The 3D Genome: Investigating the role of Pcf11 and Pol II gene loops in mammalian cell lines

by

Maria Perdiou

A thesis submitted to The University of Warwick
in partial fulfilment of the requirements for the degree of
Doctor of Philosophy in Life Sciences

The University of Warwick, School of Life Sciences

March 2022

Table of Contents

List of Figures

List of Tables

Acknowledgements

I would like to thank and acknowledge the following people:

My supervisor, Dr Daniel Hebenstreit, for his guidance and support.

Past and present members of the Hebenstreit Lab, in particular; David Edwards for his help and contribution to the computational aspect of this work in Chapter 5, Matt Jones for useful discussions and bioinformatics troubleshooting, Dr Steven Servín-González for his help with routine wet-lab techniques when I first joined the lab, and Dr Mark Walsh for advice on DNA library preparation.

Dr Chrystala Constantinidou for setting up the bioinformatics pipeline for library processing, and the rest of the Bioinformatics RTP team at SLS for helping me set up all the tools I needed.

Marieke Otten for her contribution to qPCR experiments during her time as an undergraduate student in our lab, and Poonam Shah for her help with optimising transfections.

My Advisory Panel, Professor Keith Leppard and Dr Andrew Nelson, for advice on the structure of the write-up.

Close friends and family for their continued support throughout my PhD journey.

My Doctoral program, The Midlands Integrative Biosciences Training Partnership (MIBTP), along with the BBSRC and The University of Warwick for my PhD studentship.

Declarations

I hereby declare that the present thesis titled ‘The 3D Genome: Investigating the role of Pcf11 and Pol II gene loops in mammalian cell lines’ is my own and has not been previously submitted for a degree or qualification at any university. All experiments and analyses were performed by the author, except for the following:

- Some of the qPCR experiments presented in Chapter 3 were done together with undergraduate student Marieke Otten.
- The custom scripts used to identify gene looping in Chapter 5 were written by David Edwards. David also plotted **Fig. 14** in Chapter 5 and performed the corresponding statistical test.

Abstract

Eukaryotic transcription, from initiation to termination, is a highly complex and regulated process involving the interplay of different factors. Termination is a key step in gene expression, as it does not only delimit transcription, but also affects the localisation and stability of transcripts. One of the factors involved in the termination process is the cleavage and polyadenylation factor subunit Pcf11. Human Pcf11 is recruited by RNA Polymerase II to enhance termination of transcription and 3’ end processing. Evidence shows that the formation of chromatin loops impacts transcriptional dynamics and suggests that Pcf11 is implicated in 3’-5’ end crosstalk due to its recently discovered localisation also at the TSS of genes. Simultaneously, depletion of Pcf11 has been found to reduce transcription initiation rates. Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) is a valuable tool in chromatin loop research as it allows for the identification of chromatin interactions bound by a specific protein of interest. The present work focuses on establishing quality of ChIA-PET libraries generated and analysed using the latest version of both the experimental protocol (*in-situ* ChIA-PET) and data processing pipeline. We identify and characterise gene loops bound by Pol II and Pcf11 and show that looped genes are involved in key cellular processes and are significantly more expressed than non-looped genes. Analysis of 3’ mRNA-seq data reveals that Pcf11 depletion causes significant downregulation of looped genes, providing further evidence that Pcf11 regulates aspects of transcription initiation and that its activity is crucial for the expression of looped genes.

Abbreviations

3C	Chromatin Conformation Capture
2D	2-Dimensional
3D	3-Dimensional
5C	Carbon Copy Chromosome Conformation Capture
BAM	Binary Alignment Map
BED	Browser Extensible Data
BSA	Bovine Serum Albumin
CCD	Chromatin Contact Domains
cDNA	Complementary DNA
CFIIm	Cleavage Factor II
CFIm	Cleavage Factor I
ChIA-PET	Chromatin Interaction Analysis w/ Paired-End Tag
ChIP	Chromatin Immunoprecipitation
CID	C-terminal Interacting Domain
COLA	Concatemer Ligation Assay
CPA	Cleavage and Polyadenylation
CPSF	Cleavage and Polyadenylation Specificity Factor
CRISPRi	CRISPR Interference
CstF	Cleavage Stimulation Factor
CT	Chromosomal Territory
Ct	C-terminus
CTD	C-terminal Domain
dATP	Deoxyadenosine Triphosphate

DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl Sulfoxide
DNA	Deoxyribonucleic acid
dNTPs	Deoxynucleotide Triphosphates
dsDNA	Double-stranded DNA
DSIF	DRB Sensitivity Inducing Factor
DTT	Dithiothreitol
EB	Elution Buffer
EDTA	Ethylenediaminetetraacetic Acid
EGS	Ethylene Glycol bis (succinimidyl succinate)
ENA	European Nucleotide Archive
ER-α	Oestrogen Receptor α
FBS	Foetal Bovine Serum
FC	Fold-Change
FDR	False Discovery Rate
FISH	Fluorescent in-situ Hybridization
FP	Flavopiridol
GEO	Gene Expression Omnibus
GO	Gene Ontology
HEK-293	Human Embryonic Kidney
HeLa	Henrietta Lacks
hg38	Human Genome 38 (annotation)
IGV	Integrative Genomics Viewer
Int	Internal
IP	Immunoprecipitation

KD	Knock-down
LCR	Locus Control Region
LLPS	Liquid-Liquid Phase Separation
M-MLV	Moloney-Murine Leukaemia Virus
mRNA	Messenger RNA
MT	Mitochondrial
NELF	Negative Elongation Factor
nTPM	Normalised Transcripts per Million
p-TEFb	Positive Transcription Elongation Factor B
PAS	Polyadenylation Site
PBS	Phosphate-Buffered Saline
PBST	Phosphate-Buffered Saline Tween
PCR	Polymerase Chain Reaction
PE	Paired-End
PET	Paired-End Tag
PMO	Phosphorodiamidate Morpholino Oligos
Pol II	RNA Polymerase II
PTT	Premature Transcription Termination
QA	Quality Assessment
QC	Quality Control
RISC	RNA-Induced Silencing Complex
RNA	Ribonucleic Acid
RNAi	RNA Interference
RNAPII	RNA Polymerase II
rRNA	Ribosomal RNA

RT	Room Temperature
RT-qPCR	Real Time Quantitative PCR
SD	Standard Deviation
SDS	Sodium Dodecyl Sulphate
Seq	Sequencing
Ser	Serine
SerP	Serine Phosphorylation
shRNA	Small-hairpin RNA
siRNA	Small Interfering RNA
snRNA	Small Nuclear RNA
SSC	Saline-Sodium Citrate
TAD	Topologically Associated Domains
TBST	Tris-buffered Saline Tween
TC	Tissue Culture
TES	Transcription Termination Site
TF	Transcription Factor
TPM	Transcripts per Million
Tris	Trisine
tRNA	Transfer RNA
Trp	Triptolide
TSS	Transcription Start Site

Chapter 1: Introduction

1.1 Background

1.1.2 Motivation for Research

The entire process of eukaryotic transcription, from initiation to termination, involves several events. Chromatin interactions play an essential role in the assembly of RNA polymerase and the steps that follow to generate transcripts. Transcription termination is a key step in gene expression, as this is how nascent transcripts are produced (Porrua and Libri, 2015). It is a highly complex and regulated process, involving the interplay of different factors. One of these factors is the cleavage and polyadenylation factor Pcf11. Human Pcf11 is a transcription termination factor recruited by RNA Polymerase II (Pol II) and it has been found to enhance termination of transcription and 3’ end processing in a genome-wide fashion (Kamieniarz-Gdula et al., 2019). Most research studying Pcf11 has been carried out in yeast and flies, therefore, there is still a lot to uncover about its function in human cells.

Evidence shows that the formation of chromatin loops impacts transcriptional dynamics (Ansari and Hampsey, 2005; Sing and Hampsey, 2007) and that the Pcf11 protein is implicated in 3’ - 5’ end crosstalk (Mapendano et al., 2010). With the rigorous developments in chromosome conformation capture technology, Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET) has become a valuable tool in chromatin loop research, as it allows for the identification of chromatin interactions mediated by a specific protein of interest (Lee et al., 2020 and Wang et al., 2021). My research focuses on studying Pol II and Pcf11 gene loops as well as their impact on transcription in human cells.

1.1.2 DNA, Transcription, and Gene Expression

DNA is the polymeric molecule containing the hereditary, genetic material of humans and most other organisms. It is made up of nucleotides, each of which contains one nitrogenous base. These bases are Adenine (A), Thymine (T), Guanine (G), and Cytosine (C). Almost all cells in the human body contain the same DNA, which is organised in the form of a double helix. This helical structure is made up of two complementary DNA strands interlaced with each other. In order for cells to pass on their genetic information to new cells, DNA undergoes replication.

During transcription, genes are transcribed into messenger RNA (mRNA) and are eventually translated into protein. This process is known as the “central dogma” of molecular biology (**Fig. 1**). Transcription is initiated by the activity of Pol II which binds on the promoter region of DNA, upstream of the transcription start site. In eukaryotic organisms, transcription factors (TFs) form pre-initiation complexes by binding to promoters. These complexes are recognised by Pol II and the synthesis of RNAs is initiated. Pol II moves along the DNA template and generates an RNA copy. The production of mRNA undergoes splicing. The spliceosome, made up from small nuclear RNAs (snRNA) and proteins, removes the introns and keeps the exons of the transcript. The remaining mRNA gets translated into protein in the ribosomes.

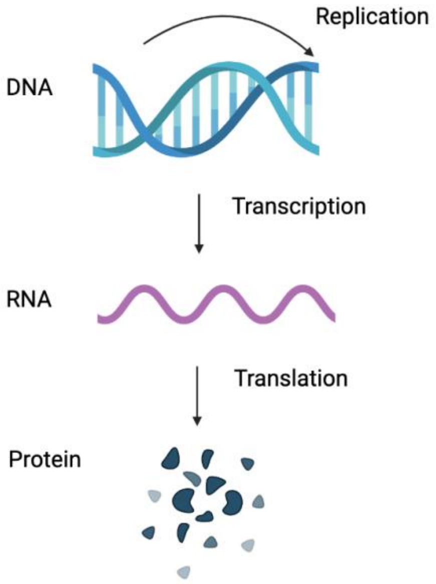


Fig. 1 The central dogma of molecular biology. DNA undergoes replication to create copies of itself. DNA is transcribed by RNA Polymerase II and mRNA is produced. mRNA is then translated into protein. (Image created on BioRender).

1.1.3 Transcription by RNA Polymerase II

In eukaryotes, transcription of genes is a process that is shared between three distinct RNA polymerase enzymes that are present in the nucleus. Each of these polymerases fulfils a different function; RNA polymerase I (PolI) is involved in the synthesis of the large ribosomal RNA (rRNA) and 5.8S rRNA, RNA polymerase II (Pol II) is involved in transcription and mRNA synthesis, and RNA polymerase III (Pol III) is used to make transfer RNA (tRNA) and 5S rRNA. Due to its important role in transcription and gene expression of protein-coding genes, Pol II has received the most attention out of the three polymerases.

Pol II is a complex of 12 subunits (Rpb1-12) that transcribes DNA into mRNA and small nuclear RNA (snRNA). Rpb 1 is the largest subunit of human Pol II and contains a unique C-terminal domain (CTD) made up of 52 heptad repeats (consensus: $Y_1S_2P_3T_4S_5P_6S_7$) that plays an essential role in the activity of Pol II. Since Pol II is responsible for mRNA synthesis in genes that are protein coding, it has been widely studied the past two decades. As opposed to prokaryotic RNA Polymerase, eukaryotic Pol II cannot initiate transcription and is not able to transcribe consistently along the DNA templates. Instead, to accomplish these functions, a number of protein complexes need to interact with it to modulate its activity. These factors control Pol II initiation, pausing, and elongation (Schier and Taatjes, 2020). The phosphorylation state of Pol II's CTD regulates its activity. A generalised model of the phosphorylation of the CTD during the process of transcription indicates that during initiation, Ser5 is phosphorylated and as Pol II progresses to elongation, Ser5 phosphorylation is gradually removed and instead Ser2 becomes increasingly phosphorylated. Although the situation is more complex, the emerging pattern involves Ser5P peaks during TSS, and Ser2P accumulation towards the TES (Hsin, 2012). Ser7 appears to be consistently phosphorylated throughout the transcriptional process, although this residue is much less studied in mammalian systems (Egloff, 2012; Mayfield et al., 2016). **Fig. 2** shows a schematic of Pol II's phosphorylation states in *S. cerevisiae*.

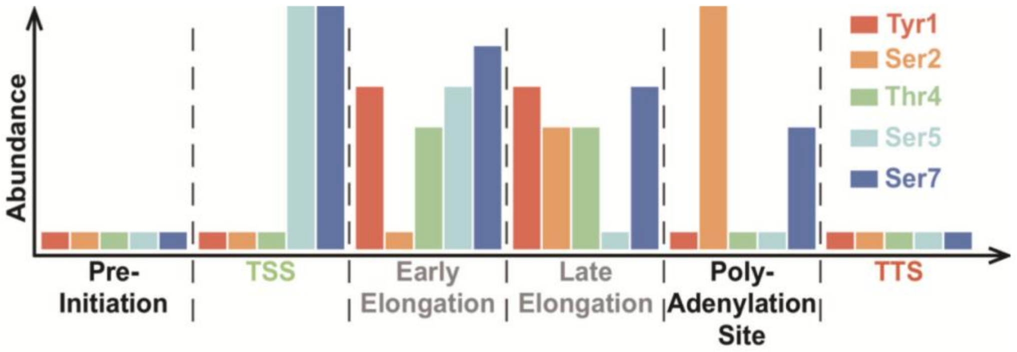


Fig. 2 CTD residue phosphorylation states through different transcriptional stages. Abundance is based on approximate enrichment from ChIP-seq data derived from *S. cerevisiae*. The x-axis represents the transcription stage from pre-initiation, transcription start site (TSS), early and late elongation, polyadenylation site, and transcription termination site (TTS) (Schematic taken from Mayfield et al., 2016).

The main steps of eukaryotic transcription involve initiation, elongation, termination, and processing. In order for Pol II to synthesise mRNA, a transcription initiation complex needs to be formed, followed by a transition to an elongation complex (Adelman et al., 2012, Sainsbury et al., 2015, and Harlen et al., 2017). Transcription initiation requires the involvement and cooperation of several other polypeptides. This is so that the promoter DNA can be effectively recognised and to also locate the transcription start site (TSS) so that initiation of synthesis of pre-mRNA can begin (Sainsbury et al., 2015). During initiation, the DNA unwinds to allow for the formation of a small open complex where Pol II can bind to the promoter. During the transition from initiation to elongation, the disordered CTD of Pol II's large subunit becomes phosphorylated by kinases (Eick et al., 2013). Elongation involves the synthesis of mRNA, achieved through the movement of Pol II along the template strand. Following the initiation of transcription, the majority of metazoan cells go through a regulatory step known as promoter proximal pausing (Muse et al., 2007 and Core et al., 2008), where Pol II pauses after it has transcribed 20-120 nucleotides downstream of the TSS.

For Pol II to enter productive elongation and generate full-length mRNAs, the activity of other factors is required; particularly, the activity of the positive transcription elongation factor b (p-TEFb) (Ni et al., 2008; Jonkers et al., 2014). After the production of full-length mRNAs, follows transcription termination. This process is coupled with mRNA processing and occurs co-transcriptionally. When the gene end is reached by Pol II, Pol II starts slowing down over the region of termination (Proudfoot, 2016). Part of the reason for this is because as soon as the nascent transcript signals polyadenylation, the 3' end cleavage and polyadenylation (CPA) complex gets recruited to Pol II. The CPA is then able to release the transcript from the template strand and allow it to eventually get translated in the cytoplasm. After the mRNA transcript is released, Pol II continues transcribing along the template, although this is short-lived due to the transcript being degraded from its 5' end by exonucleases (Proudfoot, 2016). Pol II, and particularly its CTD, is also involved in the final step of transcription which involves the processing of nascent transcripts. The CTD of Pol II extends from the core of the protein and forms a binding-site rich, tail-like structure that allows different RNA processing factors to bind to (Hsin and Manley, 2012). **Fig. 3** shows a schematic of Pol II transcription, from initiation to termination.

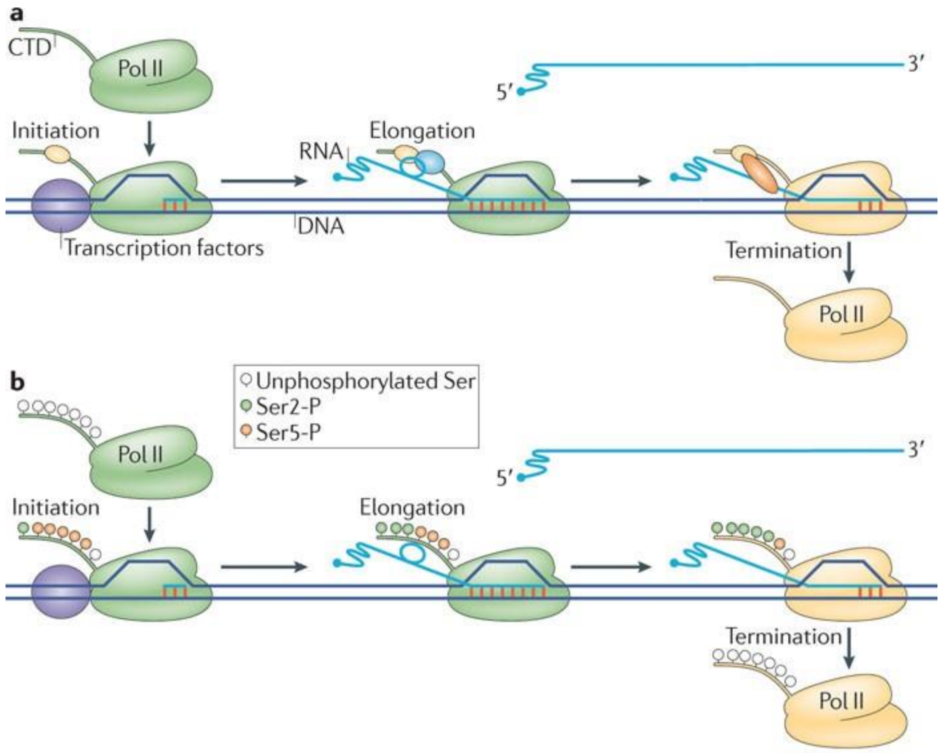


Fig. 3 Pol II transcription and its coordination through distinct patterns of CTD phosphorylation. A) Initiation: Transcription factors help Pol II to get recruited to gene promoters, the DNA is melted to expose the template strand, and RNA synthesis begins. **Elongation:** A ~ 8-9 bp hybrid of RNA and DNA is formed, and Pol II extends the transcript. **Termination:** Pol II stops the synthesis of RNA and becomes prone to termination (yellow colour). Both Pol II and nascent RNA are released from the template. Protein factors that are involved in elongation (yellow ovals), RNA processing (blue ovals), and termination (orange ovals) associate with the CTD of Pol II co-transcriptionally. **B)** The phosphorylation status of the CTD tail (heptad repeat) changes as Pol II moves along a gene. Hypophosphorylated Pol II gets recruited onto the pre-initiation complex and becomes phosphorylated on Ser5 by TFIIH CDK7 (in mammals) during initiations, and on Ser2 by CDK9 (in mammals) during elongation (schematic taken from Kuehner et al., 2011).

