) ' Replication Stalling at Friedreich's Ataxia (GAA) n Repeats In Vivo ', *Molecular and Cellular Biology*, 24(6), pp. 2286–2295. Available at: https://doi.org/10.1128/mcb.24.6.2286-2295.2004.

Kumar, R. *et al.* (2019) 'HumCFS: a database of fragile sites in human chromosomes', *BMC genomics*, 19(Suppl 9). Available at: https://doi.org/10.1186/S12864-018-5330-5.

Kuo, H.H. *et al.* (2020) 'Negligible-Cost and Weekend-Free Chemically Defined Human iPSC Culture', *Stem cell reports*, 14(2), pp. 256–270. Available at: https://doi.org/10.1016/J.STEMCR.2019.12.007.

Kurek, D. *et al.* (2015) 'Endogenous WNT signals mediate BMP-induced and spontaneous differentiation of epiblast stem cells and human embryonic stem cells', *Stem Cell Reports*, 4(1), pp. 114–128. Available at:

https://doi.org/10.1016/j.stemcr.2014.11.007.

Laco, F. *et al.* (2018) 'Unraveling the Inconsistencies of Cardiac Differentiation Efficiency Induced by the GSK3β Inhibitor CHIR99021 in Human Pluripotent Stem Cells', *Stem Cell Reports*, 10(6), pp. 1851–1866. Available at: https://doi.org/10.1016/j.stemcr.2018.03.023.

Laing, O., Halliwell, J. and Barbaric, I. (2019) 'Rapid PCR Assay for Detecting Common Genetic Variants Arising in Human Pluripotent Stem Cell Cultures', *Current Protocols in Stem Cell Biology*, 49(1), p. e83. Available at: https://doi.org/10.1002/CPSC.83.

Lamm, N. *et al.* (2015) 'Folate levels modulate oncogene-induced replication stress and tumorigenicity', *EMBO Molecular Medicine*, 7(9), pp. 1138–1152. Available at: https://doi.org/10.15252/emmm.201404824.

Langmead, B. and Salzberg, S.L. (2012) 'Fast gapped-read alignment with Bowtie 2', *Nature Methods 2012 9:4*, 9(4), pp. 357–359. Available at: https://doi.org/10.1038/nmeth.1923.

Larsen, B.D. and Sørensen, C.S. (2017) 'The caspase-activated DNase: apoptosis and beyond', *The FEBS Journal*, 284(8), pp. 1160–1170. Available at: https://doi.org/10.1111/FEBS.13970.

Lastowska, M. *et al.* (1997) 'Promiscuous translocations of chromosome arm 17q in human neuroblastomas', *Genes, Chromosomes and Cancer*, 19(3), pp. 143–149. Available at: https://doi.org/10.1002/(SICI)1098-2264(199707)19:3<143::AID-GCC2>3.0.CO;2-Y.

Lawrence, M. *et al.* (2013) 'Software for Computing and Annotating Genomic Ranges', *PLoS Computational Biology*. Edited by A. Prlic, 9(8), p. e1003118. Available at: https://doi.org/10.1371/journal.pcbi.1003118.

Lazzaro, F. *et al.* (2012) 'RNase H and postreplication repair protect cells from ribonucleotides incorporated in DNA', *Molecular Cell*, 45(1), pp. 99–110. Available at: https://doi.org/10.1016/j.molcel.2011.12.019.

Lees, J.G. *et al.* (2019) 'Oxygen regulates human pluripotent stem cell metabolic flux', *Stem Cells International*, 2019. Available at: https://doi.org/10.1155/2019/8195614.

Lensing, S. V. *et al.* (2016) 'DSBCapture: in situ capture and sequencing of DNA

breaks', *Nature Methods 2016 13:10*, 13(10), pp. 855–857. Available at: https://doi.org/10.1038/nmeth.3960.

LeTallec, B. *et al.* (2013) 'Common fragile site profiling in epithelial and erythroid cells reveals that most recurrent cancer deletions lie in fragile sites hosting large genes', *Cell Reports*, 4(3), pp. 420–428. Available at: https://doi.org/10.1016/j.celrep.2013.07.003.

Li, H. and Durbin, R. (2009) 'Fast and accurate short read alignment with Burrows–Wheeler transform', *Bioinformatics*, 25(14), pp. 1754–1760. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTP324.

Li, H. and Stillman, B. (2012) 'The origin recognition complex: A biochemical and structural view', *Subcellular Biochemistry*, 62, pp. 37–58. Available at: https://doi.org/10.1007/978-94-007-4572-8_3.

Li, S. and Wu, X. (2020) 'Common fragile sites: Protection and repair', *Cell and Bioscience*. BioMed Central Ltd. Available at: https://doi.org/10.1186/s13578-020-00392-5.

Lian, X. *et al.* (2012) 'Robust cardiomyocyte differentiation from human pluripotent stem cells via temporal modulation of canonical Wnt signaling', *Proceedings of the National Academy of Sciences of the United States of America*, 109(27), pp. E1848–E1857. Available at: https://doi.org/10.1073/PNAS.1200250109/SUPPL_FILE/COZZARELLICLASSIV-PODCAST.MP3.

Lin, C.Y. *et al.* (2012) 'Transcriptional amplification in tumor cells with elevated c-Myc', *Cell*, 151(1), pp. 56–67. Available at: https://doi.org/10.1016/j.cell.2012.08.026.

Lindahl, T. (1993) 'Instability and decay of the primary structure of DNA', *Nature 1993 362:6422*, 362(6422), pp. 709–715. Available at: https://doi.org/10.1038/362709a0.

Lindahl, T., Karran, P. and Wood, R.D. (1997) 'DNA excision repair pathways', *Current Opinion in Genetics & Development*, 7(2), pp. 158–169. Available at: https://doi.org/10.1016/S0959-437X(97)80124-4.

Lippmann, E.S., Estevez-Silva, M.C. and Ashton, R.S. (2014) 'Defined Human Pluripotent Stem Cell Culture Enables Highly Efficient Neuroepithelium Derivation Without Small Molecule Inhibitors', *Stem Cells*, 32(4), pp. 1032–1042. Available at:

https://doi.org/10.1002/STEM.1622.

Liu, L. *et al.* (2019) 'The cell cycle in stem cell proliferation, pluripotency and differentiation', *Nature Cell Biology*. Nature Publishing Group, pp. 1060–1067. Available at: https://doi.org/10.1038/s41556-019-0384-4.

Liu, L.F. and Wang, J.C. (1987) 'Supercoiling of the DNA template during transcription', *Proceedings of the National Academy of Sciences of the United States of America*, 84(20), pp. 7024–7027. Available at: https://doi.org/10.1073/PNAS.84.20.7024.

Liu, W. *et al.* (2018) 'The Suppression of Medium Acidosis Improves the Maintenance and Differentiation of Human Pluripotent Stem Cells at High Density in Defined Cell Culture Medium', *International Journal of Biological Sciences*, 14(5), pp. 485–496. Available at: https://doi.org/10.7150/ijbs.24681.

Livak, K.J. and Schmittgen, T.D. (2001) 'Analysis of relative gene expression data using real-time quantitative PCR and the 2-ΔΔCT method', *Methods*, 25(4), pp. 402–408. Available at: https://doi.org/10.1006/meth.2001.1262.

Love, M.I., Huber, W. and Anders, S. (2014) 'Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2', *Genome Biology*, 15(12), pp. 1–21. Available at: https://doi.org/10.1186/S13059-014-0550-8/FIGURES/9.

Lucà, S. *et al.* (2023) 'PATZ1 in Non-Small Cell Lung Cancer: A New Biomarker That Negatively Correlates with PD-L1 Expression and Suppresses the Malignant Phenotype', *Cancers*, 15(7), p. 2190. Available at: https://doi.org/10.3390/cancers15072190.