1.1.4 Pcf11 Overview

The human Pcf11 protein is encoded by the Pcf11 gene, also known as PCF11 Cleavage and Polyadenylation Factor Subunit, or Pre-mRNA Cleavage Complex II Protein Pcf11. The gene is located on chromosome 11 and codes for a 1,555 amino acid protein (**Fig. 4**). Pcf11 is essential for effective transcription termination by Pol II and plays a role in polyadenylation, although the mechanism is not entirely understood yet.

Although the role and function of Pcf11 in transcription and termination have mainly been studied in yeast and flies and are yet to be completely understood, it appears that Pcf11 contains an N-terminal CTD-interaction domain (CID) that, depending on phosphorylation levels, binds to the heptad repeats of Pol II's CTD and pulls apart elongating Pol II from the DNA sequence (Barilla et al., 2001, Sadowski et al., 2003; Noblet et al., 2005). Particularly, Pcf11 preferentially binds to Pol II heptad repeats that are phosphorylated on Ser2 (Licatalosi et al., 2002). Pcf11 has been found to be required for this complex to effectively terminate transcription through its CID (Grzechnik et al., 2015). Other studies have also shown that yeast Pcf11 is involved in the recruitment of mRNA export factors (Johnson et al., 2009). The CID of Pcf11 has also been shown to have RNA binding activity. It has been proposed that RNA and the process of the CTD binding to the CID compete with each other to mediate disengagement of Pol II (Zhang et al., 2005 and Hollingworth et al., 2006).

In a *drosophila* study where the Pcf11 protein had been depleted, it was revealed that Pcf11 is directly involved in termination since Pol II continued transcribing beyond the typical region of termination (Zhang and Gilmour 2006). The same experimental approach was used in a much more recent study where it was discovered that the same findings held true for vertebrates, specifically zebrafish and human cells. Again, Pcf11 was depleted using RNAi and transcription beyond usual readthrough regions was observed. It was also observed that Pcf11 selectively reduces expression of other elements involved in transcription regulation through premature termination (i.e., termination signal before the full-length transcript has been produced) along with cleavage and polyadenylation (Kamieniarz-Gdula et al., 2019).

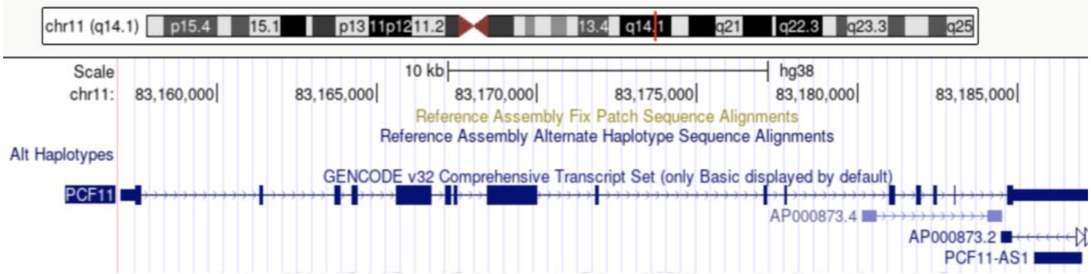


Fig. 4 Location of the Pcf11 gene on GRCh38/hg38. Pcf11 is located on the plus strand of chromosome 11 at position 83,156,988 - 83,187,451. Pcf11 is 3,464 bp long and encodes for a 1555 aa protein. (Screenshot taken from UCSC Genome Browser).

1.1.4.1 Role of Pcf11 in transcription termination

As it has already been mentioned in the previous section, termination is the last step of the transcriptional cycle, and it is crucial for correct gene expression. During termination, Pol II and RNA are being released from the DNA template and synthesis of mRNA is halted. The termination process not only ensures that RNA is released so it can be translated into protein, but that there is Pol II available for following rounds of mRNA synthesis. It also limits the phenomenon of non-coding transcription (Porrua et al., 2016 and Proudfoot, 2016).

Termination is also mechanistically involved in the 3’ end processing of pre-mRNAs. Pol II plays its part in termination after it has transcribed a poly(A) signal in the mRNA, which is then recognised by the 3’ end processing machinery of the RNA. As a result, polyadenylation and RNA cleavage occur a few nucleotides downstream of the poly-A site (PAS), leading to transcript termination (Porrua et al., 2016; Proudfoot, 2016). However, in mammals, even though the steps of cleavage and polyadenylation take place at specific genomic locations, Pol II still continues transcribing sequences that are several nucleotides downstream of the PAS (Schwalb et al., 2016).

A big CPA complex containing factors relevant to cleavage and polyadenylation processes the 3’ ends of the mRNA of higher eukaryotes, such as mammals (**Fig. 5**). These factors include the cleavage stimulation factor (CstF), the cleavage and polyadenylation specificity factor (CPSF), and cleavage factors I and II (CFIm and CFIIm), all of which are made up of several smaller subunits (Shi and Manley, 2015). CFIIm is made up of the CLP1 and Pcf11 proteins and interacts with the CPA complex in a more transient and/or weak way, as opposed to other CPA factors (Shi et al., 2009). The majority of CPA factors play their role in defined steps, however, Pcf11 appears to be important not only in 3’ end processing (Amrani et al., 1997, and Gross and Moore, 2001), but also in the termination of transcription (Zhang et al., 2005 and West and Proudfoot, 2008) where it is also involved in linking transcription with the export of mRNA (Johnson et al., 2009 and Volanakis et al., 2017). In yeast, specific domains of Pcf11, that can also be functionally separated, are responsible for Pcf11’s activities in termination and processing of the 3’ end (Sadowski et al., 2003). The interaction between Pcf11’s CID with Pol II’s CTD can pull apart elongating complexes in vitro (Zhang and Gilmour, 2006) and is necessary for normal levels of phosphorylation of Ser2 on the CTD of Pol II in yeast (Grzechnik et al., 2015).

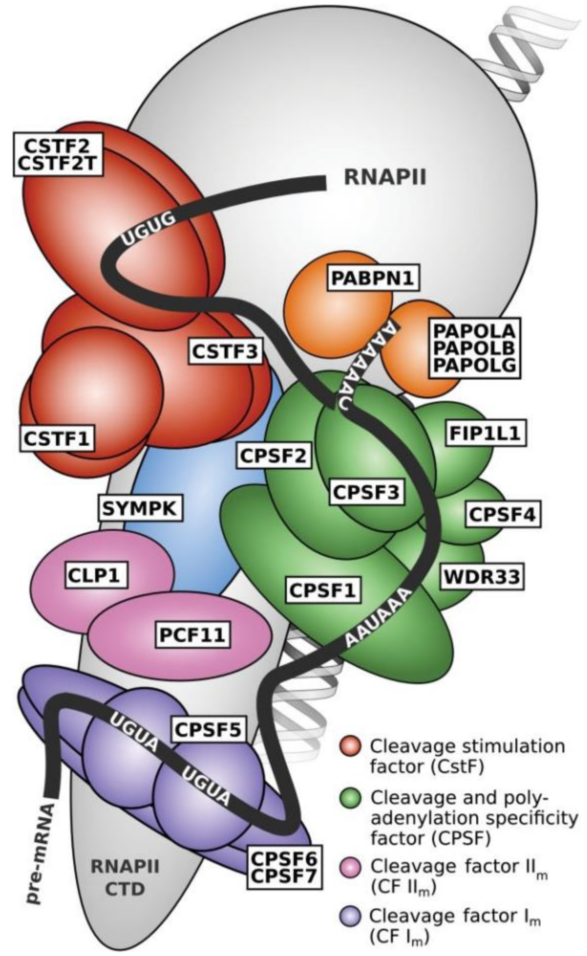

CSTF2
CSTF2T
RNAPII
UGUG
PABPN1
PAPOLA
PAPOLB
PAPOLG
AAAAAC
CSTF3
CSTF1
FIP1L1
CPSF2
CPSF3
CPSF4
SYMPK
WDR33
CLP1
CPSF1
AAUAAA
PCF11
CPSF5
UGUA
UGUA
pre-mRNA
RNAPII
CTD
CPSF6
CPSF7
● Cleavage stimulation factor (CstF)
● Cleavage and poly-adenylation specificity factor (CPSF)
● Cleavage factor II$_m$ (CF II$_m$)
● Cleavage factor I$_m$ (CF I$_m$)

Fig 5. The Cleavage and Polyadenylation Complex (CPA). The different colours indicate individual protein sub-complexes. Components of CPSF are shown in close proximity to the cleavage and polyadenylation site, where CPSF1 can recognise the AAUAAA polyadenylation signal. CPSF3 is the endonuclease that is responsible for cleavage of the mRNA. CF I$_m$ Is shown binding to UGUA motifs (upstream of the cleavage site), and the CstF complex can be shown interacting with a UG-rich region downstream of the cleavage site. Pcf11 is shown as part of the CF II$_m$ complex, together with CLP1 (schematic taken from Gruber et al., 2013).

Even though Pcf11's function in vertebrates and mammals has not been widely studied, cancer screening studies investigating mutations involved in cancer identified repeated Pcf11 mutations, mainly at the promoter region (Hornshøj et al., 2018, Kuipers et al., 2018, and Rheinbay et al., 2017). Also, expression levels of Pcf11 have been found to be predictive of neuroblastoma outcomes in patients (Ogorodnikov et al., 2018), indicating that Pcf11 might be relevant to human disease. Therefore, further research into the role of Pcf11 in human cells is relevant to understanding the mechanisms of mammalian transcription termination and gene expression. Pcf11 has also been found to affect transcription initiation rates since its depletion negatively affected Pol II levels and resulted in the downregulation in several genes (Mapendano et al., 2010).

1.2 The 3D Genome

1.2.1 3D Genome Structure

Understanding the influence of 3D genome organisation on gene regulation, evolution, and other cellular processes is a major biological question. Despite the enormous progress in the field, our knowledge regarding how chromatin structure is achieved and maintained remains limited. During the past twenty years, several studies have highlighted the importance of studying DNA and chromatin structure to understand how they affect spatial gene positioning, something that heavily impacts transcription, DNA repair, and replication (Therizols et al., 2014 and Gonzalez-Sandoval et al., 2015).

Hundreds of millions of base pairs make up the largest chromosomes. These chromosomes are then folded into different configurations. These folding stages involve nucleosomes and chromatin fibres, along with more complex structures such as chromosome domains and compartments. Eventually, the chromatin assembles itself into chromosome territories. For this reason, the organisation of 3D chromatin is a multiscale question. There remains a lot to be understood, from histone-DNA interactions at a more basic level, to chromosome-chromosome interactions at a more complex level. Additionally, this type of complex architecture can also be subjected to regulation by other elements such as non-coding RNAs, transcription factors, and proteins in order to regulate cell fate and expression of genes.

The human DNA in its unfolded and stretched-out state has a length of approximately 2 metres. In order to fit into the cell nucleus, it coils itself around histone proteins, forming nucleosomes. Histones are evolutionarily conserved proteins that assemble themselves into an octamer. Linker DNA is used to link the nucleosomes to each other. This level of organisation prevents other proteins from accessing the sequence, disabling transcription and other nuclear processes (Segal et al., 2006). The folding of the DNA into nucleosomes has been well described, however, how nucleosomes interact with each other remains unclear. At a higher resolution, loop formation of chromatin can be observed, and it sometimes involves regulatory elements. These interactions, although pivotal for cell identity, are not yet fully understood. A discovery made in the last ten years was able to show that besides individual chromatin loops, chromatin is also assembled in particular structural domains (Sexton et al., 2012, Nora et al., 2012, and Dixon et al., 2012) known as topologically associated domains (TADs), something that will be discussed in more detail in the following sections.

Even though the 3D architecture of the genome needs to be robust, it also needs to be quite flexible in order to allow for changes; for example, changes taking place before mitosis. It has been suggested that during development, the DNA structure remains robust, however, certain genes shift between chromosome compartments that are either active or inactive, while specific interactions between or within chromatin domains often alternate (Dixon et al., 2015). Genome-wide high-resolution maps of chromatin interactions that have been published in recent years have shown that organisation of 3D chromatin is a lot more complex than we previously thought (Rao et al., 2014 and Schuettengruber et al., 2014). Enhancer-promoter interactions, long-range interactions, subdomain organisation, and other significant features for development can only be accurately investigated with novel techniques and high sequencing depth. Thus, the ideal way of studying 3D genome architecture is through a combination of approaches. Microscopy methods are key when it comes to discovering information regarding the positions of genomic regions and looking at the variability in chromatin organisation between cell populations. However, microscopy methods are restricted to a small number of regions of interest. On the contrary, approaches based on chromosome conformation capture (3C) technology (Dekker et al., 2002). are genome-wide, however, the results might indicate superimpositions of different chromatin conformations instead of a single, stable structure.

Hi-C, a derivative of 3C able to produce high-throughput chromatin maps, was initially used to study the folding principles of chromatin (Lieberman-Aiden et al., 2009). Carbon copy chromosome conformation capture (5C) was later used, along with modelling approaches, to study the intercellular variability of DNA interactions within specific regions (Giorgetti et al., 2014). Single-cell Hi-C was also used to investigate the heterogeneity of 3D genome organisation within a cell population. Additionally, polymer modelling was implemented to demonstrate that chromosomes during the metaphase stage are assembled into a series of compressed loops (Naumova et al., 2013), which was in-line with previous microscopy findings (Marsden and Laemmli, 1979).

In conclusion, the configuration of the eukaryotic genome is a complex one. Its dynamic, three-dimensional (3D) organisation is closely related to several crucial biological processes, mainly DNA replication and ultimately gene expression and regulation. Modifications on this genome organisation can be destructive to organisms, potentially giving rise to a number of diseases, like cancers (Umlauf and Mourad, 2018). The study of these multiplex structures has been made possible due to developments in chromosome conformation capture (3C) methods, coupled with high-throughput sequencing (Hi-C). These advances have revealed the connection between genome organisation and nuclear architecture and how it differs between different processes, such as development and cell differentiation.

In the following sections, I briefly discuss 3C techniques, focusing on Chromatin Interaction Analysis by Paired-End Tag sequencing (ChIA-PET), before I introduce concepts relevant to eukaryotic DNA architecture.

1.2.1.1 Chromosome Conformation Capture (3C) methods

Over the past two decades, the emergence of chromosome conformation capture (3C) technology has made the high-resolution analysis of nuclear organisation possible. Back in 2002, Dekker and colleagues in the Kleckner lab at Harvard University formulated the idea that interaction frequency matrices between genomic loci could reveal the spatial organisation of different genomes. This formed the basis for the development of the 3C assay (Dekker et al., 2002). Briefly, the method involves the digestion and religation of DNA cross-links, followed by quantification of interactions between genomic loci. This reveals important information about chromosome topology and DNA-DNA interactions. The interactions could be completely random, resulting from chromatin collisions, or they can be direct and specific. For this reason, 3C methods require the use of suitable controls before any interpretations are drawn (Dekker et al., 2013). 3C technology has managed to formally show that during gene expression and repression, certain genes undergo gene looping (Tolhuis et al., 2002). Over time, different variants and upgrades of the initial 3C assay have been introduced (**Fig. 6**). These include circularised chromosome conformation capture (4C), chromosome conformation capture carbon copy (5C), high-throughput 3C (Hi-C), ChIP-3C (or ChIP-loop), and chromatin interaction analysis by paired-end tag sequencing (ChIA-PET).

ChIA-PET can be applied to detect chromosomal contact frequencies that have been mediated by specific proteins. It is similar to Hi-C, since it can detect genome-wide chromatin interactions. ChIA-PET is a candidate-specific method, as it includes an immunoprecipitation step which allows for the enrichment of DNA contacts bound by a particular protein of interest. The use of ChIA-PET was first developed and used to show that the human oestrogen receptor α (ER-α) is associated with long-range interactions between target gene promoters and regulatory ER-α binding sites (Fullwood et al., 2009).

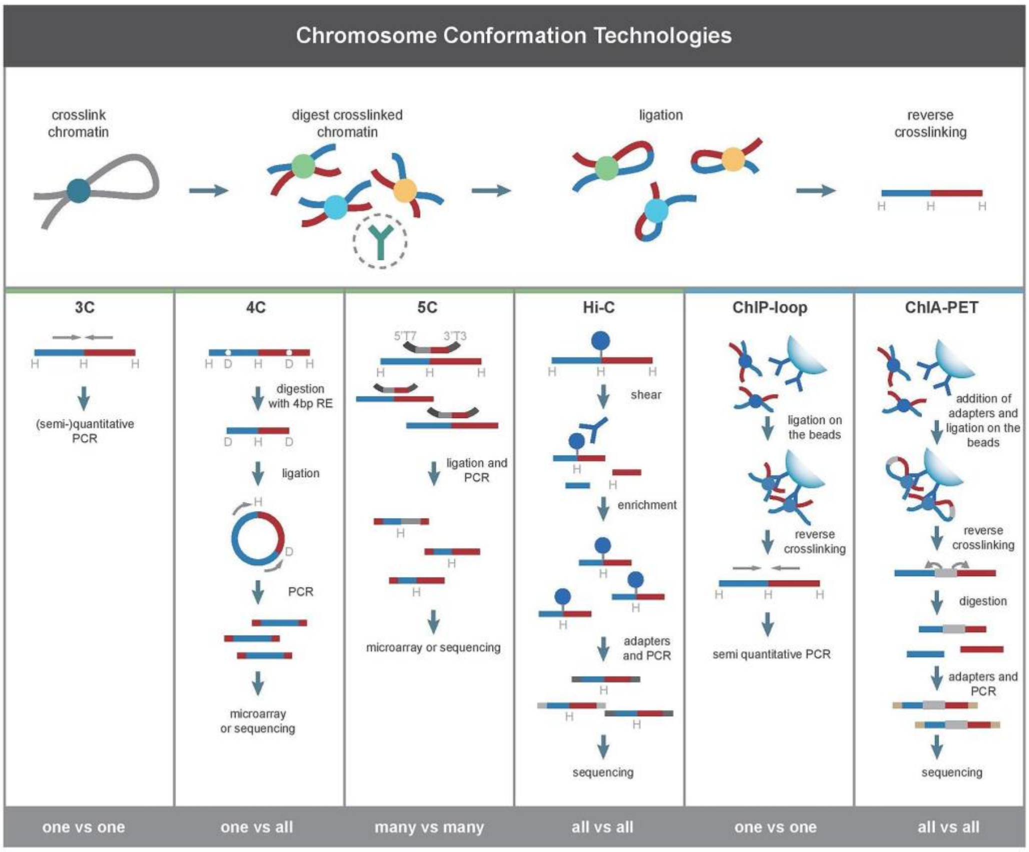


Fig. 6 Chromosome Conformation Capture (3C) technologies and derived methods. Top panel indicates the experimental reactions common in all methods. These include crosslinking of chromatin, digestion of crosslinked chromatin with a restriction enzyme, ligation of nearby fragments, and reverse crosslinking). Specific steps to each method are shown in the bottom panel under the method's name (e.g., PCR and sequencing). The ChIA-PET method is explained in depth in Chapter 2: Materials & Methods. (Schematic taken from De Wit and De Laat, 2012).