Ludwig, T.E. *et al.* (2006) 'Feeder-independent culture of human embryonic stem cells', *Nature Methods*, 3(8), pp. 637–646. Available at: https://doi.org/10.1038/nmeth902.

Luo, L., Gribskov, M. and Wang, S. (2022) 'Bibliometric review of ATAC-Seq and its application in gene expression', *Briefings in bioinformatics*, 23(3). Available at: https://doi.org/10.1093/BIB/BBAC061.

Lyu, R. *et al.* (2022) 'The specialized mitotic behavior of human embryonic stem cells', *Cell and Tissue Research*, 387(1), pp. 85–93. Available at: https://doi.org/10.1007/s00441-021-03544-2.

Lyu, X., Rowley, M.J. and Corces, V.G. (2018) 'Architectural Proteins and Pluripotency Factors Cooperate to Orchestrate the Transcriptional Response of hESCs to Temperature Stress', *Molecular Cell*, 71(6), pp. 940-955.e7. Available at: https://doi.org/10.1016/J.MOLCEL.2018.07.012.

Ma, J. and Wang, M.D. (2016) 'DNA supercoiling during transcription', *Biophysical Reviews*. Springer Verlag, pp. 75–87. Available at: https://doi.org/10.1007/s12551-016-0215-9.

Macheret, M. and Halazonetis, T.D. (2018) 'Intragenic origins due to short G1 phases underlie oncogene-induced DNA replication stress', *Nature*, 555(7694), pp. 112–116. Available at: https://doi.org/10.1038/nature25507.

Maiti, A. and Drohat, A.C. (2011) 'Thymine DNA Glycosylase Can Rapidly Excise 5-Formylcytosine and 5-Carboxylcytosine: POTENTIAL IMPLICATIONS FOR ACTIVE DEMETHYLATION OF CpG SITES', *Journal of Biological Chemistry*, 286(41), pp. 35334–35338. Available at: https://doi.org/10.1074/JBC.C111.284620.

Mali, P. *et al.* (2013) 'RNA-Guided Human Genome Engineering via Cas9', *Science (New York, N.Y.)*, 339(6121), p. 823. Available at: https://doi.org/10.1126/SCIENCE.1232033.

Mandai, M. *et al.* (2017) 'Autologous Induced Stem-Cell–Derived Retinal Cells for Macular Degeneration', *New England Journal of Medicine*, 376(11), pp. 1038–1046. Available at: https://doi.org/10.1056/NEJMOA1608368/SUPPL_FILE/NEJMOA1608368_DISCLOSURES.PDF.

Mankouri, H.W., Huttner, D. and Hickson, I.D. (2013) 'How unfinished business from S-phase affects mitosis and beyond', *EMBO Journal*. EMBO J, pp. 2661–2671. Available at: https://doi.org/10.1038/emboj.2013.211.

Manstein, F. *et al.* (2021) 'High density bioprocessing of human pluripotent stem cells by metabolic control and in silico modeling', *Stem Cells Translational Medicine*, 10(7), pp. 1063–1080. Available at: https://doi.org/10.1002/sctm.20-0453.

Marión, R.M. *et al.* (2009) 'A p53-mediated DNA damage response limits reprogramming to ensure iPS cell genomic integrity', *Nature*, 460(7259), pp. 1149–1153. Available at: https://doi.org/10.1038/nature08287.

Markouli, C. *et al.* (2019) ‘Gain of 20q11.21 in Human Pluripotent Stem Cells Impairs TGF-β-Dependent Neuroectodermal Commitment’, *Stem Cell Reports*, 13(1), pp. 163–176. Available at: https://doi.org/10.1016/j.stemcr.2019.05.005.

Mascetti, V.L. and Pedersen, R.A. (2016) ‘Contributions of Mammalian Chimeras to Pluripotent Stem Cell Research’, *Cell Stem Cell*. Cell Press, pp. 163–175. Available at: https://doi.org/10.1016/j.stem.2016.07.018.

Mathieu, J. *et al.* (2014) ‘Hypoxia-inducible factors have distinct and stage-specific roles during reprogramming of human cells to pluripotency’, *Cell Stem Cell*, 14(5), pp. 592–605. Available at: https://doi.org/10.1016/j.stem.2014.02.012.

Maynard, S. *et al.* (2008) ‘Human Embryonic Stem Cells Have Enhanced Repair of Multiple Forms of DNA Damage’, *Stem Cells*, 26(9), pp. 2266–2274. Available at: https://doi.org/10.1634/stemcells.2007-1041.

Mayshar, Y., Yanuka, O. and Benvenisty, N. (2011) ‘Teratogen screening using transcriptome profiling of differentiating human embryonic stem cells’, *Journal of cellular and molecular medicine*, 15(6), pp. 1393–1401. Available at: https://doi.org/10.1111/J.1582-4934.2010.01105.X.

McCulloch, S.D. and Kunkel, T.A. (2008) ‘The fidelity of DNA synthesis by eukaryotic replicative and translesion synthesis polymerases’, *Cell Research*. Cell Res, pp. 148–161. Available at: https://doi.org/10.1038/cr.2008.4.

Mehta, A. and Haber, J.E. (2014) ‘Sources of DNA Double-Strand Breaks and Models of Recombinational DNA Repair’, *Cold Spring Harbor Perspectives in Biology*, 6(9). Available at: https://doi.org/10.1101/CSHPERSPECT.A016428.

Meléndez-Ramírez, C. *et al.* (2021) ‘Dynamic landscape of chromatin accessibility and transcriptomic changes during differentiation of human embryonic stem cells into dopaminergic neurons’, *Scientific Reports*, 11(1), p. 16977. Available at: https://doi.org/10.1038/s41598-021-96263-1.

Méndez, J. *et al.* (2002) ‘Human origin recognition complex large subunit is degraded by ubiquitin-mediated proteolysis after initiation of DNA replication’, *Molecular Cell*, 9(3), pp. 481–491. Available at: https://doi.org/10.1016/S1097-2765(02)00467-7.

Mennen, R.H., Oldenburger, M.M. and Piersma, A.H. (2022) ‘Endoderm and mesoderm derivatives in embryonic stem cell differentiation and their use in

developmental toxicity testing', *Reproductive Toxicology*, 107, pp. 44–59. Available at: https://doi.org/10.1016/J.REPROTOX.2021.11.009.

Merkle, F.T. *et al.* (2017) 'Human pluripotent stem cells recurrently acquire and expand dominant negative P53 mutations', *Nature*, 545(7653), pp. 229–233. Available at: https://doi.org/10.1038/nature22312.

Meryet-Figuiere, M. *et al.* (2014) 'Temporal separation of replication and transcription during S-phase progression', *Cell cycle (Georgetown, Tex.)*, 13(20), pp. 3241–3248. Available at: https://doi.org/10.4161/15384101.2014.953876.

Michel, N. *et al.* (2022) 'Transcription-associated DNA DSBs activate p53 during hiPSC-based neurogenesis', *Scientific Reports 2022 12:1*, 12(1), pp. 1–14. Available at: https://doi.org/10.1038/s41598-022-16516-5.

Miller, T.C.R. *et al.* (2019) 'Mechanism of head-to-head MCM double-hexamer formation revealed by cryo-EM', *Nature*, 575(7784), pp. 704–710. Available at: https://doi.org/10.1038/s41586-019-1768-0.

Min, J. *et al.* (2023) 'Mechanisms of insertions at a DNA double-strand break', *Molecular cell* [Preprint]. Available at: https://doi.org/10.1016/J.MOLCEL.2023.06.016.