1.2.1.1.1 ChIA-PET

The experimental protocol for ChIA-PET is significantly different from the rest of the 3C family. Biotinylated linker sequences are integrated to cross-linked chromatin during proximity ligation. The chromatin is then sonicated and immunoprecipitated with a specific antibody. This results in the creation of hybrid molecules of DNA that consist of regions bound to protein factors. Biotin beads are used to pull the biotinylated linker sequences followed by paired-end tag sequencing. Through sequencing, protein factor-bound DNA interactions can be identified and quantified. The enrichment of genomic regions bound by a protein of interest allows the technique to yield high resolution reads at a lower sequencing depth and makes it possible to identify biologically relevant DNA contacts.

The use of ChIA-PET has not been as wide as that of other 3C methods, however, several significant studies that have utilised the technique have been published in recent years. Developed by the Ruan and Cheung Labs in Singapore, as previously mentioned, ChIA-PET was first applied in 2009 to comprehensively study the human chromatin interactome bound by oestrogen receptor α (ER-α). It was found that the majority of remote, high-affinity ER-α binding sites are bound at gene promoters via long-range DNA contacts, indicating that ER-α coordinates transcriptional regulation by bringing genes together through chromatin looping (Fullwood et al., 2009). It was thus proposed that one of the main mechanisms of transcription regulation in mammals is chromatin interactions. In subsequent high-impact studies during the following couple of years, the Ruan group had utilised the technique further to investigate the interactome of CTCF in pluripotent cells. They were able to discover five different chromatin domains and show that CTCF interactions affect gene expression through crosstalk of regulatory elements and gene promoters (Handoko et al., 2011). Additionally, they implemented ChIA-PET to study interactions associated with Pol II in human cells. They uncovered universal promoter-focused interactions and showed that the majority of genes in possession of promoters that interacted with one another were not only active but also transcribing in a collaborative manner (Li et al., 2012). As the technique was becoming more popular, other groups started utilising it as well. In 2013, ChIA-PET was used to study the interactome map of regulatory domains (Kieffer-Kwon et al., 2013) and cohesin-associated interactions (DeMare et al., 2013) in mice. Cohesin ChIA-PET was also performed on embryonic stem cells and revealed the existence of insulated neighbourhoods - looped formations important for appropriate expression of nearby genes (Dowen et al., 2014). It was not long before the question of whether chromatin interactions play a role in disease was posed. Frequently, cancer cells rearrange their genome and cause disorganisation with respect to standard 3D architecture. Disrupting 3D neighbourhoods has been found to alter oncogene regulation and pass on their control to regulatory elements that would normally be kept away from them, wrongly activating them (Hnisz et al., 2016). Apart from improvements with the original ChIA-PET wet-lab protocol, the emergence of such complex datasets had also prompted the design of sophisticated and robust computational tools for analysis, further facilitating the data mining process.

As a result of these advancements in 3C technology giving rise to high-resolution interaction maps, we were able to look deeper into nuclear architecture like never before.

1.2.1.2 Eukaryotic nuclear architecture

The significance of DNA compartmentalisation and higher-order nuclear structure when it comes to control and regulation of cell life is well established. The DNA of more complex eukaryotes is ordered into chromosome territories and the 3D organisation of these chromosome territories might be relevant to cellular processes such as transcription, gene expression, DNA repair, and genomic function.

Chromosomes are made up of chromatin which consists of DNA and proteins. Depending on the level of chromatin compaction as well as transcriptional activity, chromatin can be observed in two distinct and spatially separated states inside the nucleus. Therefore, the distribution of chromatin is not random. These discrete areas of chromosome occupancy are called chromosomal territories (CTs) and their distribution within the nucleus seems to depend on their gene density (Cremer and Cremer 2010). The concept of CTs has a long history and dates to the early 1900s when cell biologist Theodor Boveri coined the term (Cremer and Cremer, 2006). Cremer and Cremer (2006) used laser light to cause DNA damage and took advantage of the ability of cells to repair DNA by providing radioactively labelled nucleotides, which the cell incorporated into its DNA during the process of repair. The regions were then analysed by radiography once the cell entered the next mitosis, and its chromosomes became condensed. Highly condensed chromatin, also known as heterochromatin, is gene-poor and transcriptionally inactive. CTs made up of gene-poor chromosomes are preferentially localised in the periphery of the nucleus, close to the membrane. In contrast, euchromatin, which is gene-rich, transcriptionally active, and lightly packed, is localised within CTs that are in a more central position within the nucleus. This pattern of organisation seems to be observed in most eukaryotic cell types (Falk et al., 2019).

Another model suggests that it is not just the CTs themselves that are spatially organised, but also the genes within them. In general, inactive genes are found within interior regions of CTs, as opposed to active genes that conglomerate closer to the periphery near the interchromosomal region (Cremer et al., 2001). However, evidence suggests that this is not always the case. For example, research studying a particular locus on chromosome 11 containing different types of protein coding genes showed that the region is found within a distinct CT, and that after activation of certain genes, the genes did not relocate closer to the periphery of the CT (Mahy et al., 2002a,b). Chromosome arrangement based on size was also detected, with chromosomes that are smaller in size being localised closer to the centre of the nucleus (Sun et al., 2002; Bolzer et al., 2005, and Maharana et al., 2006). However, noticeable single-cell variations on how CTs are arranged do exist (Szczepinska et al., 2009).

In general, once the arrangement of CT neighbourhoods has been established during the beginning of interphase, it stays that way until the following prophase stage (Cremer and Cremer, 2010). Although, several genes do change positions as a response to stimuli during differentiation and disease (Volpi et al., 2000; Williams et al., 2002, Meaburn et al., 2009; Ferrai et al., 2010). Sometimes, gene loci are repositioned to new areas of the nucleus due to the movement of the entire CT, however, genes are able to relocate by the formation of chromatin loops outside the main CT body (Bridger et al., 2011). It has been shown that the chromatin that loops out of CTs becomes associated with other structures of the nucleus, such as transcription factories (Eskiw et al 2010) and Cajal bodies (Dundr et al., 2007).

1.2.1.3 Chromatin Loops

Gene loops are formed when the promoter and terminator regions of a gene come in physical contact with each other, a phenomenon that seems to be transcription dependent. A lot of long-range chromatin interactions also involve contacts between enhancers and promoters. Such interactions are thought to form active chromatin hubs where large numbers of transcription factors and Pol II are concentrated, leading to transcription. Evidence gathered through genome research indicates that regulatory DNA sequences are able to control the process of transcription from a distance by interacting with target genes via chromatin loops (Holwerda and Laat, 2012). Studies on the β-globin locus provided early evidence that chromatin loops are involved in the gene regulation of mammalian organisms. It was suggested that transcription of genes was controlled by distant regulatory sequences, known as the β-globin locus control region (LCR), due to the observation that deletion of sequences far away from the β-globin locus were suppressing the gene’s activity and, thus, causing thalassemia (Van der Ploegh et al., 1980). Direct evidence that looping and long-range interactions allow distal regulatory elements to control the β-globin locus appeared in 2002 with the use of 3C technology (Tolhuis et al., 2002).

Looped-like structures have been observed at several gene loci in a manner that juxtaposes key genetic components. Chromatin interactions have also been detected between elements located on different chromosomes, and even though these sorts of interactions do not fit the definition of chromatin loops, they might have similar functions (Kadauke and Blobel, 2009). Early electron microscopy studies were able to reveal the existence of chromosomal loops in chromosomes preparing for metaphase (Marsden and Laemmli, 1979), as well as interphase chromatin where loops had been found to be anchored on the nuclear matrix (Heng et al., 2001 and Martelli et al., 2002). The word “looping” has also been used to refer to sections of chromosomes that protrude outside of their chromosomal territory (Kosak and Groudine 2004, Cremer et al., 2006), although their impact on gene expression and regulation is not well understood.

How chromatin loops are formed is yet another not very well understood phenomenon. However, a large number of TFs have been found to be involved in the formation of loops. The GAGA factor in *drosophila*, which is coded by the *Trl* gene, is known to activate expression of several genes by blocking the suppression imposed by heterochromatin. In vitro transcription assays and reporter constructs in yeast provide evidence for the GAGA factor contributing to looping formation (Mahmoudi et al., 2002 and Petrascheck et al., 2005). Studies employing 3C technology were able to show that a combination of transcription factors was necessary for juxtaposing the β-globin LCR with the active β-globin promoter (Drissen et al., 2004 and Song et al., 2007). Therefore, these phenomena suggest long-range interactions are mediated by specific TF combinations.

1.2.1.3.1 Impact of chromatin loops in gene expression

The formation of gene loops has been shown to correlate with regulation of transcription of the interacting loci, however, it is unclear what the functional significance of long-range chromatin interactions is. A vast amount of 3C studies have established that genes and their putative regulatory elements form loops between them, but what does this mean for gene regulation? On its own, 3C only identifies DNA sequences that most frequently interact with each other, in vivo. To understand how these loops are formed or infer function, these changes in chromatin structure need to be looked at alongside other epigenetic effects, e.g., regulation of transcription, binding factors, or histone modifications. However, even then, it is hard to determine whether the structural changes are a result or a cause of gene regulation. A common theme that emerges after collating 3C studies is that chromatin loops seem to have a functional role.

It appears that, often, chromatin loops are formed between protein factor binding sites. Several proteins have been involved in regulatory loops, such as TFs, Polycomb proteins, insulator proteins, nuclear architecture proteins, and chromatin remodelling factors (Sexton et al., 2009). A number of different studies where knockouts and knockdowns of such protein factors were produced showed that looping between bound sites is disrupted (Spilianakis et al., 2004; Drissen et al., 2004; Vakoc et al., 2005; Cai et al., 2006; Splinter et al., 2006; Lanzuolo et al., 2007) and in some cases, chromatin conformation can be rescued by the expression of the protein factor in question. Kurukuti and colleagues (2006) conducted a study where they induced a mutation of a binding site for CTCF protein within a particular imprinted locus. This resulted in abrogation of looped interactions mediated by CTCF and it had also caused de novo DNA methylation at a normally interacting, distal CTCF site. Therefore, the binding of proteins to their sites can be linked to both chromatin structure and epigenetic regulation. Although not every genomic locus and protein factor exhibit the same behaviour.

1.2.1.3.2 Long-range genomic interactions

Over time, it has also become apparent that it is not just intrachromosomal interactions that are required for proper gene expression, but also interchromosomal ones. Is there a specified range of genomic distance in which regulatory elements can affect targeted genes? Studies have shown that enhancers required for gene expression have been found megabases away from the target loci, and evidence showed that they do physically interact with each other (Lettice et al., 2003 and Amano et al., 2009). Several 3C, FISH, and other high-throughput screening studies have identified even more extreme chromatin interactions. The functional significance of these long-range contacts remains unclear. Some of those interactions may be occurring so that regulatory elements can come into contact with target genes, in a similar fashion with short-range interactions involving communication between enhancers and promoters. Others may be more indirect, involving genes that are co-regulated and share common nuclear domains, and as a result are exposed to the same regulatory elements. Also, interchromosomal interactions might just be consequential events resulting from the compaction of the genome inside the nucleus, although this hypothesis has been argued against (Sexton et al., 2009).

It is therefore important to uncover the mechanisms which result in the formation of interchromosomal interactions, as this would contribute to our understanding of how chromosomal territories and regulatory DNA elements crosstalk between them. Microscopy studies published over a decade ago observed “intermingling” chromosomes as regions that overlap between chromosomal territories (CTs) (Branco and Pombo, 2006; Cremer and Cremer, 2010). One of the best known interchromosomal interaction formations involves the preassembly of ribosomes by rRNA genes, which are encoded by five distinct chromosomes, to create the nucleolus (McStay, 2016). The spatial formation of the nucleolus confirms that non-homologous chromosomes can come into contact with each other, in a manner that is not random. In a similar fashion, olfactory receptor genes are present in many distinct chromosomes. To regulate their expression, a complex topological ordering is required for all of them to aggregate in the same nuclear position - forming the “olfactosome” (Lomvardas et al., 2006; Monahan et al., 2017,2018). Additionally, chromosomal interactions have been found to be able to activate genes on other chromosomes (Spilianakis et al., 2005). Also, lncRNA loci form interchromosomal contacts affecting gene regulation processes in both disease and health (Maass et al., 2012; Hacisuleyman et al., 2014).

1.2.1.4 Topologically Associating Domains (TADs)

In recent years, lower resolution (40 kb) Hi-C maps showing the interaction frequency of large numbers of cells revealed the existence of topologically associating domains (TADs) (**Fig. 5a-d)** (Dixon et al., 2012, Nora et al., 2012, Hou et al., 2012, and Sexton et al., 2012). TADs are formed when several genomic interactions between 100 kb - 1 Mb are clustered together (Dixon et al., 2012, Nora et al., 2012). Genomic regions located within the same TAD come in contact with one another at a higher frequency than regions found in adjacent domains, making TADs hubs for intrachromosomal interactions. Topologically partitioning the genome into TADs correlates with several genomic characteristics, like synchronised gene expression, timing of DNA replication, and histone modifications (Dixon et al., 2012). It also appears that interactions between enhancers and promoters are mostly restricted within TADs (Shen et al., 2012). In more recent single-cell studies, it has been observed that TADs vary greatly within individual cells (Nagano et al., 2013, 2017, Ramani et al., 2017; Stevens et al., 2017). Deleting the boundaries that separate TADs from one another has been found, in some cases, to cause disease phenotypes (Franke et al., 2016; Taberlay et al., 2016). Although, in other studies, such perturbations were not found to lead to biologically significant phenotypic changes (Barutcu et al., 2018), with the exception of when longer regions ranging from 200-400 kb were deleted (Nora et al., 2012; Rodríguez-Carballo et al., 2017).

A model known as "loop-extrusion" (**Fig. 7e**) can explain how the genome is organised, as well as the interactions it forms within TAD structures. The model explains how long-range, intrachromosomal interactions bring regulatory elements - that are otherwise far away - into close proximity to targeted loci (Fudenberg et al., 2016). The model proposes that *cis*-acting factors (factors that are in close proximity), such as cohesin, begin to form loops that get progressively bigger and stall at the boundaries of TADs due to contacts with boundary molecules, such as CTCF (CCCTC-binding factor). These *cis*-acting elements move towards one another while the intervening DNA is being extruded. The researchers showed that each TAD results from several loops formed by extrusion, as opposed to one single static loop (Fudenberg et al., 2016). This model has been shown to be mediating intrachromosomal interactions and is responsible for the formation of most TADs (Fudenberg et al., 2016; Golodoborodko et al., 2016; Ganji et al., 2018). A study that was published not too long ago has introduced the likelihood that large-scale genomic regions could also be influenced by the mechanism of loop extrusion (Nuebler et al., 2018).

Compartmentalisation is another mechanism that contributes towards the maintenance or establishment of chromatin domains in higher eukaryotes. Compartments were first identified through their 'plaid'-like pattern of extremely long-range intrachromosomal and interchromosomal interactions (Lieberman-Aiden et al., 2009). It is hypothesised that his pattern of partitioning represents the subdivision of the human genome into two types of compartments, defined as A and B. A compartments contain genes that are actively transcribed and active histone marks, whereas B compartments contain genes that are inactive along with repressive marks (Lieberman-Aiden et al., 2009). Initially, lower resolution Hi-C maps had suggested that several TADs on the megabase scale are nested inside one single, contiguous region of either an A or B compartment. Although, it appears that when higher resolution heatmaps emerged, the mammalian genome partitions into six or more much smaller sub compartments instead. These sub compartments are made up of different combinations of both active and inactive chromatin modifications (Rao et al., 2014). Remarkably, ultra-high resolution Hi-C maps in invertebrates have revealed the existence of 'compartment domains', which refer to very faint compartments (Rowley et al., 2007).

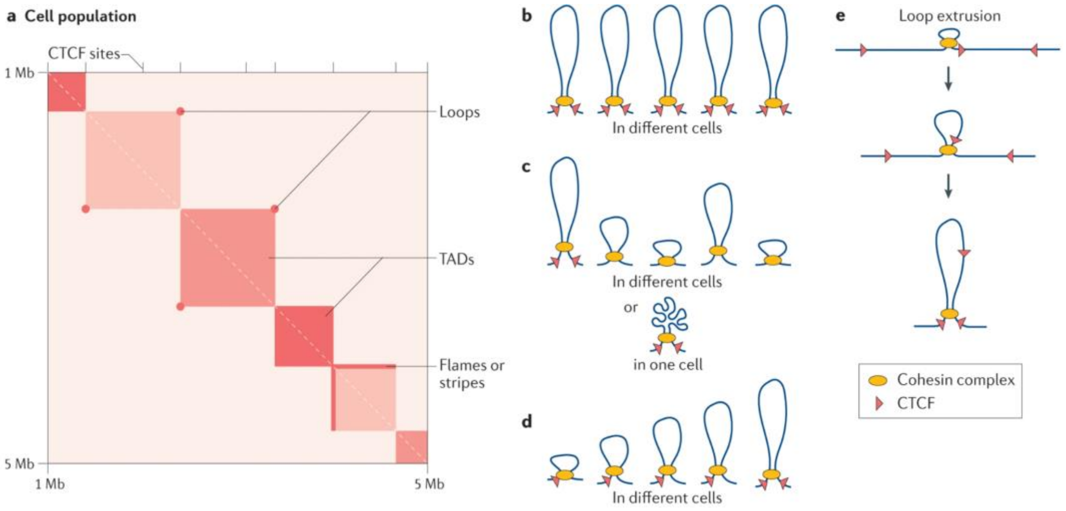


Fig. 7 Genome organisation and the loop extrusion model. A) Schematic of a 4 Mb region in a Hi-C map of a single chromosome, indicating how loops and TADs are arranged across a homogenous mammalian cell population. Vertical lines at the top of the matrix correspond to CTCF enrichment as identified by ChIP peaks. **B)** Dots (loops) on a Hi-C map represent chromatin loops formed through interactions occurring amongst the same genomic sequences in several cells. **C)** Triangular TADs on a Hi-C map are a result of chromosome domains which contain genomic sequences that appear to interact between each other more frequently than they interact with sequences outside that particular chromosomal domain. This TAD pattern could be the result of the mean of many loops occurring at different locations in several cells,

within the same TAD (top). Otherwise, it could be a result of loops that come into contact with chromatin at several positions along their length (bottom). In the case where a loop spanning the entirety of a TAD is present at several cells, this will be represented as a dot on the corner of the TAD. **D)** Flames, or stripes, represent the occurrence of a specific sequence found to be interacting with several sequences in a population of cells. **E)** Loop extrusion model mediated by cohesin. The cohesin complex extrudes a loop upon association with chromatin, until it encounters CTCF. This results in a chromatin loop with CTCF molecules at its base and cohesin on top. (Schematic taken from Davidson and Peters, 2021).

Another theme that has been recently introduced to justify the formation of cellular substructures is that of liquid-liquid phase separation. DNA and RNA both come into contact with proteins harbouring regions of low complexity (Van der Lee et al., 2014 and Protter at al., 2018) and form aggregates in either solid, gel, or liquid form that could be compartmentalising and shaping the genome (Erdel and Rippe, 2018, Langdon et al., 2018, and Maharana et al., 2018). Droplet-like structures, created through condensation, are formed inside the nucleus resulting in the rearrangement of chromatin.

Further technical advances in Hi-C technology allowing for higher resolution (1-4 kb) revealed the existence of smaller chromatin domains, called sub-TADs. In Hi-C maps of mammalian cells, sub-TADs are hierarchically located inside TADs (Phillips-Cremins et al., 2013 and Rao et al., 2014). These nested sub-TADs have a similar structure to TADs, however, the boundaries they are demarcated by show lower insulation robustness. This is indicated by their weaker capacity to reduce long-range interactions between domains (Dixon et al., 2012 and Norton et al., 2018).