Montgomery, S.B. *et al.* (2013) 'The origin, evolution, and functional impact of short insertion–deletion variants identified in 179 human genomes', *Genome Research*, 23(5), p. 749. Available at: https://doi.org/10.1101/GR.148718.112.

Montilla-Rojo, J. *et al.* (2023) 'Teratoma Assay for Testing Pluripotency and Malignancy of Stem Cells: Insufficient Reporting and Uptake of Animal-Free Methods—A Systematic Review', *International Journal of Molecular Sciences*, 24(4), p. 3879. Available at: https://doi.org/10.3390/IJMS24043879/S1.

Moore, L.D., Le, T. and Fan, G. (2012) 'DNA Methylation and Its Basic Function', *Neuropsychopharmacology 2013 38:1*, 38(1), pp. 23–38. Available at: https://doi.org/10.1038/npp.2012.112.

Morimoto, S. *et al.* (2019) 'Type II DNA Topoisomerases Cause Spontaneous Double-Strand Breaks in Genomic DNA', *Genes*, 10(11). Available at: https://doi.org/10.3390/GENES10110868.

Moris, N. *et al.* (2020) 'An in vitro model of early anteroposterior organization during

human development', *Nature 2020 582:7812*, 582(7812), pp. 410–415. Available at: https://doi.org/10.1038/s41586-020-2383-9.

Mourad, R. *et al.* (2018) 'Predicting double-strand DNA breaks using epigenome marks or DNA at kilobase resolution', *Genome Biology*, 19(1), pp. 1–14. Available at: https://doi.org/10.1186/S13059-018-1411-7/TABLES/2.

Mummery, C.L. *et al.* (1987) 'Cell cycle analysis during retinoic acid induced differentiation of a human embryonal carcinoma-derived cell line', *Cell Differentiation*, 20(2–3), pp. 153–160. Available at: https://doi.org/10.1016/0045-6039(87)90429-5.

Nagata, S. *et al.* (2003) 'Degradation of chromosomal DNA during apoptosis', *Cell death and differentiation*, 10(1), pp. 108–116. Available at: https://doi.org/10.1038/SJ.CDD.4401161.

Naidoo, K. *et al.* (2018) 'Evaluation of CDK12 Protein Expression as a Potential Novel Biomarker for DNA Damage Response-Targeted Therapies in Breast Cancer', *Molecular cancer therapeutics*, 17(1), pp. 306–315. Available at: https://doi.org/10.1158/1535-7163.MCT-17-0760.

Närvä, E. *et al.* (2010) 'High-resolution DNA analysis of human embryonic stem cell lines reveals culture-induced copy number changes and loss of heterozygosity', *Nature Biotechnology*, 28(4), pp. 371–377. Available at: https://doi.org/10.1038/nbt.1615.

Närvä, E. *et al.* (2013) 'Continuous hypoxic culturing of human embryonic stem cells enhances SSEA-3 and MYC levels', *PLoS ONE*, 8(11). Available at: https://doi.org/10.1371/journal.pone.0078847.

Nasto, L.A. *et al.* (2013) 'Mitochondrial-derived reactive oxygen species (ROS) play a causal role in aging-related intervertebral disc degeneration', *Journal of orthopaedic research : official publication of the Orthopaedic Research Society*, 31(7), pp. 1150–1157. Available at: https://doi.org/10.1002/JOR.22320.

Natrajan, R. *et al.* (2009) 'Tiling path genomic profiling of grade 3 invasive ductal breast cancers', *Clinical Cancer Research*, 15(8), pp. 2711–2722. Available at: https://doi.org/10.1158/1078-0432.CCR-08-1878.

Negrini, S., Gorgoulis, V.G. and Halazonetis, T.D. (2010) 'Genomic instability an evolving hallmark of cancer', *Nature Reviews Molecular Cell Biology*. Nature

Publishing Group, pp. 220–228. Available at: https://doi.org/10.1038/nrm2858.

Nguyen, H.T. *et al.* (2014) 'Gain of 20q11.21 in human embryonic stem cells improves cell survival by increased expression of Bcl-xL', *Molecular Human Reproduction*, 20(2), pp. 168–177. Available at: https://doi.org/10.1093/MOLEHR/GAT077.

Niakan, K.K. *et al.* (2012) 'Human pre-implantation embryo development', *Development (Cambridge, England)*, 139(5), pp. 829–841. Available at: https://doi.org/10.1242/DEV.060426.

Nichols, J. *et al.* (1998) 'Formation of pluripotent stem cells in the mammalian embryo depends on the POU transcription factor Oct4', *Cell*, 95(3), pp. 379–391. Available at: https://doi.org/10.1016/S0092-8674(00)81769-9.

Nichols, J. and Smith, A. (2009) 'Naive and Primed Pluripotent States', *Cell Stem Cell*, 4(6), pp. 487–492. Available at: https://doi.org/10.1016/j.stem.2009.05.015.

Nick McElhinny, S.A. *et al.* (2010) 'Abundant ribonucleotide incorporation into DNA by yeast replicative polymerases', *Proceedings of the National Academy of Sciences of the United States of America*, 107(11), pp. 4949–4954. Available at: https://doi.org/10.1073/pnas.0914857107.

Nickoloff, J.A. *et al.* (2021) 'The Safe Path at the Fork: Ensuring Replication-Associated DNA Double-Strand Breaks are Repaired by Homologous Recombination', *Frontiers in Genetics*, 12, p. 1887. Available at: https://doi.org/10.3389/FGENE.2021.748033/BIBTEX.

Niehrs, C. and Luke, B. (2020) 'Regulatory R-loops as facilitators of gene expression and genome stability', *Nature Reviews Molecular Cell Biology*. Nature Research, pp. 167–178. Available at: https://doi.org/10.1038/s41580-019-0206-3.

Nimonkar, A. V. *et al.* (2011) 'BLM–DNA2–RPA–MRN and EXO1–BLM–RPA–MRN constitute two DNA end resection machineries for human DNA break repair', *Genes & Development*, 25(4), pp. 350–362. Available at: https://doi.org/10.1101/GAD.2003811.

Nit, K., Tyszka-Czochara, M. and Bobis-Wozowicz, S. (2021) 'Oxygen as a Master Regulator of Human Pluripotent Stem Cell Function and Metabolism', *Journal of Personalized Medicine 2021, Vol. 11, Page 905*, 11(9), p. 905. Available at: https://doi.org/10.3390/JPM11090905.

Noe Gonzalez, M., Blears, D. and Svejstrup, J.Q. (2021) 'Causes and consequences of RNA polymerase II stalling during transcript elongation', *Nature reviews. Molecular cell biology*, 22(1), pp. 3–21. Available at: https://doi.org/10.1038/S41580-020-00308-8.

O'Leary, N.A. *et al.* (2016) 'Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation', *Nucleic acids research*, 44(D1), pp. D733–D745. Available at: https://doi.org/10.1093/NAR/GKV1189.

Ohle, C. *et al.* (2016) 'Transient RNA-DNA Hybrids Are Required for Efficient Double-Strand Break Repair', *Cell*, 167(4), pp. 1001-1013.e7. Available at: https://doi.org/10.1016/j.cell.2016.10.001.

Okita, K. *et al.* (2008) 'Generation of mouse induced pluripotent stem cells without viral vectors', *Science (New York, N.Y.)*, 322(5903), pp. 949–953. Available at: https://doi.org/10.1126/SCIENCE.1164270.

Olariu, V. *et al.* (2010) 'Modeling the evolution of culture-adapted human embryonic stem cells', *Stem Cell Research*, 4(1), pp. 50–56. Available at: https://doi.org/10.1016/j.scr.2009.09.001.