The discovery of the loop extrusion model along with the mechanism of compartmentalisation has caused intense competition among looping mechanisms regarding how domains are formed. As a result, a topic that is now under hot debate is how to best update the definitions of TADs and sub-TADs (Rowley et al., 2017, Rao et al., 2017, and Schwarzer et al., 2017). Research suggests that compartmentalisation and loop extrusion are both separate and competing forces, therefore highlighting the importance of being able to clearly and uniquely define the two mechanisms. In a recent review on

TADs, the authors have suggested adding additional qualifiers to reflect the newly discovered mechanistic models (Beagan and Philips-Cremins, 2020). Certainly, as 3C and Hi-C technologies advance allowing us to look deeper and deeper into genome structure in ultra-high resolution, the need for more accurate and precise definitions will only grow.

1.2.1.4.1 Impact of chromosome (TAD) rearrangement

Existing TADs can be disrupted due to genetic rearrangements. For example, inversions, deletions, duplications, or translocations, they can all result in the formation of de novo TADs. These newly formed TADs usually show functionality issues due to misregulation of gene expression, resulting in diseased states such as cancer. A comprehensive study investigating structural variations in cancer genomes using Hi-C, whole genome sequencing, and next-generation optical mapping, revealed that cancer genomes showed several structural variations when compared to normal cells which did not have any of those variations (Dixon et al., 2018). It was also observed that, based on an overall analysis of all Hi-C maps, newly formed TADs are often formed due to large-scale chromatin rearrangements in cancerous cells. An example of this is the de novo TAD appearing in a human cancer cell line as a result of chromosomes 9 and 18 fusing with each other (Dixon et al., 2018). Additionally, further research on how newly formed TADs affect gene expression has revealed that in several cancer cell lines, genes that are found inside TADs with structural rearrangements show higher allelic biases in comparison to genes that are found in TADs that have not been rearranged. This strongly argues that at least to a certain degree, structural variations result in changes in gene expression (Dixon et al., 2018).

Another Hi-C study involving the use of normal and prostate cancer cells has also revealed that cancer cells tend to generate additional TAD boundaries (Achinger-Kawecka et al., 2016). It was also demonstrated that the majority of boundary domains found in normal cells were also found in cancer cells. Unexpectedly, it was shown that a significant proportion of the newly formed, smaller TADs that were generated in both cell lines had similar new boundaries. The boundaries of newly formed TADs usually correlate with changes in copy number variations and portray bias in levels of gene expression that is also observed in cancer genomes (Achinger-Kawecka et al., 2016). The same study also revealed that new, cancer-specific chromatin loops exist inside the smaller TADs. These were found to be enriched for promoters, enhancers, and binding sites for CTCF. These findings appear to confirm the theory of enhancers participating in regulating local interactions separating cancer and normal cells (Achinger-Kawecka et al., 2016).

1.3 The Present Work

The present thesis will focus on identifying gene looping in human cell lines and seeking to understand how the termination factor Pcf11 localises at the TSS of genes. To do this, I will first present experiments performed to establish a Pcf11 knock-down, as well as optimise the *in-situ* ChIA-PET technology to generate quality libraries. I will perform extensive QC on all libraries used in this work before proceeding to isolate and characterise gene loops bound by Pcf11 and Pol II. To enhance and support the findings of this work, I will analyse publicly available datasets from ChIP-seq and RNA-seq experiments. The main goal of this work is to determine the quality of ChIA-PET samples by applying best practices for QC, as well as attempt to answer the question of whether Pcf11's localisation at the start sites of genes is a result of genome-wide 3' - 5' end crosstalk.

Chapter 2: Materials and Methods

2.1 Cell Culture and Dual Crosslinking

HEK-293, HeLa, and U2OS cells were thawed from - 80 °C for initial culturing. Cells were grown in 500 cm^2 tissue culture-treated culture dishes (Corning®) with Dulbecco's Modified Eagle Medium and foetal bovine serum (DMEM,10 % FBS) at 37 ^oC, 5 % $CO_{2.}$ At around 80 % confluency, cells were harvested and crosslinked – [for Flavopiridol-treated HEK-293 cells, 300 nM of Flavopiridol were added for 45 minutes prior to crosslinking. Equal amounts of DMSO were added to the control cells for equal time]. The media was removed from the culture dishes and the monolayer was rinsed with 30 ml of phosphate-buffered saline (PBS), twice. After rinsing, a solution of 50 ml phosphate buffer saline (PBS) and 1.4 ml of 37 % formaldehyde was added to each dish. Dishes were placed on a shaker (100 rpm) for 20 minutes. To quench the crosslinking reaction, 0.2 M of glycine was added to each dish for 10 minutes. Cells were then washed with 30 ml of PBS. For the next crosslinking reaction, 2 mM of ethylene glycol bis (EGS) in 50 ml of PBS (37 °C) were added to each dish before placing on the shaker (100 rpm) for 45 minutes. To quench the crosslinking reaction, 0.2 M of glycine was added to each dish (10 minutes, 100 rpm). The solution was removed, and cells were washed with 30 ml of PBS, twice. Cells were scraped from the dish with cell scrapers and collected in 50 ml Falcon tubes. Tubes were then centrifuged at 2000 rpm for 10 minutes at 4 °C. The supernatant was discarded, and crosslinked pellets were stored in - 80 °C for future use.

2.2 ChIA-PET Library Preparation

The following steps were adapted and optimised accordingly based on the in-situ version of the original ChIA-PET protocol. This protocol was kindly sent to me by Dr Ping Wang, a postdoctoral fellow in the Ruan Lab of The Jackson Laboratory, CT, USA, whose lab along with the Cheung Lab pioneered the technique back in 2009. This version of the protocol with minor changes was later officially published in Current Protocols (Wang et al., 2021).

DAY 1

Cell Lysis

Crosslinked pellets (~ 10M cells) were taken out of the - 80 °C freezer and thawed on ice for 20 minutes. Pellets were then centrifuged at 200 rpm for 5 minutes to discard any residual buffer supernatant. Cells were lysed using 0.1 % SDS Cell Lysis Buffer (**Table 1**) containing a mini proteinase inhibitor tablet (Roche Inc.). The tubes containing 1 ml of 0.1 % Cell Lysis buffer were placed on a rotator for 1 hour at 4 °C. Tubes were then centrifuged at 2500 g for 5 minutes and the supernatant was discarded. For nuclear lysis, the pellet was suspended in 100 µl of 0.55 % SDS solution containing a mini protease inhibitor tablet. The tubes were incubated at room temperature for 10 minutes, then at 62 °C for another 10 minutes, followed by 10 minutes at 37 °C. To quench the SDS reaction, 270 µl of water and 5 µl of Triton X100 were added to the tubes. Tubes were placed on the shaker for 15 minutes at 37 °C. **Fig. 1** shows a schematic of the full in-situ ChIA-PET protocol.

Table 1 Cell lysis buffer composition.

Cell Lysis Buffer	50 mM HEPES-KOH pH 7.5
	150 mM NaCl
	1 mM EDTA
0.1 % SDS	1% Triton X-100
	0.1% Sodium Deoxycholate
	0.1 % SDS
	Molecular Grade Water

Restriction Digestion

For the restriction digestion, 50 µl of 10x CutSmart buffer (NEB) and 30 µl of *AluI* restriction enzyme (NEB) were added to each of the cell pellet tubes from the previous step. Restriction digestion tubes were incubated in the thermomixer at 37 °C overnight.

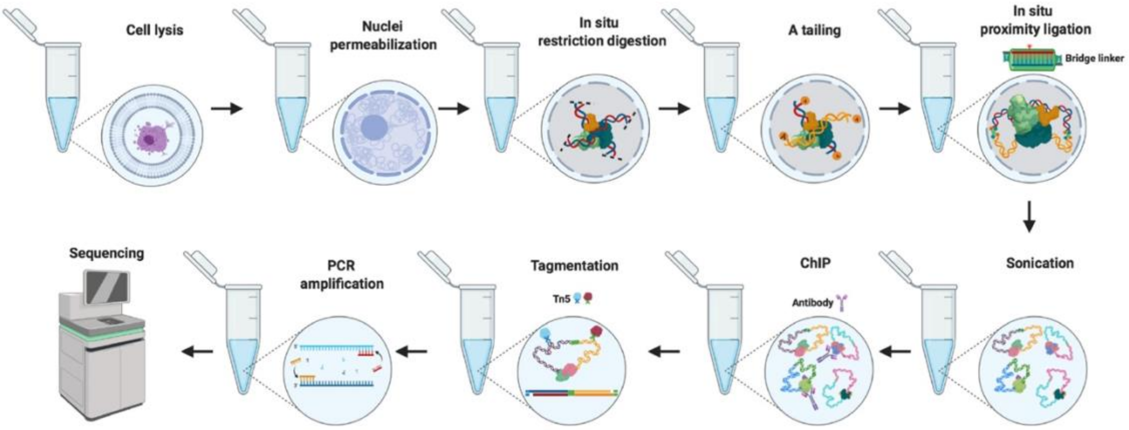


Fig. 1 Schematic of the *in-situ* ChIA-PET protocol. Isolated nuclei from dual crosslinking (EGS and Formaldehyde) are permeabilized, and *AluI* (four-nucleotide cutter restriction enzyme) is added to the reaction to perform *in-situ* digestion. The proximity ligation reaction is also performed *in-situ*, with the use of a bridge linker, following the A-tailing step. After chromatin sonication, ChIP against either Pcf11 or Pol II is carried out. The de-crosslinked ChIP DNA is tagmented using Tn5 transpose, and PCR is performed to amplify the ChIA-PET library. The library is then ready for sequencing (taken from Wang et al., 2021).

DAY 2

Single-tube A-tailing/Proximity Ligation

A-tailing

The digested samples were left at room temperature and then the following reagents were added to the tubes:

Reagent	Amount (µl)
BSA (20 mg/ml)	11
10 mM dATP	11
10x Cutsmart buffer	4
H_2O	3
Klenow fragment (3’→5’ exo) (NEB 5U/µl)	11

Tubes were mixed and incubated at RT for 1 hour in the thermomixer. The tubes were then incubated at 37 °C for an hour on the tube revolver rotator (ThermoFisher).

Proximity Ligation

After the A-tailing reaction, tubes were left at RT. The following reagents were added to each of the tubes for the proximity ligation reaction:

Reagent	Amount (µl)
H_2O	260
NEB 5x quick ligation reaction buffer	200
Bridge Linker (200 ng/µl) (*See Bridge Linker Prep section for details*)	4
T4 DNA ligase (NEB, 400U/µl)	10

Tubes were incubated at RT for 1 hour and then transferred to a 16 °C shaking incubator for overnight ligation.

Coating Protein G beads with Antibody

Protein G beads (ThermoFisher) were removed from the 4 °C fridge and mixed thoroughly. For each sample, 100 µl of G beads were prepared for the purpose of immunoprecipitation (IP). The beads were washed with 1x 500 µl PBST buffer for 3 times and were then resuspended in 1x PBST buffer. For each library, 20 µg of antibody was mixed with the washed protein G beads and tubes were left rotating overnight at 4 °C for the purpose of the antibody binding to the protein G beads. Antibody product numbers are listed below:

Antibody	Product number
Pcf11	ab134391 (Abcam)
Pol II [1C7]	ab252854 (Abcam)
Pol II [4E12]	ab252853 (Abcam)
Pol II [3E8]	ab252852 (Abcam)

DAY 3

Sonication

Tubes were removed from the 16 °C incubator and spun at 5500 g for 10 minutes. The supernatant was removed, and the ligation mix pellets were gently resuspended in 500 µl of 1x PBST buffer. Tubes were sonicated using Diagenode's Bioruptor on 'High' setting, for 15-30 minutes, with 30 second on/off intervals. Sonication conditions were optimised to generate chromatin fragments within the 2-4 kb range.

"Preclearing" Protein G beads wash

For each sample, a corresponding 100 µl of G beads were washed 3 times with 1 ml 1x PBST buffer. The wash buffer was discarded using the magnetic separation rack (NEB). Beads were left in 200 µl of 1x PBST at 4 °C until use for "preclearing".

Chromatin Preclearing

For the preclearing process, the sonicated chromatin from step 6 was mixed with the newly washed beads from step 7. Tubes were incubated at 4 °C for 1 hour with rotation to remove any non-specific binding material. After the preclearing reaction, ~1 ml of chromatin was collected and transferred from each tube to a new tube and kept on ice. Preclearing beads were discarded.

Preparation of Antibody-coated beads

The antibody-coated beads from step 5 were washed twice with 1 ml 1xPBST buffer and left on ice with 1xPBST buffer until use.

Immunoprecipitation (IP)

The precleared chromatin suspension from step 8 was added to the washed antibody-coated beads for IP. Tubes were incubated overnight at 4 °C.

DAY 4

Washing IP product and purifying antibody-enriched IP DNA

After overnight IP, the supernatant of each tube was removed and discarded using the magnetic rack. The antibody-captured chromatin on the G beads was washed via the following procedure:

Table 2 details the composition of each buffer.

- **3x washes with Low Salt buffer:** Beads were resuspended in 1 ml of low salt buffer and placed on the rotating rack at 4 °C for 5 minutes. Tubes were short spun and placed on the magnetic rack and supernatant was discarded.

- **2x washes with High Salt buffer:** Beads were resuspended in 1 ml of high salt buffer and placed on the rotating rack at 4 °C for 5 minutes. Tubes were short spun and placed on the magnetic rack and supernatant was discarded.

- **1x wash with LiCl buffer:** Beads were resuspended in 1 ml of LiCl buffer and placed on the rotating rack at 4 °C for 5 minutes. Tubes were short spun and placed on the magnetic rack and supernatant was discarded.

- **2x washes with Tris/EDTA (TE) buffer:** Beads were resuspended in 1 ml of Tris/EDTA buffer and placed on the rotating rack at 4 °C for 5 minutes. Tubes were short spun and placed on the magnetic rack and supernatant was discarded.

TE buffer was discarded, and DNA was eluted from the beads with 200 µl of ChIP Elution buffer at 65 °C for 30 minutes in the thermomixer, shaking at 900 rpm. The tubes were then placed on the magnetic rack and the supernatants were transferred to new tubes. Beads were washed with a 100 µl elution buffer EB (Qiagen). The 100 µl wash buffer and the 200 µl solution were then combined together and 20 µl of Proteinase K (10 mg/ml) was added to the buffer. Tubes were placed in the thermomixer at 65 °C overnight for decrsosslinking.

Table 2 Composition of washing and elution buffers.

Low Salt Buffer (Cell lysis buffer)	50 mM HEPES-KOH pH 7.5 150 mM NaCl 1 mM EDTA 1% Triton X-100 0.1% Sodium Deoxycholate 0.1 % SDS Molecular Grade Water
High Salt Buffer	50 mM Tris-HCl pH 7.5 350 mM NaCl 1 mM EDTA 1% Triton X-100 0.1% Sodium Deoxycholate 0.1 % SDS Molecular Grade Water
LiCl Wash Buffer	10 mM Tris-HCl pH 8.0 250 mM LiCl 1 mM EDTA 0.5 % IGEPAL® - CA-630 0.5 % Sodium Deoxycholate
TE Buffer	10 mM Tris-HCl 1 mM EDTA•Na_2
ChIP Elution Buffer	50 mM Tris-HCl pH 7.5 10 mM EDTA 1% SDS

	Molecular Grade Water

DAY 5 – DAY 6

Tagmentation of ChIP-DNA

In a PCR tube, the following mixture was prepared for each sample using Ilumina’s Tagment DNA Enzyme and Buffer kit:

Tagmentation Component	Amount
Proximity ligated DNA	30-50 ng
Tagmentation Buffer	25 µl
Transposase Enzyme (TDE1)	5 µl
Nuclease-free Water	X µl to raise volume to 50 µl
Total	***50 µl***

Tubes were short spun and incubated at 55 °C for 5 minutes and then at 10 °C for 10 minutes in a PCR machine.

Tagmentation DNA was purified using the Zymo DNA Clean & Concentrator kit (Zymo Research) according to manufacturer’s instructions. The tagmentation reaction was repeated for the remainder of proximity ligated DNA. All purified tagmented DNA was combined in preparation for the next step.

Immobilisation of DNA library to streptavidin dynabeads

M280 Streptavidin Dynabeads (ThermoFisher) were taken out of the 4 °C fridge and placed on the bench for 30 minutes to come to RT. The beads were mixed well and for each sample, 30 µl of suspended dynabeads were pipetted out and into a new 1.5 ml Eppendorf tube. Tubes were placed on the magnetic rack to discard the supernatant and were then washed with 150 µl 2x Binding & Washing buffer (**Table 3**), twice. Beads were resuspended in 100 µl of iBlock buffer, mixed, and incubated at RT for 45 minutes with rotation. After incubation, tubes were short spun and placed on the magnetic rack to discard the iBlock buffer. Beads were then washed with 1x Binding & Washing buffer, twice. The buffer was discarded and 100 µl of genomic DNA mixture were added to each tube for blocking. Tubes were mixed well and then incubated at RT for 30 minutes with rotation. The blocking DNA mixture was then discarded, and beads were washed with 200 µl of 1x Binding & Washing buffer, twice. All of the tagmented DNA library product of each sample was added to the tubes containing the washed streptavidin beads. An equal amount of 2x Binding & Washing buffer was added to each tube. Tubes were mixed well and incubated at RT for 45 minutes with rotation. After incubation, tubes were short spun and placed on the magnetic rack to discard the supernatant. Beads were then washed with 500 µl 0.5 % SDS/2X SSC buffer, five times. The beads were washed with 500 µl 1x Binding & Washing buffer, twice. All buffer was then discarded, and beads were gently resuspended in 30 µl of EB buffer. **Table 3** details the composition of each of the buffers used for this step.

Table 3 Composition of Streptavidin washing and elution buffers.

1x Binding & Washing Buffer	5 mM Tris-HCl pH 7.5 0.5 mM EDTA 1M NaCl Molecular Grade Water
2x Binding & Washing Buffer	10 mM Tris-HCl pH 7.5 1 mM EDTA 2M NaCl Molecular Grade Water
iBlock Buffer (100 ml solution)	2 g I-Block™ Protein-Based Blocking Reagent (ThermoFisher) 95 ml Molecular Grade Water 5 ml 10% (wt/vol) SDS

0.5 % SDS/2X SSC Buffer	20X SSC buffer 1 % SDS Molecular Grade Water
Genomic DNA mixture	500 ng (in 50 µl of water) of sheared genomic DNA (300-500 bp fragments) in 50 µl of 2x Binding & Washing Buffer

Library PCR amplification

The following reaction mixture was prepared in a PCR tube, using Illumina's Nextera DNA Library Prep kit:

PCR Component	Amount (µl)
DNA Library-Coated Beads	10
Nuclease Free Water	10
NPM Mix	15
PPC PCR Primer	5
Index Primer 1 (i7)	5
Index Primer 2 (i5)	5

The following program was then run:

Temperature	Time	Cycles
72 °C	03:00 min	
98 °C	30 sec	
98 °C	10 sec	11 – 13
63 °C	30 sec	
72 °C	40 sec	
72 °C	05:00 min	
4 °C	hold	

PCR Product Purification

AMPure XP beads for PCR purification (Beckman Coulter) were taken out of the 4 °C fridge and allowed to come to RT for 30 minutes before using. The 50 µl PCR product supernatant was transferred from the reaction tube to a new 1.5 ml tube using the magnetic rack. AMPure beads were vortexed to resuspend. Equal volume of AMPure beads was added to the reaction tube. The mix was pipetted well, around 10 times. The mixture was incubated at RT for 5 minutes using the tube revolver rotator. The tubes were then spun down briefly and placed on the magnetic rack to allow the beads to clear from the solution (~ 3-5 minutes). The supernatant was discarded and 200 µl of freshly prepared 80 % ethanol was added into the tubes to wash the beads. The ethanol was discarded, and the process was repeated once more. All ethanol was removed, and tubes were left open on the magnetic rack to air dry for ~ 5 minutes. DNA was eluted from the beads by washing with 20 µl of TE buffer. The PCR product was then quantified using the Qubit fluorometer (Qubit™ dsDNA HS assay) (ThermoFisher) and fragment size distribution was determined using Agilent's 2100 Bioanalyzer. The PCR reaction and purification was performed again with the remaining material. All purified PCR product (ChIA-PET library) was pooled together in a single tube, ready to be sent for sequencing.

All libraries underwent paired-end 150 bp (PE150) sequencing using Illumina's NovaSeq6000 platform (200-300M reads per library). Sequencing was outsourced to Novogene's facilities in Cambridge, UK.

2.2.1 Bridge Linker Preparation and Quality Control Steps

Bridge Linker Preparation

Bridge linker oligos for proximity ligation:

Bridge linker-F: 5'-/5Phos/CGCGATATC/**iBIOdT**/TATCTGACT -3'
Bridge linker-R: 5'-/5Phos/GTCAGATAAGATATCGCGT -3'.

Bridge linkers were HPLC purified (250 nmole) and purchased from IDT (Integrated DNA Technologies). The following procedure was then followed to prepare the bridge linkers for use during the proximity ligation reaction.

1x TNE buffer (Tris-NaCl-EDTA) was added to dissolve the top and bottom bridge linker oligos to a concentration of 100 µM. Oligos were vortexed for 10 seconds and then left at RT for 30 minutes to ensure complete resuspension. Five different ratios were prepared of top (100 µM) and bottom (100 µM) bridge linker oligos as detailed in **Table 4**.

Table 4 Ratio of top and bottom oligos for bridge linker preparation.

Ratio (vol/vol)	Top oligo (Forward) (µl)	Bottom oligo (Reverse) (µl)
A) 1:1	5	5
B) 1:1.5	5	7.5
C) 1.5:1	7.5	5
D) 1:2	5	10
E) 2:1	10	5

The following PCR program was then run:

Cycle number	Temperature and duration
1	95 °C, 2 mins
2-71	Decreasing 1 °C per cycle and holding for 1 min
72	25 °C, 5 mins
73	4 °C, 5 mins

Annealed bridge linkers were diluted to 200 ng/µl. Each single stranded oligo (200 ng) was run alongside 200 ng (10 µl) of annealed adapters from the previous step on a 4-20% (wt/vol) TBE gel. The gel was then immersed in gelRed dye for ~1hr and viewed on the Dark Reader Transilluminator (BioRad) (**Fig. 2**).

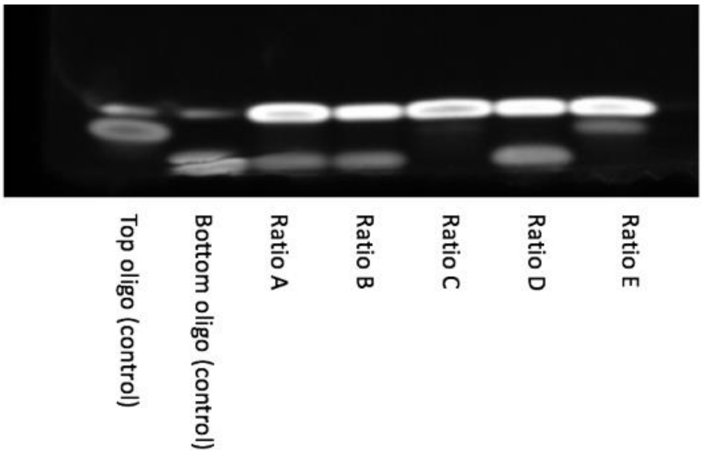


Fig. 2 Gel image after bridge linker ligation. Lanes 1 and 2 indicate the top and bottom oligos as controls, respectively. The product at the top of the two lanes is due to leakage from well 3 and can be ignored. Lane 3 corresponds to Ratio A (1:1), lane 4 to Ratio B (1:1.5), and lane 6 to Ratio D (1:2). In all three of these lanes, excess bottom oligo can be observed. Lane 7 corresponds to Ratio E (2:1) and excess top oligo can be observed. Ratio C (1.5:1) on lane 5 appears to be the most suitable ratio for mixing the top and bottom oligos as no excess unannealed DNA can be detected.

The optional ratio was then determined (**Ratio C** 1.5:1) based on the absence of no detectable unannealed top or bottom oligo on the lane. The rest of the top and bottom oligo stocks were mixed with the optional ratio and PCR was performed again, according to the PCR program shown above. The annealing linker was quantified with NanoDrop Spectrophotometer (Thermo Fisher) and then the bridge linker mixture was diluted to a final concentration of 200 ng/μl and aliquoted for storage at – 20 °C for future use.

Quality Control After Restriction Digestion with *AluI*

After overnight *AluI* digestion, 10 μl of the digested sample were pipetted into a new tube, along with 90 μl of Tris pH 8.0 and 2 μl proteinase K enzyme. The solution was mixed and incubated at 65 °C for 1 hour. The DNA was purified using Qiagen's PCR purification kit and fragment distribution was determined using Agilent’s 2100-HS Bioanalyzer. The profile after *AluI* enzyme digestion in HEK-293 cells is shown in **Fig. 3**.

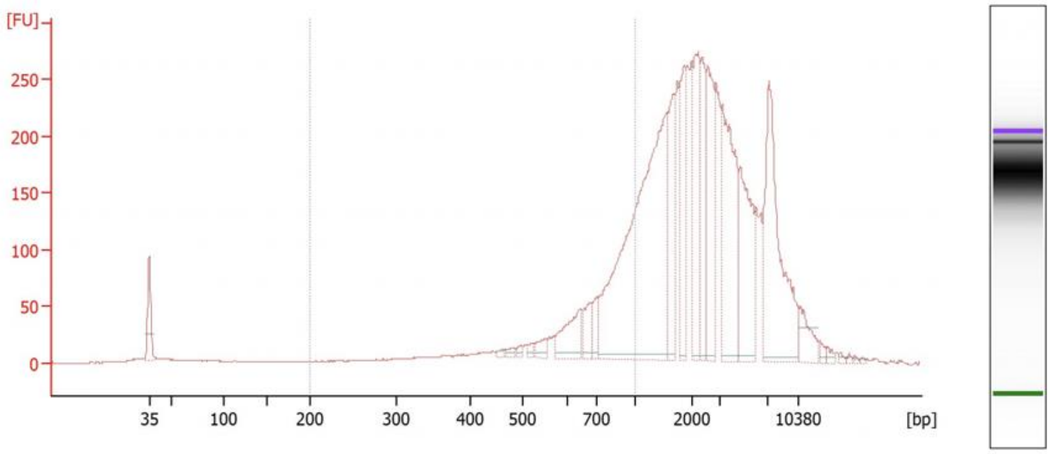


Fig. 3 Bioanalyzer profile after *AluI* digestion. The DNA fragment distribution after digestion appears to be between 450 bp – 10.4 kb, with fragments close to 2 kb being the average.

Quality Control After Proximity Ligation

The same procedure as with QC after restriction digestion was followed. Due to the addition of bridge linkers, ligated DNA is expected to shift slightly to the right in comparison to the digested chromatin. **Fig 4** shows the Bioanalyzer profile of HEK-293 after proximity ligation.

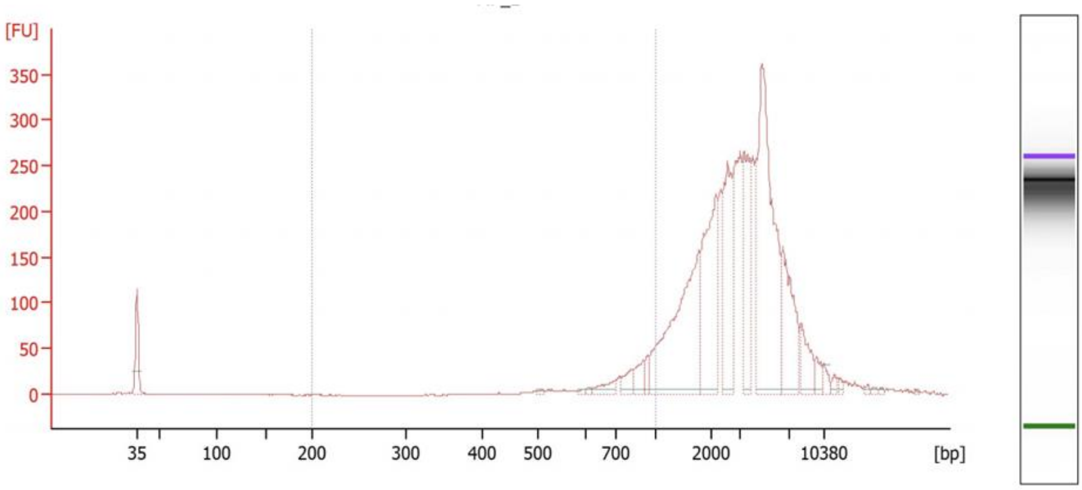


Fig. 4 Bioanalyzer profiles after bridge-linker ligation. The fragment distribution shows a small shift to the right in comparison to the *AluI* digested sample. This is an indication of

successful incorporation of the linker between fragments as the fragment distribution size has now increased from ~450 bp – 10.4 kb, to ~600 bp – 10.4 kb.

Quality Control After Tagmentation Reaction

After the initial tagmentation reaction and DNA purification using Zymo's Genomic DNA Clean & Concentrator kit, the profile is checked using Bioanalyzer. The acceptable profile is shown in **Fig. 5**. If the fragments are found to be too large or too short, the DNA and/or transposase enzyme concentrations are adjusted accordingly.

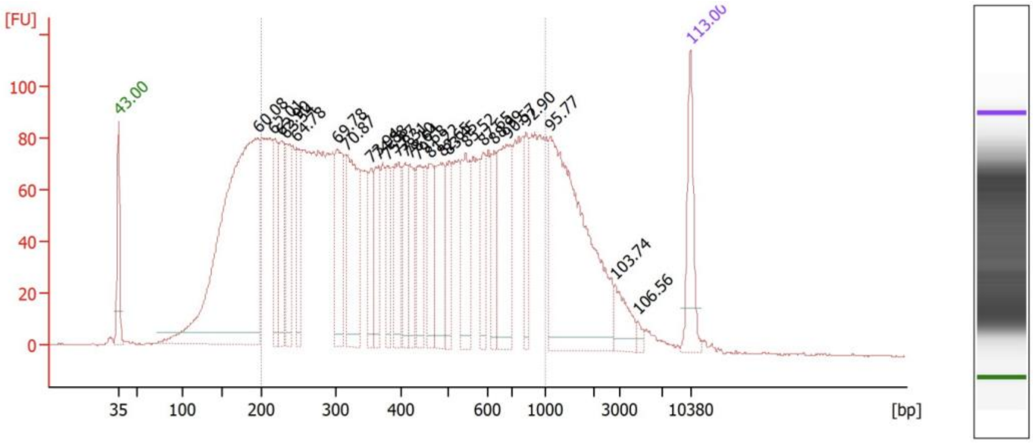


Fig. 5 Bioanalyzer profile after DNA tagmentation reaction. Tn5 transposase enzyme is used to uniformly tagment the DNA before immobilisation on streptavidin beads and PCR amplification.

2.3 siRNA Transfections

For siRNA transfection, cells (HEK-293 and HeLa) were cultured in 175 cm^2 flasks with growth media (DMEM,10 % FBS) at 37 ^{o}C, 5 % CO_2. At around 80 % confluency, cells were split via the following process: media was aspirated from the flask and the monolayer was rinsed with 10 ml of PBS. PBS was aspirated and 10 ml of trypsin/EDTA mix was added to the flask. After 3-5 minutes cells were checked under the microscope to ensure that they were detaching from the surface. 10 ml of growth media were added to the flask and cells were pipetted up and down to create a dispersed single cell suspension.

The suspension was transferred to a 50 ml Falcon tube and cells were counted before being centrifuged at 500 g for 5 minutes. The supernatant was aspirated, and cells were resuspended in growth media. For forward transfection, cells (~ 200,000 in 2.5 ml of DMEM/FBS) were seeded in 6-well plates 24 hours prior to siRNA transfection. For reverse transfection, cells were split, seeded, and transfected on the same day.

For HEK-293 cells, pre-designed siRNA duplexes were obtained from Merck. **Table 1** indicates the target sequences. Mission® siRNA Universal Negative Control (Merck) was used as control. For transfection, 35-55 nM of siRNA were mixed with 4 µl of Lipofectamine® RNAiMAX (Thermo Fisher) along with 494.5 µl Opti-MEM in a 1.5 ml Eppendorf tube. The mixture was vortexed and then incubated for 20 minutes at room temperature before being added on a 6-well plate. Plates were incubated for 24-48 hours before RNA extraction and RT-qPCR. For HeLa cells, the same process was applied, except the Pcf11 siRNA duplexes used were the ON-TARGETplus SMARTpool Pcf11 siRNA (Dharmacon) as indicated in **Table 1**. For control, the ON-TARGETplus Non-targeting Pool siRNA (Dharmacon) was used. Cells were incubated for 72 hours before lysis and protein blotting.

Table 1 Target sequences of siRNA duplexes for HEK-293 and HeLa cells.

Pcf11 siRNA Target sequences	**Direction: 5'- 3'**
HEK-293	**GUACCUUAUGGAUUCUAUU**
HeLA	**GAUACAAAUCAGCGACUUA**
	GUGUGCAAAUUUAACGAAA
	AAGUUAAGGAAGAACGAAU
	GGAUAAGACCGAUGGCAAA

2.4 RNA Extraction, cDNA Preparation, and RT-qPCR

HEK-293 cells were harvested, and RNA was extracted using Qiagen’s Rneasy Mini Kit according to the manufacturer’s instructions. RNA concentration was determined using NanoDrop™ 2000 (ThermoFisher). For cDNA preparation, 500 µg of RNA were added to PCR tubes for each sample along with 0.25 µl of random (hexamer) primers (Promega) and Rnase-free water to make up a total volume of 10 µl. Samples were annealed at 60 °C for 5 minutes. Then, for each sample the following reagents were added:

- 2 µl DTT (2M)
- 1 µl dNTPs (Promega)
- 2 µl Rnase-free water
- 4 µl M-MLV Reverse Transcriptase x5 Buffer (First strand) (Promega)
- 1 µl M-MLV Reverse Transcriptase (Promega)

Samples were incubated at 42 °C for 1 hour. Each cDNA sample was then diluted with 80 µl water. For qPCR reaction prep, the following protocol was carried out in qPCR tubes:

- 0.25 µl Pcf11 Primers (**Table 2**)
- 4.25 µl water
- 5 µl 2x SensIFAST™ SYBR MasterMix – No-Rox (Bioline)
- 0.5 µl Diluted cDNA Sample

Total volume = 10 µl

GAPDH controls and non-template controls were also prepared in the same way, and qPCR runs were completed using Qiagen’s Rotor-Gene Q with the following cycle conditions:

95 °C for 15 seconds
60 °C for 15 seconds
72 °C for 20 seconds

Samples were run for 40-45 cycles and values were normalised against GAPDH expression. All primer sets (Easy™ Oligos Sigma-Aldrich) were selected from PrimerBank and sequences were confirmed with BLAST (Madden, 2002). PrimerBank suggested three sets for Pcf11 and all three were used. **Table 2** lists the primer sequences for Pcf11 and GAPDH.

Table 2 Pcf11 primer sets used for RT-qPCR.

Primer Set	Forward (5'- 3')	Reverse (5'- 3')
Pcf11 1	GTTGGAAGAGAGTATCTCACTGC	GCTAGACGTATTCACATTGGGG
Pcf11 2	TGGTCAGTTCCCCTAGCATCT	GCCTTAGCTTGCTCTAGCTCAA
Pcf11 3	AGCTAGAGCAAGCTAAGGCAC	TGCACAGGAACCTGATGAGGA
GAPDH	GGAGCGAGATCCCTCCAAAAT	GGCTGTTGTCATACTTCTCATGG

2.4 Western Blotting

6-well plates were placed on ice and washed with ice-cold PBS. PBS was aspirated and ice-cold lysis buffer (130 µl of RIPA buffer (ThermoFisher) with 1.6 µl Halt™ Protease and Phosphatase inhibitor cocktail (ThermoFisher)) was added into each well. Adherent cells were scraped off the dish using a plastic cell scraper and were gently transferred into microcentrifuge tubes. Samples were in constant agitation for 30 minutes at 4 °C and then centrifuged for 15 minutes at 12000 rpm. The supernatant was aspirated and placed in a fresh tube, and the pellet was discarded. Protein concentration was quantified using Pierce™ BCA Protein Assay Kit (ThermoFisher) according to manufacturer's instructions. The volume for 30-40 mg of protein was calculated and an equal volume of 2x Laemmli sample buffer (Thermo Fisher) was added. Samples were reduced and denatured by boiling each lysate in sample buffer at 95 °C for 5 minutes. Samples were then loaded into the wells of a precast 4-20% SDS-PAGE gel (abcam), along with a molecular weight marker (abcam). The gel ran at 120 V for 90 - 120 minutes. To transfer proteins from the gel to a nitrocellulose membrane, semi-dry method was used at 15 V for 15 minutes.

2.4.1 Antibody staining

The membrane was blocked overnight at 4 °C in 5% milk/TBST buffer. The membrane was then incubated with primary polyclonal Pcf11 antibody* (abcam134391) at 1:1000 dilution and a control GAPDH antibody (Proteintech) at 1:1000 dilution overnight at 4 °C. Membrane was washed with TBST buffer 3 times, for 5 minutes each. The membrane was then incubated with an anti-rabbit secondary antibody to Pcf11 (abcam) at 1:10,000 dilution, and an anti-mouse secondary antibody to GAPDH (Santa Cruz) at 1:10,000, and was then washed 3 times with TBST, 5 minutes each. For signal development, Pierce™ ECL Western Blotting Substrate (Thermo Fisher) was used according to the manufacturer’s recommendations. The membrane was covered in transparent plastic wrap and the image was acquired using ImageQuant™LAS 4000 (GE Healthcare) for chemiluminescence. ImageJ software was used for digital visualisation of the membranes (Schindelin et al., 2015).