Ostling, O. and Johanson, K.J. (1984) 'Microelectrophoretic study of radiation-induced DNA damages in individual mammalian cells', *Biochemical and biophysical research communications*, 123(1), pp. 291–298. Available at: https://doi.org/10.1016/0006-291X(84)90411-X.

Owen-Smith, J., Scott, C.T. and Mccormick, J.B. (2012) 'Expand and Regularize Federal Funding for Human Pluripotent Stem Cell Research', *Journal of policy analysis and management : [the journal of the Association for Public Policy Analysis and Management]*, 31(3), p. 714. Available at: https://doi.org/10.1002/PAM.21607.

Paeschke, K., Capra, J.A. and Zakian, V.A. (2011) 'DNA Replication through G-Quadruplex Motifs Is Promoted by the Saccharomyces cerevisiae Pif1 DNA Helicase', *Cell*, 145(5), pp. 678–691. Available at: https://doi.org/10.1016/j.cell.2011.04.015.

Palmerola, K.L. *et al.* (2022) 'Replication stress impairs chromosome segregation and preimplantation development in human embryos', *Cell*, 0(0). Available at: https://doi.org/10.1016/J.CELL.2022.06.028.

Papaioannou, V.E. *et al.* (1975) 'Fate of teratocarcinoma cells injected into early

mouse embryos', *Nature*, 258(5530), pp. 70–73. Available at: https://doi.org/10.1038/258070A0.

Park, P.J. (2009) 'ChIP-Seq: advantages and challenges of a maturing technology', *Nature reviews. Genetics*, 10(10), p. 669. Available at: https://doi.org/10.1038/NRG2641.

Paull, T.T. *et al.* (2000) 'A critical role for histone H2AX in recruitment of repair factors to nuclear foci after DNA damage', *Current Biology*, 10(15), pp. 886–895. Available at: https://doi.org/10.1016/S0960-9822(00)00610-2.

Pepe, A. and West, S.C. (2014) 'MUS81-EME2 Promotes Replication Fork Restart', *Cell Reports*, 7(4), pp. 1048–1055. Available at: https://doi.org/10.1016/J.CELREP.2014.04.007.

Petermann, E. *et al.* (2010) 'Hydroxyurea-Stalled Replication Forks Become Progressively Inactivated and Require Two Different RAD51-Mediated Pathways for Restart and Repair', *Molecular Cell*, 37(4), pp. 492–502. Available at: https://doi.org/10.1016/j.molcel.2010.01.021.

Petryk, N. *et al.* (2016) 'Replication landscape of the human genome', *Nature Communications*, 7(1), pp. 1–13. Available at: https://doi.org/10.1038/ncomms10208.

Pick, M. *et al.* (2009) 'Clone- and Gene-Specific Aberrations of Parental Imprinting in Human Induced Pluripotent Stem Cells', *STEM CELLS*, 27(11), pp. 2686–2690. Available at: https://doi.org/10.1002/STEM.205.

Polo, S.E. and Jackson, S.P. (2011) 'Dynamics of DNA damage response proteins at DNA breaks: a focus on protein modifications', *Genes & Development*, 25(5), p. 409. Available at: https://doi.org/10.1101/GAD.2021311.

Pommier, Y. (2006) 'Topoisomerase I inhibitors: camptothecins and beyond', *Nature Reviews Cancer 2006 6:10*, 6(10), pp. 789–802. Available at: https://doi.org/10.1038/nrc1977.

Pommier, Y. *et al.* (2016) 'Roles of eukaryotic topoisomerases in transcription, replication and genomic stability', *Nature reviews. Molecular cell biology*, 17(11), pp. 703–721. Available at: https://doi.org/10.1038/NRM.2016.111.

Pommier, Y. *et al.* (2022) 'Human topoisomerases and their roles in genome stability and organization', *Nature Reviews Molecular Cell Biology*. Nature Research, pp. 407–

427. Available at: https://doi.org/10.1038/s41580-022-00452-3.

Prendergast, L. *et al.* (2020) 'Resolution of R-loops by INO80 promotes DNA replication and maintains cancer cell proliferation and viability', *Nature Communications*, 11(1). Available at: https://doi.org/10.1038/s41467-020-18306-x.

Price, C.J. *et al.* (2021) 'Genetically variant human pluripotent stem cells selectively eliminate wild-type counterparts through YAP-mediated cell competition', *Developmental cell*, 56(17), pp. 2455-2470.e10. Available at: https://doi.org/10.1016/J.DEVCEL.2021.07.019.

Promonet, A. *et al.* (2020) 'Topoisomerase 1 prevents replication stress at R-loop-enriched transcription termination sites', *Nature Communications*, 11(1). Available at: https://doi.org/10.1038/s41467-020-17858-2.

Qiu, S. *et al.* (2021) 'Replication Fork Reversal and Protection', *Frontiers in Cell and Developmental Biology*. Frontiers Media S.A., p. 670392. Available at: https://doi.org/10.3389/fcell.2021.670392.

Quinlan, A.R. and Hall, I.M. (2010) 'BEDTools: a flexible suite of utilities for comparing genomic features', *Bioinformatics (Oxford, England)*, 26(6), pp. 841–842. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTQ033.

Ramírez, F. *et al.* (2016) 'deepTools2: a next generation web server for deep-sequencing data analysis', *Nucleic Acids Research*, 44(W1), pp. W160–W165. Available at: https://doi.org/10.1093/NAR/GKW257.

Rao, S.S.P. *et al.* (2014) 'A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping', *Cell*, 159(7), pp. 1665–1680. Available at: https://doi.org/10.1016/j.cell.2014.11.021.

Ray, S. *et al.* (2022) 'A mechanism for oxidative damage repair at gene regulatory elements', *Nature 2022 609:7929*, 609(7929), pp. 1038–1047. Available at: https://doi.org/10.1038/s41586-022-05217-8.

Reimers, M.A. *et al.* (2020) 'Clinical Outcomes in Cyclin-dependent Kinase 12 Mutant Advanced Prostate Cancer', *European urology*, 77(3), pp. 333–341. Available at: https://doi.org/10.1016/J.EURURO.2019.09.036.

Rizzo, F. *et al.* (2016) 'Selective mitochondrial depletion, apoptosis resistance, and increased mitophagy in human Charcot-Marie-Tooth 2A motor neurons', *Human*

molecular genetics, 25(19), pp. 4266–4281. Available at: https://doi.org/10.1093/HMG/DDW258.

Roberts, S.A. *et al.* (2012) 'Clustered mutations in yeast and in human cancers can arise from damaged long single-strand DNA regions', *Molecular cell*, 46(4), pp. 424–435. Available at: https://doi.org/10.1016/J.MOLCEL.2012.03.030.

Robinson, J.T. *et al.* (2011) 'Integrative genomics viewer', *Nature Biotechnology*. Nature Publishing Group, pp. 24–26. Available at: https://doi.org/10.1038/nbt.1754.

Rogakou, E.P. *et al.* (1998) 'DNA double-stranded breaks induce histone H2AX phosphorylation on serine 139', *Journal of Biological Chemistry*, 273(10), pp. 5858–5868. Available at: https://doi.org/10.1074/jbc.273.10.5858.

Ross-Innes, C.S. *et al.* (2012) 'Differential oestrogen receptor binding is associated with clinical outcome in breast cancer', *Nature*, 481(7381), pp. 389–393. Available at: https://doi.org/10.1038/NATURE10730.

Rouhani, F.J. *et al.* (2016) 'Mutational History of a Human Cell Lineage from Somatic to Induced Pluripotent Stem Cells', *PLoS genetics*, 12(4). Available at: https://doi.org/10.1371/JOURNAL.PGEN.1005932.