**

**Prior, other Pcf11 antibodies (Santa Cruz and Thermo Fisher) were also used but generated no detectable bands after several attempts.*

3.3 Data Analysis

All raw fastq ChIA-PET files were downloaded from Novogene’s servers after sequencing and stored on secure servers managed by The University of Warwick. For pre-processing, samples were sub-sampled accordingly to ensure that replicates had a similar number of reads between each other and their corresponding control. For sub-sampling, seqtk (https://github.com/lh3/seqtk) was used as follows:

```
seqtk sample -s read1.fastq <number of reads> > sub1.fastq
```

3.3.1 Initial Processing of ChIA-PET Samples

Fastq files were processed using the ChIA-PIPE pipeline (Lee et al., 2020) with the hg38/GRCh38 genome annotation as reference. ChIA-PIPE combines both independently available as well as custom-made python and perl scripts for the analysis of ChIA-PET data. In summary, it takes the two fastq files (read 1 and read 2) as input files and then the reads are scanned for the bridge linker sequence to determine the number of valid PETs (i.e., read pairs with linker and two PETs). Then, sequence alignment is performed using bwa (Li and Durbin, 2009), and files are sorted using Samtools (Li et al., 2009). The resulting BAM file contains uniquely mapped, non-redundant PETs. ChIA-PIPE then generates 2D chromatin interaction maps in standardised file formats compatible with publicly available high-resolution visualisation interfaces, such as Juicebox (Robinson et al., 2018), Epigenome WashU Browser (Li et al., 2019), Integrative Genomics Viewer (IGV) (Robinson et al., 2011), and HiGlass (Kerpedjiev et al., 2018). Next, loops are called and are annotated with peak support. Peak calling is performed either using SPP (Kharchenko et al., 2008) or MACS2 (Zhang et al., 2008), and with or without an input control, depending on the software and experiment. All peak annotations in the present work were called using MACS2 on broadpeak mode and input controls were used where available. Chromatin contact domains (CCDs) are called using the peak-supported loops. The pipeline then collates metrics from each step and produces a comprehensive table (.tsv) with QA (quality assessment) information. Once ChIA-PIPE was installed and all the software and dependencies were in place, a custom configuration shell script was written for each individual library and the ChIA-PIPE bash script was run for each sample.

Interaction correlation plots between sample replicates were produced with HiCExplorer (Ramírez et al., 2018; Wolff et al., 2018; Wolff et al., 2020) for command line with the function hicCorrelate and default options.

3.3.2 RNA-seq and ChIP-seq Analysis

Pcf11 ChIP-seq and siPcf11 3’ mRNA data were downloaded from GEO (Accession: GSE123105). HeLa and U2OS bulk-RNAseq datasets were downloaded from the ENA Browser (HeLa) and GEO (U2OS) with Accession numbers PRJNA245463 and GSM4943751, respectively. Pcf11 ChIP-seq libraries were aligned to the human genome (hg38/GRCh38) using bwa (Li and Durbin, 2009). Files were sorted and indexed using Samtools (Li et al., 2009). MACS2 was used to call for peaks on BroadPeak mode, with all other options kept as default. RNA-seq and 3’ mRNA-seq datasets were aligned, sorted, and indexed in the same way. LiBiNorm was used in htseq compatible mode to produce counts files. For bulk RNA-seq, option -j was used to produce TPM values. For differential gene expression analysis of siLuc and siPcf11 3’ mRNA-seq samples, the R package *DESeq2* (v.3.14) (Love et al., 2014) was used.

3.3.3 ChIP Peak Analysis and Metagene Plots

The R package *ChIPpeakAnno* v3.28.0 (Zhu et al., 2010) was used for calling and overlapping matching peaks between samples and replicates using the built-in function findOverlapsOfPeaks and using the MACS2 bed file outputs as inputs.

For the metagene (coverage) plots, bigwig files were generated from the bam files and normalised against their corresponding control using BEDtools v.2.25.0 (Quinlan and Hall, 2010) via the following command:

```
bamCompare -b1 TREATMENT.bam -b2 CONTROL.bam -o NORMALISED.bw -bl
ENCFF356LFX.bed -p 2
```

The option -bl was used to exclude blacklisted regions. These regions correspond to the hg38/GRCh38 assembly and can be accessed on https://www.encodeproject.org/files/ENCFF356LFX/. Blacklisted regions are considered to be artefact regions (i.e., abnormally high signal, abnormal shapes, or read coverage).

The normalised bigwig file was then used to generate a compressed matrix of the coverage regions with the following deepTools (Ramírez et al., 2016) command:

```
computeMatrix scale-regions -S NORMALISED.bw -R hg38_RefSeq.bed --
beforeRegionStartLenght 3000 --regionBodyLength 5000 --afterRegionStartLength 3000 -o
MATRIX.mat.gz
```

In scale-regions mode, all regions in the BED file are plotted. The profile is then plotted using the command:

```
plotProfile -m MATRIX.mat.gz -out MATRIX.png
```

3.3.4 Gene Loop identification

A custom written Python3 script was used for the identification of looped genes. The peak-supported interaction files were taken as input alongside the hg38/GRCh38 gtf files corresponding to protein-coding (pc) genes (n = 19,941). The interaction coordinates of each interaction file were scanned against the start and end sites of pc genes allowing for a span of +/- 500 bp on either side. Looped genes were isolated and a score for each looped gene was assigned based on how close the interaction anchors were to the real TSS/TES, along with the number of PETs mapped to that region.

3.3.5 GO Analysis

GO Analysis was performed using BiNGO on Cytoscape (v3.9.1) with the following settings: hypergeometric test was selected as statistical test and Benjamin & Hochberg False Discovery Rate (FDR) was applied as a method for multiple testing correction. Significance level was set at < 0.05 and the categories visualised corresponded to overrepresented categories of GO Biological Processes after correction. For organism/annotation, Homo sapiens was selected.

3.3.6 Differential Looping

The R package *diffloop* v3.14 (Lareau and Aryee, 2021) was used for identification of differential loops between the siPcf11 and N.S siRNA ChIA-PET samples. BED files corresponding to the coordinates of each sample's raw interactions were loaded into R with the function loopsMake to create a loops object. The function mangoCorrection with option FDR = 0.01 was used on the loops object to filter loops that may be biased due to proximity or low PET counts. To further eliminate bias as per the library's authors' notes, we filtered out loops that were overrepresented in one replicate (>5 PETs), but absent in the other (=0 PETs). Differential looping was performed with the quickAssocVoom function. Loops with $padj \leq 0.05$ were considered significantly differentially looped between conditions.

Chapter 3: Transient Knock-down of Pcf11

3.1 Introduction

3.1.1 Pcf11 and Transcription

Human Pcf11 is one of several factors implicated in the machinery of pre-mRNA 3’ end processing. Pcf11 has been shown to be directly involved in the process of transcriptional termination in model organisms, such as *Drosophila* (Zhang and Gilmour, 2006), as well as human cells (Kamieniarz-Gdula et al., 2019). In both studies, Pcf11 depletion via RNA interference (RNAi) appears to cause Pol II to continue transcribing beyond normal termination regions, highlighting the dependence of effective transcript termination on the function of Pcf11. Depletion of Pcf11 has also been found to decrease transcription initiation of endogenous genes, suggesting that successful formation of the 3’ end also stimulates the initiation of transcription (Mapendano et al., 2010). This could suggest that factors found at the TES are recycled back to the TSS as a way of maintaining ongoing transcription. Short-range chromatin interactions, or gene loops, could be a possible way through which physical contacts between the 5’ and 3’ ends occur (Ansari and Hampsey, 2005, O'Sullivan et al., 2004), thus facilitating the recycling of Pol II and other factors.

Pcf11 ChIP-seq data in HeLa cells (Kamieniarz-Gdula et al., 2019) have confirmed the localisation of Pcf11 at the TSS, further adding onto the evidence that there is crosstalk between the TES and TSS. ChIA-PET technology provides a way of capturing bona fide chromatin interactions mediated by a factor of interest, something that would enable us to have a close look at gene loop formation and confirm whether Pcf11 localisation at the promoter region is a result of looping.

Most ChIA-PET experiments have been performed on factors such as Pol II and CTCF, and without input controls for the ChIP reactions (Lee et al., 2020). For the purposes of this work, I determined that Pcf11 knock-down samples would be useful controls. Although we do not expect knock-down of Pcf11 to disturb overall 3D conformation, these samples would be used as input controls during the peak-annotation stage. The following experiments also reflect the lengthy process to identify a working Pcf11 antibody for use during ChIP reactions.

3.1.2 Objectives

As mentioned previously, performing Pcf11 ChIA-PET on human cells would help us answer the question of whether localisation of Pcf11 at the promoter is a result of gene looping. This would in turn suggest that Pcf11 is implicated in other aspects of the transcriptional cycle besides termination, and that perhaps other factors function in a similar manner. These results could also provide interesting insights in terms of Pcf11's role and function in the context of chromatin interactions in general. For these purposes of generating a transient Pcf11 knock-down, we utilised siRNA technology and performed a number of optimisation experiments. In the following pages, I present the results from RT-qPCR and Western Blot experiments carried out to establish Pcf11 knock-down dynamics on the mRNA and protein level in human cell lines.

3.2 Results

3.2.1 Pcf11 mRNA Reduction

One of the parameters I wanted to establish was the optimal siRNA concentration to achieve maximum Pcf11 knock-down without impacting cell viability. **Fig. 1** shows the results after a 48-hour incubation post-transfection for HEK-293 cells transfected with 35 nM and 55 nM of siRNA.

A significant reduction in Pcf11 mRNA is observed at both concentrations. There is no significant difference between samples transfected with 35 nM or 55 nM, indicating that an siRNA concentration between 35-55 nM is effective at reducing Pcf11 mRNA levels by 70-90%. These results confirm the effectiveness of the siRNA sequences used.

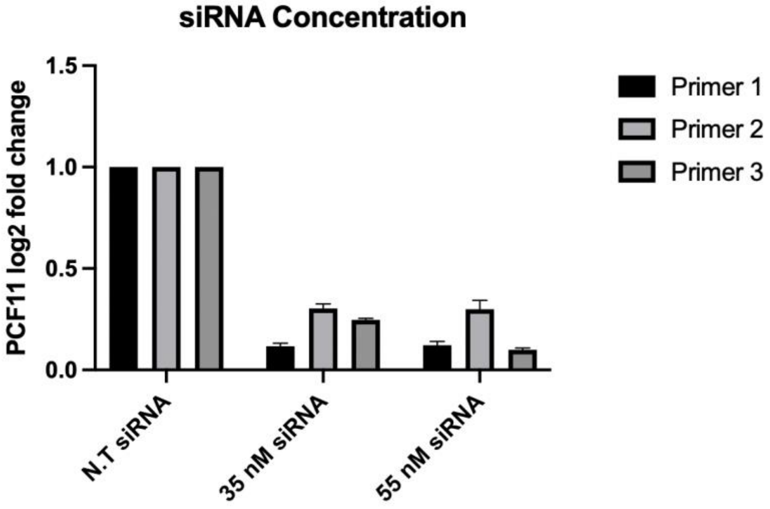


Fig. 1 Log2 fold change of Pcf11 mRNA in HEK-293 cells at two different siRNA concentrations. Successful reduction of mRNA is shown for samples transfected with 35 nM and 55 nM of Pcf11 siRNA (n=2) after a 48-hour incubation period. Values were normalised first against a corresponding GAPDH control, and then against the values for non-targeting (N.T) siRNA. Error bars correspond to SD between the two replicates.

Another parameter that I wanted to examine the effect of was that of incubation time. It is usually recommended to incubate cells between 24-72 hours before assessing knock-down efficiency (Haiyong, 2019), although the level and duration of the knock-down also depend on factors such as cell type, cell health, confluency, and siRNA concentration. **Fig. 2A and 2B** show the results 24 and 48 hours post transfection, respectively.

Successful knock-down efficiency can be observed for both conditions, with an overall ~ 84% knock-down for samples incubated for 24 hours, and ~ 65% for samples incubated for 48 hours. Although knock-down efficiency appears to be less successful after 48 hours, this does not necessarily suggest new transcript generation since the same result cannot be observed in a separate 48-hour incubation experiment as previously shown in **Fig. 1** above. The same parameters were used for both 24- and 48-hour experiments, however, inconsistency in cell confluency might have impacted siRNA uptake in this instance.

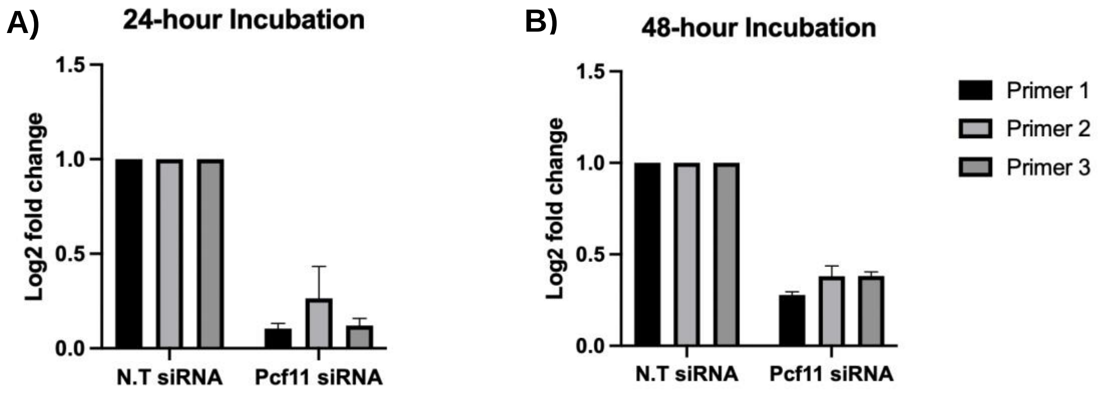


Fig. 2 Log2 fold change of Pcf11 mRNA levels in HEK-293 cells at 24- and 48-hours post-transfection. A) Relative reduction of Pcf11 mRNA in samples transfected with 55 nM of siRNA after a 24-hour incubation (n=2) **B)** Relative reduction of Pcf11 mRNA in samples transfected with 55 nM of siRNA after a 48-hour incubation (n=2). Values were normalised against GAPDH expression and then against N.T siRNA. Error bars correspond to SD between the two replicates.

I also wanted to explore the effects of forward and reverse transfection on knock-down efficiency. **Fig. 3A and B** show the results for forward and reverse transfection, respectively. Pcf11 mRNA reduction can be observed for both transfection methods, with an overall ~ 65% knock-down in forward transfection and ~ 80% in reverse transfection. Uptake of nucleic acids seems to be more efficient in reverse transfection, however, this could also be attributed to cell confluency discrepancies since the experiments had to be carried out on different days. Overall, reverse transcription was preferred since it saves a day of work and shows high knock-down efficiency.

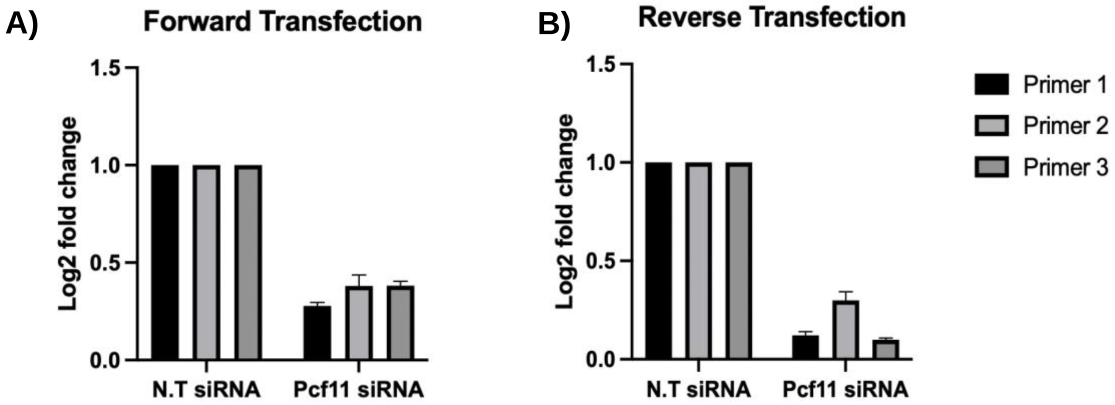


Fig. 3 Log2 fold change of Pcf11 mRNA levels in HEK-293 cells after forward and reverse transfection. A) Relative reduction of Pcf11 mRNA in samples after 48 hours post forward transfection with 55 nM of siRNA (n=2). **B)** Relative reduction of Pcf11 mRNA in samples after 48 hours post reverse transfection with 55 nM of siRNA (n=2). Error bars correspond to SD between the two replicates.

Given these results and parameters tested, we determined that reverse transfecting HEK-293 cells with 35-55 nM of siRNA followed by a 24–48-hour incubation time, results in Pcf11 mRNA reduction of up to 90%.

3.2.2 Pcf11 Protein Knock-down

Although we were able to establish knock-down of Pcf11 through RT-qPCR experiments, this also had to be confirmed on the protein level. Early attempts to visualise the effects of the knock-down on a protein blot proved unsuccessful due to ineffective antibodies. **Fig. 4** shows one of many Western blot experiments to detect Pcf11. Presence of GAPDH is clear in both control and Pcf11 siRNA samples, however, Pcf11 protein is not detectable. After several repeats using higher antibody ratios and more lysate, a different Pcf11 antibody was used, however, it yielded similarly unsuccessful results.

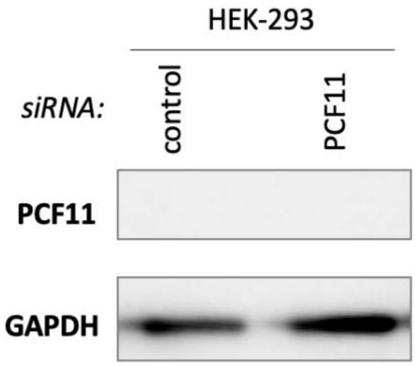


Fig. 4 Western blot analysis of Pcf11. The blot was probed with antibodies against Pcf11 (sc-515669) and GAPDH for HEK-293 cell samples transfected with non-targeting control siRNA and Pcf11 siRNA. The upper panel shows the region where Pcf11 bands should have appeared, and the lower panel shows the presence of GAPDH as control.

Given that attempts to confirm the knock-down on the protein level and establish an effective Pcf11 antibody for immunoprecipitation purposes were continuously unsuccessful, this work was put on-hold to work on other aspects of the project, such as establishing and optimising the ChIA-PET protocol. In light of new published data on Pcf11 ChIP-seq (Kamieniarz-Gdula et al., 2019), the knock-down work was later revisited and revised to replicate the methodology of Kamieniarz-Gdula et al., 2019. Therefore, although early RT-qPCR results confirmed the validity of the siRNA duplexes used, for the following work I switched to pooled Pcf11 siRNAs made up of four different duplexes as described in the Methods section. I also used HeLa cells and the Pcf11 antibody that was used in the relevant publication, as I wanted my findings to be comparable.

Pcf11 protein is detectable and a ~ 90% knock-down of the protein can be observed in the sample transfected with Pcf11 siRNA in HeLa cells (**Fig. 5)**.

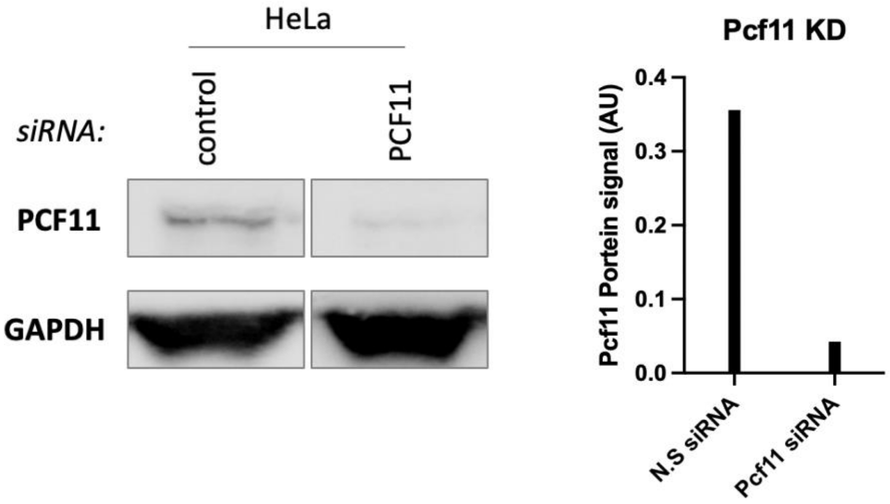


Fig. 5 Western blot analysis of Pcf11. The blot (left) was probed with antibodies against Pcf11 (ab134391, Abcam) and GAPDH for HeLa cell samples transfected with either non-targeting control siRNA or 50 nM Pcf11 siRNA. The upper panel shows Pcf11 bands, and the lower panel shows the presence of GAPDH as control of expression. The bar chart (right) quantifies the expression signal of Pcf11 protein in control and Pcf11 siRNA samples. Values were normalised by each sample's corresponding GAPDH protein expression. The Pcf11 siRNA sample shows ~ 90% reduced expression compared to the control.

Given that the Pcf11 antibody and knock-down were validated through protein blotting, I resumed with the initial research plan of doing Pcf11 ChIA-PET on control and knock-down samples.

3.3. Discussion

There are a number of technologies that can be used to induce gene silencing, both *in vitro* and *in vivo*. Such methods include RNAi (siRNA or shRNA), phosphorodiamidate morpholino oligos (PMOs), External Guide Sequences (EGSs), and CRISPR interference (CRISPRi). Knock-down by RNAi is a widely used method for transient gene silencing in mammalian cells. The use of RNAi technology to knock-down Pcf11 in human cell lines has been successfully reported in the past (West and Proudfoot, 2008, Mapendano et al., 2010), therefore, this method was preferred. The limitation that comes with using siRNA for gene silencing mainly concerns the risk of off-target effects. However, as we discover more about the causes of these effects, we are able to tackle them more effectively. The duplexes purchased were designed and modified in a way to minimise off-target activity by editing the antisense strand seed region to destabilise off-target activity.

HEK-293 cells were initially chosen for this work, however, after Pcf11 ChIP-seq data in HeLa cells were published (Kamieniarz-Gdula et al., 2019), HeLa cells were chosen for purposes of ChIA-PET library prep instead. Early attempts to validate Pcf11 protein degradation on HEK-293 cells were unsuccessful due to lack of a working antibody, therefore research plans had to be amended. For this reason, a “back-up” dataset was generated which will be further discussed in the following Chapters. This dataset refers to Flavopiridol (FP)-treated and untreated HEK-293 Pcf11 ChIA-PET libraries, generated directly after we were able to observe Pcf11 bands on a protein blot, but prior to revisiting the Pcf11 knock-down work that would have allowed us to proceed with the initial plan of performing ChIA-PET on untreated vs knock-down samples. FP is a potent inhibitor of transcription that works by blocking the function of positive elongation factor b (P-TEFb; cyclin-dependent kinase 9 CDK9/cyclin T) (Jonkers et al., 2014). P-TEFb phosphorylates the CTD of Pol II on Ser2 (Baumli et al., 2008), which is the phosphorylated amino acid position that preferentially interacts with Pcf11’s CID. Although not ideal to investigate Pcf11’s direct role in gene looping, this data would still be able to provide insights into how Pcf11’s looping profile and localisation are affected under conditions where Pcf11-Pol II interactions are minimised. However, as Flavopiridol is a very potent inhibitor and p-TEFb interacts with other factors, such as DRB Sensitivity Inducing Factor (DSIF) and Negative Elongation Factor (NELF) (Baumli et al., 2008), interpretations about Pcf11’s direct role in looping formation and 3’ - 5’ end crosstalk would be inconclusive. This highlighted the importance of revising the research plan and confirming Pcf11 protein depletion to generate a more suitable dataset.

Ultimately, the best way to confirm successful silencing of genes is through Western blot. However, this was not immediately possible due to lack of protein blotting equipment that only became available later. For this reason, several RT-qPCR experiments were carried out (**Fig. 2-4**) to monitor knock-down efficiency on the mRNA level and optimise certain parameters such as concentration, incubation time, and transfection method (forward or reverse). Monitoring gene silencing through RT-qPCR can also provide useful insights. For example, when mRNA reduction is observed without a corresponding decrease in protein levels, this could indicate that protein half-time is too long. On the other hand, if protein reduction is observed without decrease in mRNA levels, this could indicate that the siRNA's effects occur post-translation. Although the siRNA duplex and primers used to test knock-down efficiency on HEK-293 cells were effective, Pcf11 protein reduction could not be validated through Western blot. Earlier published data on Pcf11 Western blots on human cell lines (West and Proudfoot 2008 and Mapendano et al., 2010) used a Pcf11 antibody that was sourced from David Gilmour's lab, potentially indicating that an effective antibody able to detect human Pcf11 was lacking from commercial vendors. However, after the publication of HeLa Pcf11 ChIP-seq data (Kamieniarz-Gdula et al., 2019) using a commercially available Pcf11 antibody (abcam) I was able to revisit and complete this aspect of the project. Since I wanted the ChIA-PET data to be directly comparable to that of published ChIP-seq data, I used the same cell line, siRNA, and antibodies used in the published data. Hence, after several optimisation experiments and through testing a number of antibodies, I was able to confirm ~ 90% depletion of Pcf11 protein in HeLa cells ~72 hours post siRNA transfection as was shown in **Fig. 5**. Once the knock-down was validated, ChIA-PET library prep followed. The results from these experiments are explored in the following Chapters, where I first thoroughly assess the quality of the libraries before performing more specific analyses to identify gene looping.

Chapter 4: Assessment of ChIA-PET Data Quality and Preliminary Analysis

4.1 Introduction

4.1.1 ChIA-PET - A Multi-Dimensional Dataset

The diameter of the nucleus of a human cell is around five orders of magnitude smaller than the length of the DNA contained within it (Pal et al., 2019). Therefore, this poses both a structural and functional challenge as the genome needs to densely compact itself, while simultaneously maintaining its function and remaining accessible to regulatory elements. The way this is achieved is through the DNA associating itself with proteins that serve functional and structural roles (e.g., histones). This complex of DNA and proteins makes up chromatin and its organisation within the nucleus is closely regulated (Cavalli and Mistelli, 2013).

Chromatin Interaction Analysis by Paired-End Tag (ChIA-PET) sequencing is a complex experimental technique that involves several steps. As the technique combines chromosome conformation capture (3C) technology along with proximity ligation and chromatin immunoprecipitation, a ChIA-PET library carries three sets of genomic information. These sets of information include (i) whole-genome chromatin interactions (i.e Hi-C), (ii) chromatin interactions between genomic regions bound by a specific protein of interest, and (iii) the binding profile of the protein of interest (i.e., ChIP-seq). Given that ChIA-PET captures specific sites of interaction in a genome-wide fashion, it is able to achieve 1 kb resolution at a much lower sequencing depth than Hi-C, which normally requires at least a billion reads to achieve the same level of resolution in higher eukaryotes (Rao et al., 2014, Lieberman-aiden et al., 2009). However, due to the multilevel nature of the dataset and the number of experimental reactions that take place, background noise can be captured during different stages throughout the protocol.

As it was detailed in *Chapter 2: Materials and Methods*, fragments of DNA that are in close proximity are *in-situ* ligated using a biotinylated bridge linker. However, as the cell nucleus would have undergone a number of disruptive reactions already (e.g., cell and nuclear lysis), there exists the probability that digested chromatin fragments might have moved from their original position and false interactions could be captured during this stage. Literature suggests that the ratio of intra to inter-chromosomal interactions could be an indicator for noisy data (Wang et al., 2021), and that inter-chromosomal interactions would likely be the result of non-specific, random ligation events. This excludes the possibility of legitimate translocations often observed in cancer cells. According to the experience of the protocol's original authors, a quality dataset would be expected to have a majority of intra-chromosomal paired-end tags (PETs) (Wang et al., 2021). Non-specific binding can also be captured during immunoprecipitation (IP), as well as during pull-down of the biotinylated linker. Hence, a fraction of interactions that are not specific to the factor of interest, as well as fragments of DNA that do not contain the bridge linker, can still be captured and sequenced. Such non-specific interactions can be identified and filtered out during downstream steps of the computational analysis.

Initial processing and additional downstream analyses can help us examine the quality of the data and hence determine data reliability. ChIA-PIPE is the latest available tool for the use of initial and downstream processing of ChIA-PET libraries (Lee et al., 2020). ChIA-PIPE provides an automated pipeline that is able to sort read pairs based on whether or not the linker sequence is found between two tags, align them to a reference genome, quantify PETs, and determine interaction frequencies (i.e., looping between loci). ChIA-PIPE also incorporates already established tools, such as MACS2 (Zhang et al., 2008) to perform peak-calling for binding site identification. To then identify interactions that are specific to the protein of interest, peaks are overlapped with chromatin interaction anchors. A number of quality assessment metrics can be generated and used to identify potential weaknesses of the dataset, something that will be further explored in this Chapter.

4.1.2 Visualising and Annotating Hi-C and ChIA-PET data

In order to understand the quality of the dataset in terms of resolution and to also extract biologically relevant information, the combined integration and analysis of different types of annotations is required. Sophisticated tools for visualisation are essential to identifying relevant patterns in the data regarding genomic coordinates. The visualisation of data produced from high-throughput experiments, such as Hi-C and ChIA-PET, is even more complex given that each of the data points is representative of the interaction between two distal sites. There are different tools and software that can be applied for the generation of high-quality plots showing ChIA-PET contact maps, along with annotations for chromatin contact domains and loops. It is important to note here that since a ChIA-PET matrix is sparser than a Hi-C matrix as it does not contain the entire interactome of the genome and is not sequenced as deep, the identification of TADs using common Hi-C algorithms is not ideal. Instead, briefly, the ChIA-PIPE pipeline computes chromatin contact domains (CCDs) by sorting the peak-supported loops by PET count, and then only using the top ⅓ (67th percentile) as a cutoff for calling CCDs. The span of the CCD is computed using the midpoint of the corresponding loop's anchors. The output CCDs file, along with the loops and other annotation files, can be viewed on interactive interfaces allowing for genome browsing and loading additional tracks. Juicebox is a visualisation system that allows for maximum zoom of contact maps generated from Hi-C experiments and is widely used by researchers in the field of 3D genome architecture (Durand et al., 2016). The visualisation of ChIA-PET matrices is not only useful for identifying patterns and domains, but also determining data resolution and therefore quality.

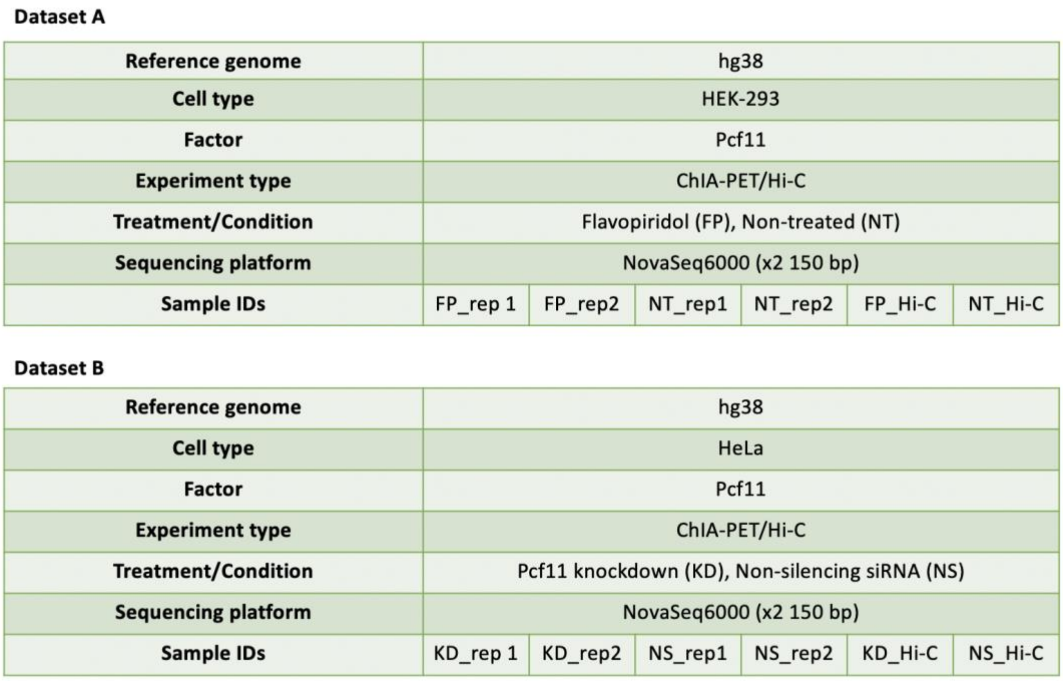
Given the considerations explored in the previous paragraphs, it is important to establish data quality, as well as determine whether replicates correlate with one another. It is also useful to identify factors that could be compromising the data early-on, before drawing any biological conclusions from downstream analyses. The *in-situ* ChIA-PET protocol (Wang et al., 2021), which is an improved version of the long-read ChIA-PET protocol (Li et al., 2017), was only very recently made publicly available. All the data presented in this work were generated via *in-situ* ChIA-PET using the version of the protocol that was kindly sent to me by the Ruan Lab prior to its official publication. Several months were dedicated to pilot and optimisation experiments in order to ensure library quality within the given timeframe. In the following pages I present results from quality assessment and preliminary analyses performed mainly on two of the datasets of my work: Pcf11 ChIA-PET on HEK-293 and HeLa cells. This work also includes Pol II ChIA-PET libraries from U2OS cells that will also be explored in this chapter, however, I will mainly utilise these data in Chapter 5.

4.2 Results

4.2.1 Metrics

The first Pcf11 ChIA-PET dataset that was generated is made up of six samples - two replicates of untreated HEK-293 cells, two replicates of Flavopiridol (FP)-treated HEK-293 cells, and a corresponding “input” control for each of the two conditions, equivalent to a Hi-C library. I will refer to this dataset as **Dataset A**. The second Pcf11 ChIA-PET dataset that was generated is also made up of six samples - two replicates of untreated HeLa cells, two replicates of Pcf11 siRNA HeLa cells (knock-down), and a corresponding Hi-C sample for each condition. I will refer to this dataset as **Dataset B**. Relevant details of each of the two experiments are listed in **Table 1**.

Table 1 Experimental information of Datasets A and B.



Dataset A

Reference genome	hg38					
Cell type	HEK-293					
Factor	Pcf11					
Experiment type	ChIA-PET/Hi-C					
Treatment/Condition	Flavopiridol (FP), Non-treated (NT)					
Sequencing platform	NovaSeq6000 (x2 150 bp)					
Sample IDs	FP_rep 1	FP_rep2	NT_rep1	NT_rep2	FP_Hi-C	NT_Hi-C

Dataset B

Reference genome	hg38					
Cell type	HeLa					
Factor	Pcf11					
Experiment type	ChIA-PET/Hi-C					
Treatment/Condition	Pcf11 knockdown (KD), Non-silencing siRNA (NS)					
Sequencing platform	NovaSeq6000 (x2 150 bp)					
Sample IDs	KD_rep 1	KD_rep2	NS_rep1	NS_rep2	KD_Hi-C	NS_Hi-C

I first wanted to determine the distribution of important mapping metrics that would mostly provide information on (i) whether the bridge linker ligation step of the protocol was successful, and (ii) what proportion of PETs are redundant. To establish data quality, as a benchmark for comparison, I used the published quality assessment metrics from CTCF ChIA-PET libraries from HFFc6 cells (Lee et al., 2020). General guidelines from Wang et al., (2021) were also taken into consideration to determine dataset quality. **Fig.1A-C** shows mapping metrics for datasets A and B, along with the equivalent metrics of the benchmark dataset.

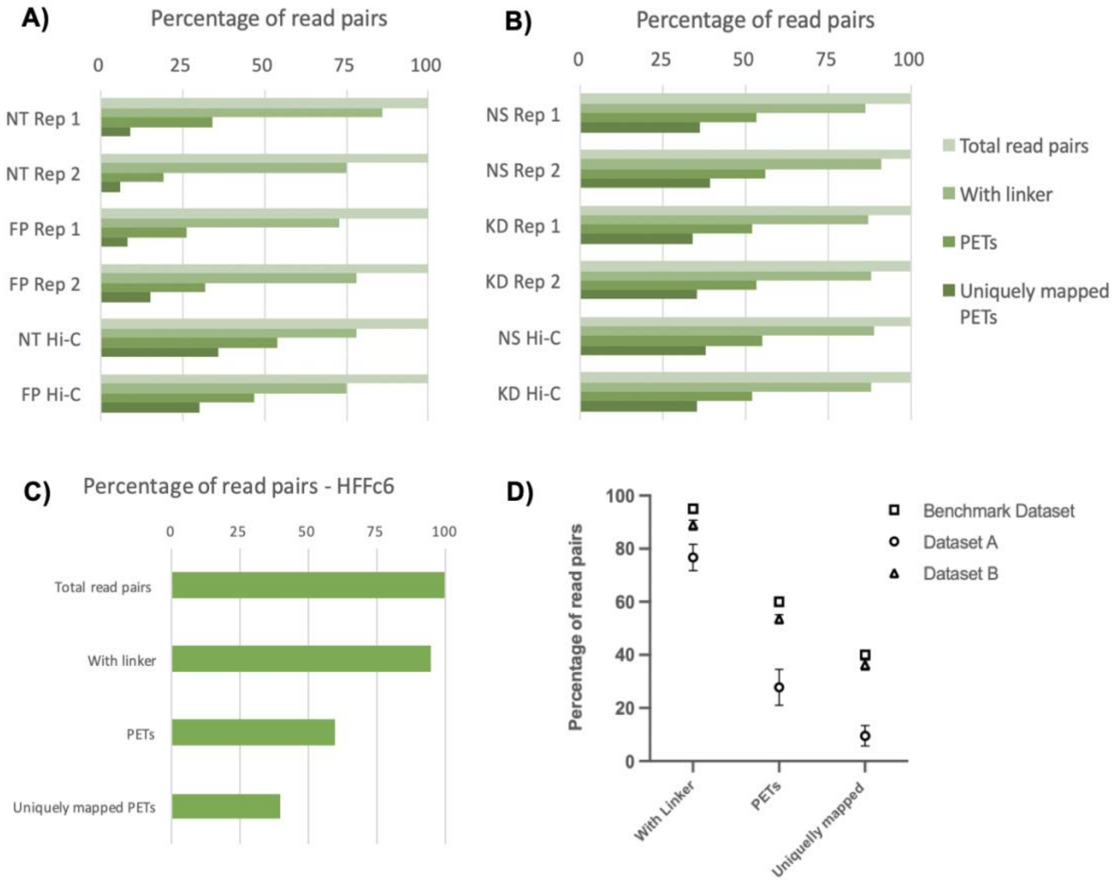


Fig. 1 Mapping metrics of Datasets A and B. Total read pairs refer to the total number of reads per library (200-300M), represented as a percentage (100%). **A)** Percentage of read pairs for Dataset A. Bar plot showing the percentage of reads that contain the bridge linker, along with the percentage of all and uniquely mapped PETs. NT refers to non-treated samples and FP refers to Flavopiridol-treated samples. **B)** Equivalent to A, but for Dataset B. NS corresponds to non-silencing siRNA control samples, and KD corresponds to Pcf11 knock-down samples. **C)** Benchmark dataset metrics as published in Lee et al., 2020. The data corresponds to CTCF ChIA-PET in HFFc6 cells, available on the 4DN portal (**Accession number:** 4DNESCQ7ZD21). **D)** Comparison of benchmark dataset and Datasets A and B.