Rouhani, F.J. *et al.* (2022) 'Substantial somatic genomic variation and selection for BCOR mutations in human induced pluripotent stem cells', *Nature Genetics 2022 54:9*, 54(9), pp. 1406–1416. Available at: https://doi.org/10.1038/s41588-022-01147-3.

Roukos, V. and Misteli, T. (2014) 'The biogenesis of chromosome translocations', *Nature cell biology*, 16(4), p. 293. Available at: https://doi.org/10.1038/NCB2941.

Rowe, R.G. and Daley, G.Q. (2019) 'Induced pluripotent stem cells in disease modelling and drug discovery', *Nature Reviews Genetics 2019 20:7*, 20(7), pp. 377–388. Available at: https://doi.org/10.1038/s41576-019-0100-z.

Rozowsky, J. *et al.* (2009) 'PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls', *Nature Biotechnology 2009 27:1*, 27(1), pp. 66–75. Available at: https://doi.org/10.1038/nbt.1518.

Ruiz, S. *et al.* (2015) 'Limiting replication stress during somatic cell reprogramming reduces genomic instability in induced pluripotent stem cells', *Nature Communications*, 6(1), p. 8036. Available at: https://doi.org/10.1038/ncomms9036.

Rybin, M.J. *et al.* (2021) ‘Emerging Technologies for Genome-Wide Profiling of DNA Breakage’, *Frontiers in Genetics*. Frontiers Media S.A., p. 610386. Available at: https://doi.org/10.3389/fgene.2020.610386.

Saayman, X. and Esashi, F. (2022) ‘Breaking the paradigm: early insights from mammalian DNA breakomes’, *The FEBS journal*, 289(9), pp. 2409–2428. Available at: https://doi.org/10.1111/FEBS.15849.

Safari, M. *et al.* (2021) ‘R-Loop–Mediated ssDNA Breaks Accumulate following Short-Term Exposure to the HDAC Inhibitor Romidepsin’, *Molecular Cancer Research*, 19(8), pp. 1361–1374. Available at: https://doi.org/10.1158/1541-7786.MCR-20-0833.

Salaverria, I. *et al.* (2008) ‘Chromosomal alterations detected by comparative genomic hybridization in subgroups of gene expression-defined Burkitt’s lymphoma’, *Haematologica*, 93(9), pp. 1327–1334. Available at: https://doi.org/10.3324/haematol.13071.

Sameshima, Y. *et al.* (1994) ‘Frequent Allelic Losses on Chromosomes 2q, 18q, and 22q in Advanced Non-Small Cell Lung Carcinoma’, *Cancer Research*, 54(21), pp. 5643–5648.

Sanz, L.A. and Chédin, F. (2019) ‘High-resolution, strand-specific R-loop mapping via S9.6-basedDNA:RNA ImmunoPrecipitation and high-throughput sequencing.’, *Nature protocols*, 14(6), p. 1734. Available at: https://doi.org/10.1038/S41596-019-0159-1.

Sartori, A.A. *et al.* (2007) ‘Human CtIP promotes DNA end resection’, *Nature 2007 450:7169*, 450(7169), pp. 509–514. Available at: https://doi.org/10.1038/nature06337.

Sawyer, J.R. *et al.* (1995) ‘Cytogenetic findings in 200 patients with multiple myeloma’, *Cancer Genetics and Cytogenetics*, 82(1), pp. 41–49. Available at: https://doi.org/10.1016/0165-4608(94)00284-I.

Schaarschmidt, D. *et al.* (2004) ‘An episomal mammalian replicon: Sequence-independent binding of the origin recognition complex’, *EMBO Journal*, 23(1), pp. 191–201. Available at: https://doi.org/10.1038/sj.emboj.7600029.

Schindelin, J. *et al.* (2012) ‘Fiji: an open-source platform for biological-image analysis’, *Nature Methods 2012 9:7*, 9(7), pp. 676–682. Available at: https://doi.org/10.1038/nmeth.2019.

Schultz, L.B. *et al.* (2000) ‘p53 binding protein 1 (53BP1) is an early participant in the

cellular response to DNA double-strand breaks', *Journal of Cell Biology*, 151(7), pp. 1381–1390. Available at: https://doi.org/10.1083/jcb.151.7.1381.

Sedletska, Y., Radicella, J.P. and Sage, E. (2013) 'Replication fork collapse is a major cause of the high mutation frequency at three-base lesion clusters', *Nucleic acids research*, 41(20), pp. 9339–9348. Available at: https://doi.org/10.1093/NAR/GKT731.

Sela, Y. *et al.* (2012) 'Human Embryonic Stem Cells Exhibit Increased Propensity to Differentiate During the G1 Phase Prior to Phosphorylation of Retinoblastoma Protein', *STEM CELLS*, 30(6), pp. 1097–1108. Available at: https://doi.org/10.1002/stem.1078.

Seltmann, S. *et al.* (2016) 'hPSCreg—the human pluripotent stem cell registry', *Nucleic Acids Research*, 44(Database issue), p. D757. Available at: https://doi.org/10.1093/NAR/GKV963.

Sen, T. and Thummer, R.P. (2022) 'CRISPR and iPSCs: Recent Developments and Future Perspectives in Neurodegenerative Disease Modelling, Research, and Therapeutics', *Neurotoxicity Research*, 40(5), p. 1597. Available at: https://doi.org/10.1007/S12640-022-00564-W.

Serra, M. *et al.* (2012) 'Process engineering of human pluripotent stem cells for clinical application', *Trends in biotechnology*, 30(6), pp. 350–359. Available at: https://doi.org/10.1016/J.TIBTECH.2012.03.003.

Simara, P. *et al.* (2017) 'DNA double-strand breaks in human induced pluripotent stem cell reprogramming and long-term in vitro culturing', *Stem Cell Research and Therapy*, 8(1), pp. 1–13. Available at: https://doi.org/10.1186/S13287-017-0522-5/FIGURES/6.

Sinai, M.I.T. *et al.* (2019) 'AT-dinucleotide rich sequences drive fragile site formation', *Nucleic acids research*, 47(18), pp. 9685–9695. Available at: https://doi.org/10.1093/NAR/GKZ689.

Singh, A.M. and Dalton, S. (2009) 'The Cell Cycle and Myc Intersect with Mechanisms that Regulate Pluripotency and Reprogramming', *Cell Stem Cell*, pp. 141–149. Available at: https://doi.org/10.1016/j.stem.2009.07.003.

Singh, Sandeep *et al.* (2020) 'Pausing sites of RNA polymerase II on actively transcribed genes are enriched in DNA double-stranded breaks', *Journal of Biological Chemistry*, 295(12), pp. 3990–4000. Available at: https://doi.org/10.1074/JBC.RA119.011665.

Singh, Sheetal, Shih, S.J. and Vaughan, A.T.M. (2020) ‘Detection of DNA double-strand breaks and chromosome translocations using ligation-mediated PCR and inverse PCR’, *Methods in Molecular Biology*, 2102, pp. 271–288. Available at: https://doi.org/10.1007/978-1-0716-0223-2_15/FIGURES/4.

Skotheim, R.I. *et al.* (2002) ‘New insights into testicular germ cell tumorigenesis from gene expression profiling.’, *Cancer research*, 62(8), pp. 2359–64. Available at: http://www.ncbi.nlm.nih.gov/pubmed/11956097 (Accessed: 26 November 2019).

Slimen, I.B. *et al.* (2014) ‘Reactive oxygen species, heat stress and oxidative-induced mitochondrial damage. A review’, *International journal of hyperthermia : the official journal of European Society for Hyperthermic Oncology, North American Hyperthermia Group*, 30(7), pp. 513–523. Available at: https://doi.org/10.3109/02656736.2014.971446.

Sollier, J. *et al.* (2014) ‘Transcription-Coupled Nucleotide Excision Repair Factors Promote R-Loop-Induced Genome Instability’, *Molecular Cell*, 56(6), pp. 777–785. Available at: https://doi.org/10.1016/J.MOLCEL.2014.10.020.

Sollier, J. and Cimprich, K.A. (2015) ‘Breaking bad: R-loops and genome integrity’, *Trends in Cell Biology*, 25(9), pp. 514–522. Available at: https://doi.org/10.1016/j.tcb.2015.05.003.

Sondka, Z. *et al.* (2018) ‘The COSMIC Cancer Gene Census: describing genetic dysfunction across all human cancers’, *Nature Reviews Cancer 2018 18:11*, 18(11), pp. 696–705. Available at: https://doi.org/10.1038/s41568-018-0060-1.

Song, J. and Chen, K.C. (2015) ‘Spectacle: Fast chromatin state annotation using spectral learning’, *Genome Biology*, 16(1), pp. 1–18. Available at: https://doi.org/10.1186/S13059-015-0598-0/COMMENTS.

Speck, C. *et al.* (2005) ‘ATPase-dependent cooperative binding of ORC and Cdc6 to origin DNA’, *Nature Structural and Molecular Biology*, 12(11), pp. 965–971. Available at: https://doi.org/10.1038/nsmb1002.

spicuglia, salvatore and Vanhille, L. (2012) ‘Chromatin signatures of active enhancers’, *Nucleus (Austin, Tex.)*, 3(2), pp. 126–131. Available at: https://doi.org/10.4161/NUCL.19232.

Spits, C. *et al.* (2008) ‘Recurrent chromosomal abnormalities in human embryonic

stem cells', *Nature Biotechnology*, 26(12), pp. 1361–1363. Available at: https://doi.org/10.1038/nbt.1510.

Stadtfeld, M. *et al.* (2008) 'Induced pluripotent stem cells generated without viral integration', *Science (New York, N.Y.)*, 322(5903), pp. 945–949. Available at: https://doi.org/10.1126/SCIENCE.1162494.

Stafford, N. (2008) 'Germany liberalises law on stem cell research', *BMJ : British Medical Journal*, 336(7649), p. 851. Available at: https://doi.org/10.1136/BMJ.39552.538356.DB.

Starmer, J. and Magnuson, T. (2016) 'Detecting broad domains and narrow peaks in ChIP-seq data with hiddenDomains', *BMC Bioinformatics*, 17(1), pp. 1–10. Available at: https://doi.org/10.1186/S12859-016-0991-Z/FIGURES/4.

Stirling, D.R. *et al.* (2021) 'CellProfiler 4: improvements in speed, utility and usability', *BMC Bioinformatics*, 22(1), pp. 1–11. Available at: https://doi.org/10.1186/S12859-021-04344-9/FIGURES/6.

Stork, C.T. *et al.* (2016) 'Co-transcriptional R-loops are the main cause of estrogen-induced DNA damage', *eLife*, 5(AUGUST). Available at: https://doi.org/10.7554/eLife.17548.

Strano, A. *et al.* (2020) 'Variable Outcomes in Neural Differentiation of Human PSCs Arise from Intrinsic Differences in Developmental Signaling Pathways', *Cell Reports*, 31(10). Available at: https://doi.org/10.1016/J.CELREP.2020.107732.

Streubel, G. *et al.* (2017) 'Fam60a defines a variant Sin3a-Hdac complex in embryonic stem cells required for self-renewal', *The EMBO Journal*, 36(15), pp. 2216–2232. Available at: https://doi.org/10.15252/EMBJ.201696307.

Strumberg, D. *et al.* (2000) 'Conversion of Topoisomerase I Cleavage Complexes on the Leading Strand of Ribosomal DNA into 5′-Phosphorylated DNA Double-Strand Breaks by Replication Runoff', *Molecular and Cellular Biology*, 20(11), p. 3977. Available at: https://doi.org/10.1128/MCB.20.11.3977-3987.2000.

Sun, Y. *et al.* (2023) 'A graph neural network-based interpretable framework reveals a novel DNA fragility–associated chromatin structural unit', *Genome Biology*, 24(1), p. 90. Available at: https://doi.org/10.1186/S13059-023-02916-X.

Szlachta, K. *et al.* (2020) 'Topoisomerase II contributes to DNA secondary structure-

mediated double-stranded breaks', *Nucleic Acids Research*, 48(12), p. 6654. Available at: https://doi.org/10.1093/NAR/GKAA483.

Taapken, S.M. *et al.* (2011) 'Karotypic abnormalities in human induced pluripotent stem cells and embryonic stem cells', *Nature biotechnology*, 29(4), pp. 313–314. Available at: https://doi.org/10.1038/NBT.1835.

Taei, A. *et al.* (2020) 'Signal regulators of human naïve pluripotency', *Experimental cell research*, 389(2). Available at: https://doi.org/10.1016/J.YEXCR.2020.111924.

Takahashi, K. *et al.* (2007) 'Induction of Pluripotent Stem Cells from Adult Human Fibroblasts by Defined Factors', *Cell*, 131(5), pp. 861–872. Available at: https://doi.org/10.1016/j.cell.2007.11.019.

Takahashi, K. and Yamanaka, S. (2006) 'Induction of Pluripotent Stem Cells from Mouse Embryonic and Adult Fibroblast Cultures by Defined Factors', *Cell*, 126(4), pp. 663–676. Available at: https://doi.org/10.1016/j.cell.2006.07.024.

Taverna, M. *et al.* (2013) 'Specific antioxidant properties of human serum albumin', *Annals of Intensive Care*, pp. 1–7. Available at: https://doi.org/10.1186/2110-5820-3-4.

Tchurikov, N.A. *et al.* (2015) 'Genome-wide mapping of hot spots of DNA double-strand breaks in human cells as a tool for epigenetic studies and cancer genomics', *Genomics data*, 5, pp. 89–93. Available at: https://doi.org/10.1016/J.GDATA.2015.05.018.

Tena, A. *et al.* (2020) 'Induction of recurrent break cluster genes in neural progenitor cells differentiated from embryonic stem cells in culture', *Proceedings of the National Academy of Sciences of the United States of America*, 117(19), pp. 10541–10546. Available at: https://doi.org/10.1073/PNAS.1922299117.

Teng, Y.H.F. *et al.* (2011) 'Mutations in the epidermal growth factor receptor (EGFR) gene in triple negative breast cancer: Possible implications for targeted therapy', *Breast Cancer Research*, 13(2), pp. 1–9. Available at: https://doi.org/10.1186/BCR2857/TABLES/5.

Thomas, R. *et al.* (2017) 'Features that define the best ChIP-seq peak calling algorithms', *Briefings in bioinformatics*, 18(3), pp. 441–450. Available at: https://doi.org/10.1093/BIB/BBW035.

Thompson, O. *et al.* (2020) 'Low rates of mutation in clinical grade human pluripotent stem cells under different culture conditions', *Nature Communications*, 11(1), pp. 1–14. Available at: https://doi.org/10.1038/s41467-020-15271-3.

Thomson, J.A. *et al.* (1998) 'Embryonic stem cell lines derived from human blastocysts', *Science*, 282(5391), pp. 1145–1147. Available at: https://doi.org/10.1126/science.282.5391.1145.

Tian, M. and Alt, F.W. (2000) 'Transcription-induced Cleavage of Immunoglobulin Switch Regions by Nucleotide Excision Repair Nucleases in Vitro', *Journal of Biological Chemistry*, 275(31), pp. 24163–24172. Available at: https://doi.org/10.1074/JBC.M003343200.

Toledo, L.I. *et al.* (2013) 'XATR prohibits replication catastrophe by preventing global exhaustion of RPA', *Cell*, 155(5), p. 1088. Available at: https://doi.org/10.1016/j.cell.2013.10.043.

Trowsdale, J. and Knight, J.C. (2013) 'Major histocompatibility complex genomics and human disease', *Annual Review of Genomics and Human Genetics*. Annu Rev Genomics Hum Genet, pp. 301–323. Available at: https://doi.org/10.1146/annurev-genom-091212-153455.

Tubbs, A. *et al.* (2018) 'Dual Roles of Poly(dA:dT) Tracts in Replication Initiation and Fork Collapse', *Cell*, 174(5), pp. 1127-1142.e19. Available at: https://doi.org/10.1016/j.cell.2018.07.011.

Tuduri, S. *et al.* (2009) 'Topoisomerase I suppresses genomic instability by preventing interference between replication and transcription', *Nature cell biology*, 11(11), pp. 1315–1324. Available at: https://doi.org/10.1038/NCB1984.

Turocy, J. *et al.* (2022) 'DNA Double Strand Breaks cause chromosome loss through sister chromatid tethering in human embryos', *bioRxiv*, p. 2022.03.10.483502. Available at: https://doi.org/10.1101/2022.03.10.483502.

Uusküla-Reimand, L. *et al.* (2016) 'Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders', *Genome Biology 2016 17:1*, 17(1), pp. 1–22. Available at: https://doi.org/10.1186/S13059-016-1043-8.

Vallabhaneni, H. *et al.* (2018) 'High Basal Levels of γH2AX in Human Induced Pluripotent Stem Cells Are Linked to Replication-Associated DNA Damage and

Repair', *Stem Cells*, 36(10), pp. 1501–1513. Available at: https://doi.org/10.1002/stem.2861.

Vallier, L., Alexander, M. and Pedersen, R.A. (2005) 'Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells', *Journal of cell science*, 118(Pt 19), pp. 4495–4509. Available at: https://doi.org/10.1242/JCS.02553.

Vanoosthuyse, V. (2018) 'Strengths and weaknesses of the current strategies to map and characterize R-loops', *Non-coding RNA*. MDPI AG. Available at: https://doi.org/10.3390/ncrna4020009.

Varum, S. *et al.* (2011) 'Energy Metabolism in Human Pluripotent Stem Cells and Their Differentiated Counterparts', *PLoS ONE*. Edited by M. Ludgate, 6(6), p. e20914. Available at: https://doi.org/10.1371/journal.pone.0020914.

Vidricaire, G., Jardine, K. and McBurney, M.W. (1994) 'Expression of the Brachyury gene during mesoderm development in differentiating embryonal carcinoma cell cultures', *Development (Cambridge, England)*, 120(1), pp. 115–122. Available at: https://doi.org/10.1242/DEV.120.1.115.

Vitillo, L. *et al.* (2023) 'The isochromosome 20q abnormality of pluripotent cells interrupts germ layer differentiation', *Stem cell reports*, 18(3), pp. 782–797. Available at: https://doi.org/10.1016/J.STEMCR.2023.01.007.

Vrtis, K.B. *et al.* (2021) 'Single-strand DNA breaks cause replisome disassembly', *Molecular Cell*, 81(6), pp. 1309-1318.e6. Available at: https://doi.org/10.1016/j.molcel.2020.12.039.

Wang, D. *et al.* (2022) 'Active DNA demethylation promotes cell fate specification and the DNA damage response', *Science*, 378(6623), pp. 983–989. Available at: https://doi.org/10.1126/SCIENCE.ADD9838/SUPPL_FILE/SCIENCE.ADD9838_MOVIE_S1.ZIP.

Wang, K. *et al.* (2019) 'Ultra-High-Frequency Reprogramming of Individual Long-Term Hematopoietic Stem Cells Yields Low Somatic Variant Induced Pluripotent Stem Cells', *Cell reports*, 26(10), pp. 2580-2592.e7. Available at: https://doi.org/10.1016/J.CELREP.2019.02.021.

Wang, M. *et al.* (2020) 'Increased Neural Progenitor Proliferation in a hiPSC Model of Autism Induces Replication Stress-Associated Genome Instability', *Cell Stem Cell*,

26(2), pp. 221-233.e6. Available at: https://doi.org/10.1016/j.stem.2019.12.013.

Wang, X. and Wu, Q. (2022) 'The Divergent Pluripotent States in Mouse and Human Cells', *Genes*. MDPI, p. 1459. Available at: https://doi.org/10.3390/genes13081459.

Wang, Y. *et al.* (2001) *Antitumor Drug Adozelesin Differentially Affects Active and Silent Origins of DNA Replication in Yeast Checkpoint Kinase Mutants 1*, *CANCER RESEARCH*. Available at: http://aacrjournals.org/cancerres/article-pdf/61/9/3787/3253875/ch090103787p.pdf (Accessed: 15 August 2023).

Wang, Y. *et al.* (2022) 'Digital Counting of Breaks Labeling In Situ: A Fast and Absolute Quantification Method for Measurement of DNA Double-Strand Breaks Based on Digital Polymerase Chain Reaction', *Analytical Chemistry* [Preprint]. Available at: https://doi.org/10.1021/ACS.ANALCHEM.2C03985/ASSET/IMAGES/LARGE/AC2C03985_0007.JPEG.

Warmflash, A. *et al.* (2014) 'A method to recapitulate early embryonic spatial patterning in human embryonic stem cells', *Nature Methods*, 11(8), pp. 847–854. Available at: https://doi.org/10.1038/NMETH.3016.

Warnock, M. (1985) 'Moral thinking and government policy: the Warnock Committee on Human Embryology.', *The Milbank Memorial Fund quarterly. Health and society*, 63(3), pp. 504–522. Available at: https://doi.org/10.2307/3349845.

Warren, L. *et al.* (2010) 'Highly efficient reprogramming to pluripotency and directed differentiation of human cells with synthetic modified mRNA', *Cell stem cell*, 7(5), pp. 618–630. Available at: https://doi.org/10.1016/J.STEM.2010.08.012.

Watanabe, K. *et al.* (2007) 'A ROCK inhibitor permits survival of dissociated human embryonic stem cells', *Nature Biotechnology*, 25(6), pp. 681–686. Available at: https://doi.org/10.1038/nbt1310.

Wei, P.C. *et al.* (2016) 'Long Neural Genes Harbor Recurrent DNA Break Clusters in Neural Stem/Progenitor Cells', *Cell*, 164(4), pp. 644–655. Available at: https://doi.org/10.1016/j.cell.2015.12.039.

Wei, P.C. *et al.* (2018) 'Three classes of recurrent DNA break clusters in brain progenitors identified by 3D proximity-based break joining assay', *Proceedings of the National Academy of Sciences of the United States of America*, 115(8), pp. 1919–1924. Available at: https://doi.org/10.1073/pnas.1719907115.

Weissbein, U. *et al.* (2016) ‘Analysis of chromosomal aberrations and recombination by allelic bias in RNA-Seq’, *Nature Communications*, 7. Available at: https://doi.org/10.1038/ncomms12144.

Werbowetski-Ogilvie, T.E. *et al.* (2009) ‘Characterization of human embryonic stem cells with features of neoplastic progression’, *Nature Biotechnology*, 27(1), pp. 91–97. Available at: https://doi.org/10.1038/nbt.1516.

Wilson, H.K. *et al.* (2015) ‘Exploring the effects of cell seeding density on the differentiation of human pluripotent stem cells to brain microvascular endothelial cells’, *Fluids and Barriers of the CNS*, 12(1), pp. 1–12. Available at: https://doi.org/10.1186/S12987-015-0007-9/FIGURES/4.

Woodward, A.M. *et al.* (2006) ‘Excess Mcm2–7 license dormant origins of replication that can be used under conditions of replicative stress’, *Journal of Cell Biology*, 173(5), pp. 673–683. Available at: https://doi.org/10.1083/JCB.200602108.

Wu, J. *et al.* (2023) ‘Cohesin maintains replication timing to suppress DNA damage on cancer genes’, *Nature genetics* [Preprint]. Available at: https://doi.org/10.1038/S41588-023-01458-Z.

Wu, J. and Izpisua Belmonte, J.C. (2015) ‘Dynamic Pluripotent Stem Cell States and Their Applications’, *Cell Stem Cell*, 17(5), pp. 509–525. Available at: https://doi.org/10.1016/J.STEM.2015.10.009.

Wu, T. *et al.* (2021) ‘clusterProfiler 4.0: A universal enrichment tool for interpreting omics data’, *Innovation*, 2(3). Available at: https://doi.org/10.1016/j.xinn.2021.100141.

Wu, W. *et al.* (2021) ‘Neuronal enhancers are hotspots for DNA single-strand break repair’, *Nature*, 593(7859), pp. 440–444. Available at: https://doi.org/10.1038/s41586-021-03468-5.

Wu, Y. *et al.* (2015) ‘EEPD1 Rescues Stressed Replication Forks and Maintains Genome Stability by Promoting End Resection and Homologous Recombination Repair’, *PLOS Genetics*, 11(12), p. e1005675. Available at: https://doi.org/10.1371/JOURNAL.PGEN.1005675.

Xiao, H. *et al.* (2003) ‘Acidic pH induces topoisomerase II-mediated DNA damage’, *Proceedings of the National Academy of Sciences of the United States of America*, 100(9), pp. 5205–5210. Available at: https://doi.org/10.1073/pnas.0935978100.

Xu, F. *et al.* (2021) 'Lysophosphatidic acid shifts metabolic and transcriptional landscapes to induce a distinct cellular state in human pluripotent stem cells', *Cell Reports*, 37(9), p. 110063. Available at: https://doi.org/10.1016/J.CELREP.2021.110063.

Xu, H. *et al.* (2017) 'Epidermal growth factor receptor in glioblastoma (Review)', *Oncology Letters*, 14(1), pp. 512–516. Available at: https://doi.org/10.3892/ol.2017.6221.

Yan, P. *et al.* (2020) 'Genome-wide R-loop Landscapes during Cell Differentiation and Reprogramming', *Cell Reports*, 32(1), p. 107870. Available at: https://doi.org/10.1016/j.celrep.2020.107870.

Yan, W.X. *et al.* (2017) 'BLISS is a versatile and quantitative method for genome-wide profiling of DNA double-strand breaks', *Nature Communications*, 8. Available at: https://doi.org/10.1038/ncomms15058.

Yang, F., Kemp, C.J. and Henikoff, S. (2015) 'Anthracyclines induce double-strand DNA breaks at active gene promoters', *Mutation research*, 773, pp. 9–15. Available at: https://doi.org/10.1016/J.MRFMMM.2015.01.007.

Yilmaz, A. and Benvenisty, N. (2019) 'Defining Human Pluripotency', *Cell Stem Cell*. Cell Press, pp. 9–22. Available at: https://doi.org/10.1016/j.stem.2019.06.010.

Yin, J. *et al.* (2019) 'Optimizing genome editing strategy by primer-extension-mediated sequencing', *Cell Discovery*, 5(1). Available at: https://doi.org/10.1038/s41421-019-0088-8.

Yin, Z. *et al.* (2001) 'Limiting the location of putative human prostate cancer tumor suppressor genes on chromosome 18q', *Oncogene*, 20(18), pp. 2273–2280. Available at: https://doi.org/10.1038/sj.onc.1204310.

Ying, Q.L. *et al.* (2008) 'The ground state of embryonic stem cell self-renewal', *Nature 2008 453:7194*, 453(7194), pp. 519–523. Available at: https://doi.org/10.1038/nature06968.

Ying, S. *et al.* (2013) 'MUS81 promotes common fragile site expression', *Nature Cell Biology*, 15(8), pp. 1001–1007. Available at: https://doi.org/10.1038/ncb2773.

Yuan, J. *et al.* (2000) *Diminished DNA Repair and Elevated Mutagenesis in Mammalian Cells Exposed to Hypoxia and Low pH 1*, *CANCER RESEARCH*.

Available at: http://aacrjournals.org/cancerres/article-pdf/60/16/4372/2479797/ch160004372.pdf (Accessed: 9 August 2023).

Yuan, Z. and Li, H. (2020) 'Molecular mechanisms of eukaryotic origin initiation, replication fork progression, and chromatin maintenance', *Biochemical Journal*. Portland Press Ltd, pp. 3499–3525. Available at: https://doi.org/10.1042/BCJ20200065.

Zampetidis, C.P. *et al.* (2021) 'A recurrent chromosomal inversion suffices for driving escape from oncogene-induced senescence via subTAD reorganization', *Molecular Cell*, 81(23), pp. 4907-4923.e8. Available at: https://doi.org/10.1016/J.MOLCEL.2021.10.017.

Zang, C. *et al.* (2009) 'A clustering approach for identification of enriched domains from histone modification ChIP-Seq data', *Bioinformatics*, 25(15), pp. 1952–1958. Available at: https://doi.org/10.1093/BIOINFORMATICS/BTP340.

Zeman, M.K. and Cimprich, K.A. (2014) 'Causes and Consequences of Replication Stress', *Nature cell biology*, 16(1), p. 2. Available at: https://doi.org/10.1038/NCB2897.

Zhang, H.Y. *et al.* (2009) 'In benign Barrett's epithelial cells, acid exposure generates reactive oxygen species that cause DNA double-strand breaks', *Cancer Research*, 69(23), pp. 9083–9089. Available at: https://doi.org/10.1158/0008-5472.CAN-09-2518/655608/P/IN-BENIGN-BARRETT-S-EPITHELIAL-CELLS-ACID-EXPOSURE.

Zhang, T. *et al.* (2022) 'Replication collisions induced by de-repressed S-phase transcription are connected with malignant transformation of adult stem cells', *Nature Communications 2022 13:1*, 13(1), pp. 1–17. Available at: https://doi.org/10.1038/s41467-022-34577-y.

Zhang, Y. *et al.* (2008) 'Model-based analysis of ChIP-Seq (MACS)', *Genome Biology*, 9(9), pp. 1–9. Available at: https://doi.org/10.1186/GB-2008-9-9-R137/FIGURES/3.

Zhao, Y. *et al.* (2021) 'Resolvin D1 attenuates acid-induced DNA damage in esophageal epithelial cells and rat models of acid reflux', *European Journal of Pharmacology*, p. 174571. Available at: https://doi.org/10.1016/J.EJPHAR.2021.174571.

Zhou, J. *et al.* (2023) 'Induction and application of human naive pluripotency', *Cell Reports*. Elsevier B.V., p. 112379. Available at:

https://doi.org/10.1016/j.celrep.2023.112379.

Zhu, Y. *et al.* (2019) ‘qDSB-Seq is a general method for genome-wide quantification of DNA double-strand breaks using sequencing’, *Nature Communications 2019 10:1*, 10(1), pp. 1–11. Available at: https://doi.org/10.1038/s41467-019-10332-8.