The metrics of the benchmark dataset (**Fig. 1C**), show that out of all sequenced read pairs (345M), 95% contain the linker. The fraction of linker reads with PETs is 60%, and uniquely mapped PETs make up 40% of the total reads. Just as was mentioned in the previous Chapter, Dataset A was produced as a "back-up" dataset after attempts to validate Pcf11 knock-down on the protein level were repeatedly unsuccessful due to ineffective antibodies. As a result, Dataset A (**Fig. 1A**) was generated immediately after pilot experiments had confirmed that the ChIA-PET protocol was established enough to reach the library production step, but prior to the completion of further optimisation to maximise data quality. Across all of Dataset A samples, ~88% of total reads contain the linker. On average, only ~26% of NT and FP replicates are linker reads with PETs, and only ~10% of the total reads are uniquely mapped PETs. In contrast, 51% of the reads of the Hi-C libraries of the same dataset are PETs, and 33% are uniquely mapped. This discrepancy is a result of low starting material for the ChIA-PET libraries (low DNA quantity), which in turn resulted in PCR overamplification before sequencing, an issue that will be further discussed in the Discussion section of this chapter. Dataset B was generated several months later, after both further optimisation and Pcf11 knock-down were confirmed. As shown in **Fig. 1B**, on average, ~ 90% of reads contain the linker, ~ 55% are PETs, and ~ 37% are uniquely mapped PETs. These metrics closely resemble those of the benchmark dataset's **(Fig. 1D**), confirming that Dataset B is of adequate quality and that further optimisation to the experimental protocol had indeed improved library quality.

Another important factor to consider when determining ChIA-PET data quality is the ratio of intra (*cis*) to inter (*trans*) chromosomal interactions. ChIA-PET datasets are expected to have a higher (>1) ratio of intra to inter chromosomal interactions, something that would indicate higher quality interaction data (Wang et al., 2021). **Fig. 2A-C** shows the average distribution of intra/inter chromosomal interactions for Datasets A and B, along with those of the U2OS Pol II datasets. On average across all HEK-293 ChIA-PET samples, the ratio of intrachromosomal to interchromosomal interactions is 1:1, meaning that there are equal numbers of interchromosomal to intrachromosomal interactions (**Fig. 2A top**). Although this suggests that there is a level of noise in the data, it does not necessarily mean that there is no meaningful information to be extracted downstream. HeLa ChIA-PET libraries (**Fig. 2A bottom**), on average, have a ratio of 1.22:1 (intra/inter). This is in line with the suggested ratio of >1:1. Relevant steps of the protocol (nuclear lysis conditions) were further optimised prior to the generation of the HeLa ChIA-PET libraries, resulting in a more favourable amount of meaningful chromatin contacts, although not as close as that of the benchmark dataset (ratio 2.2:1). This ‘noise’, however, appears to be more biological rather than technical. This becomes apparent from the Pol II datasets that were generated following to the same protocol. The lower ratio of total *cis* interactions captured in the Pcf11 datasets appears to be due to less total Pcf11 in the cell lines themselves (**Fig. 2D**), resulting in less pull-down of Pcf11-immunoprecipitated DNA fragments, and thus affecting the overall ratio of *cis* to *trans* interactions. This possibility is further confirmed from the Hi-C datasets (i.e., non-IP’d libraries) of the Pcf11 datasets which had a ratio of >2:1 intra: inter. **Fig. 2D** shows the gene expression of Pcf11 and Pol II in human cell lines. In HEK-293 cells, Pol II appears to be almost 8 times more expressed than Pcf11, and a 3-fold expression ratio of Pol II to Pcf11 can be observed for HeLa cells.

For the current research, interchromosomal interactions were filtered out of downstream analysis and only *cis* interactions were used to address further research questions.

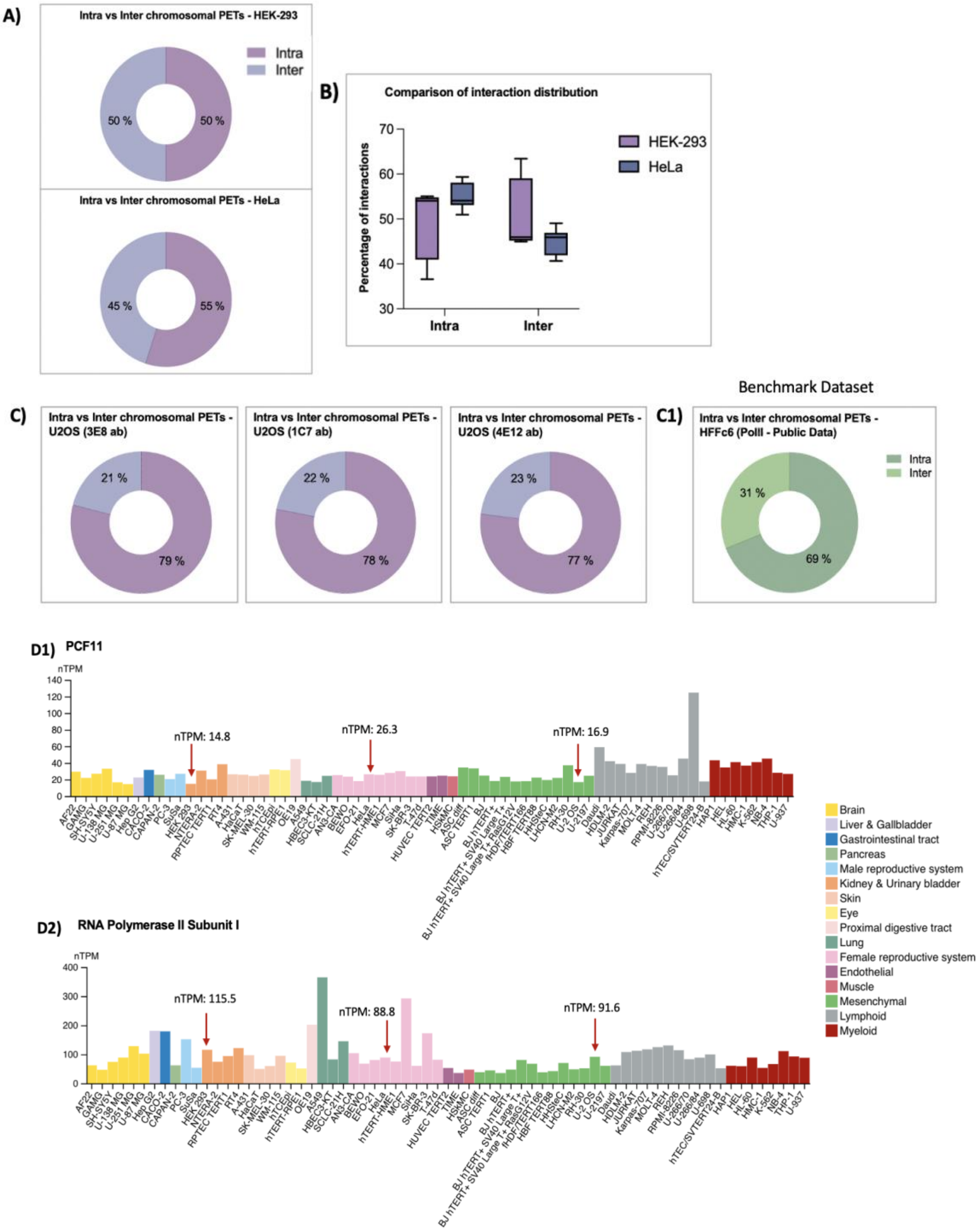


Fig. 2 Distribution of intra/inter chromosomal PETs in Datasets A and B. A) Donut plots of the average distribution of intra vs inter chromosomal interactions in Pcf11 HEK-293, HeLa, and Pol II U2OS ChIA-PET libraries. **B)** Boxplot representation of PET (intra/inter) distribution across the two datasets (four ChIA-PET libraries for each - HEK-293 SD= 8.912, HeLa SD = 3.021). **C)** PET (intra/inter) distribution of Pol II U2OS ChIA-PET libraries immunoprecipitated with three different Pol II antibodies. **C1)** Benchmark dataset (Pol II ChIA-PET in HFFc6 cells) PET distribution (Lee et al., 2020). **D1-2)** Total mRNA expression in normalised transcripts

per million (nTPM) of Pcf11 (D1) and Pol II (subunit I) (D2), in HEK-293, HeLa, and U2OS cell lines, as published on The Human Protein Atlas (available at: https://www.proteinatlas.org).

To further explore the datasets, principal component analysis (PCA) was performed on the basis of the peak-supported loops identified for each sample (**Fig. 3A-B**). As the samples are from different cell lines and given the dimensionality of a ChIA-PET dataset, variation can be attributed to a number of factors. Multivariate analysis reveals that a two-component reduction of the data allows for categorisation primarily on the basis of cell line. However, I can observe that sample FP_rep2 from the Pcf11 HEK-293 dataset seems to deviate from the rest of the HEK-293 libraries, indicating that another element could be affecting the clustering of this library with the rest of the samples. Library FP_rep2 yielded a much larger number of loops than the rest of the HEK-293 samples, something that could be contributing towards poor clustering with the rest of the HEK-293 samples. In a separate PCA plot of Dataset B only (**Fig. 3B**), as expected, variance can be further broken down into condition (Knock-down vs non-silencing control siRNA), with the knock-down samples showing greater variability.

I also looked at the proportion of PET counts per loop span distance. The distribution for Dataset B as shown in **Fig. 3C.** More PET counts map with loops that span shorter distances, indicating that most interactions occur between loci that are between 0-200 kb apart. This distribution is in line with that of the benchmark dataset's (Lee et al., 2020), as well as with published Pol II ChIA-PET data for the human cell lines MCF7 and K562 (Lareau and Aryee, 2021). It can also be observed that since Hi-C libraries represent the whole interactome map of DNA-DNA contacts, they appear to have a higher number of PETs, especially for longer-spanning loops.

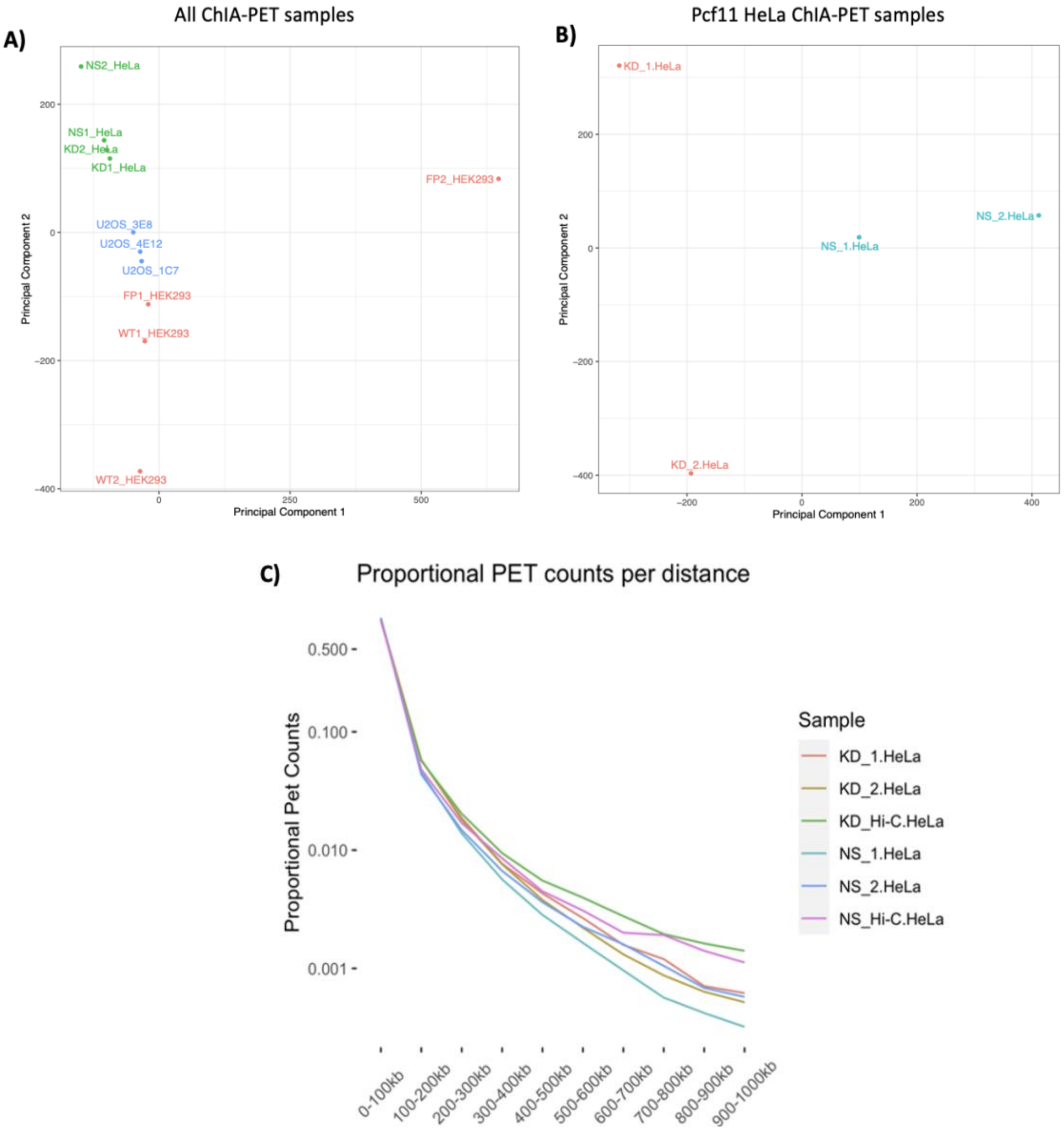


Fig. 3 Principal component analysis of Datasets A and B and PET count distribution of Dataset B. A) PCA plot of HEK-293 (pink), U2OS (blue), and HeLa (green) ChIA-PET libraries. **B)** PCA plot of HeLa ChIA-PET libraries only (knockdown samples = pink, non-silencing siRNA control = blue). **C)** Proportional PET counts per binned loop distance of HeLa libraries.

4.2.2 Visualisation of 2D Contact Maps, Loops, and Domains

The most common way to represent data derived from Hi-C and ChIA-PET experiments is in matrix format. The matrix format is perhaps the most helpful and well-established way of exploring the resolution of Hi-C libraries, identifying patterns such as chromatin contact domains (CCDs), and determining the depth of information that can be extracted downstream. **Fig. 4A-B** shows example matrices from some of our samples at different resolutions.

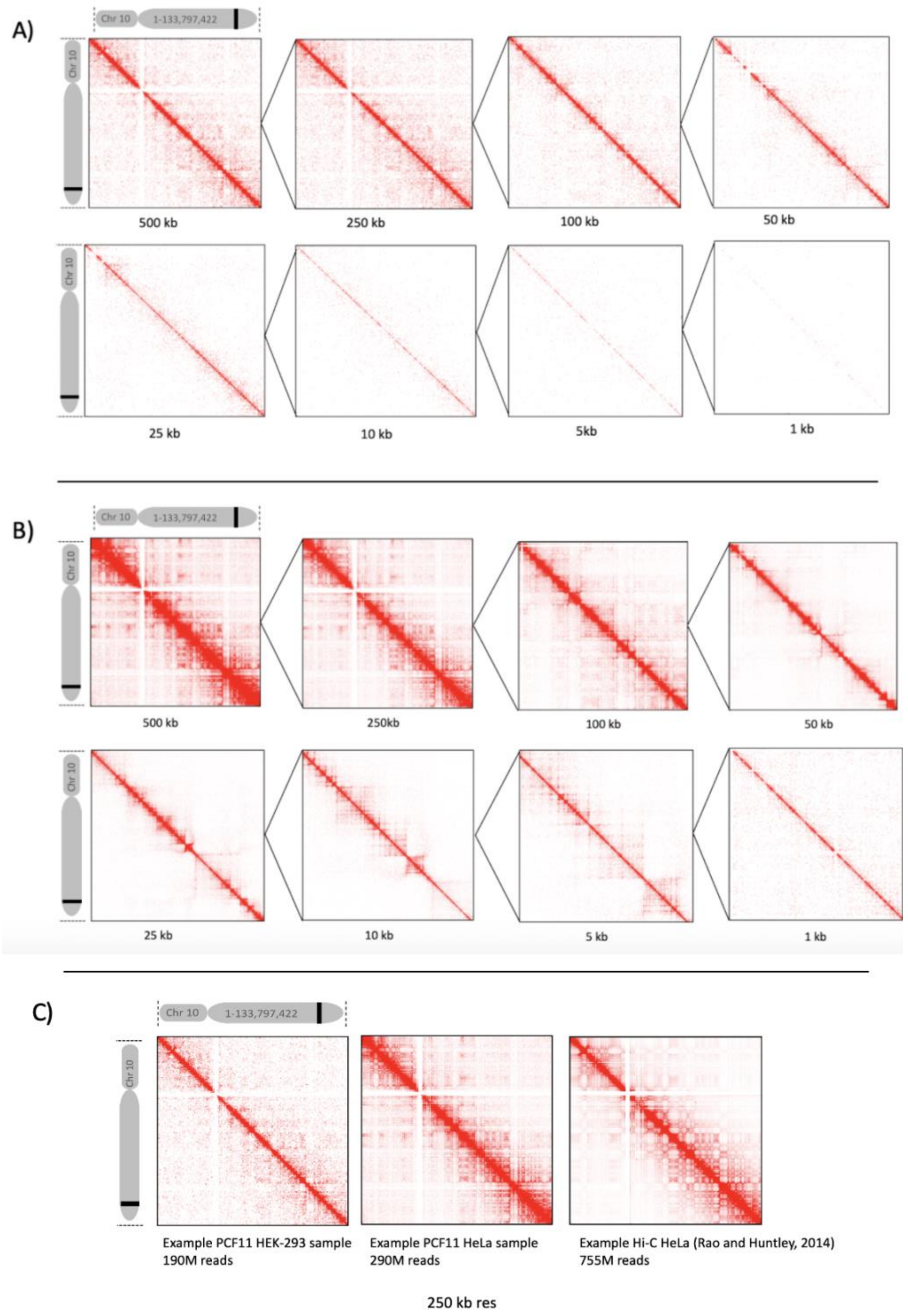


Fig. 4 Example matrices of Pcf11 ChIA-PET samples in HEK-293 and HeLa samples. A) Example interaction maps of sample FP_rep1 (HEK-293) for chromosome 10 across different resolutions (total sample reads = 190M). **B)** Example interaction maps of sample NS_rep1 (HeLa) for chromosome 10 across different resolutions (total sample reads = 290M). Maps visualised using Juicer (Durand et al., 2016). **C)** Closer comparison of HEK-293 and HeLa matrices with a Hi-C HeLa map generated through *in-situ* Hi-C (Rao and Huntely, 2014).

These maps highlight the key differences between the two datasets which are (i) sequencing depth, and (ii) total number of uniquely mapped PETs. Total numbers of interactions, as well as how many of these are peak supported will be presented in the next chapter. The example library from Dataset A (**Fig. 4A**) reveals a lower library resolution in comparison to the example sample from Dataset B (**Fig. 4B**). The matrices of the FP_rep1 library appear to be much sparser than those of NS_rep1, even at higher resolutions (>100 kb). This sparsity indicates the lack of high numbers of PETs, which translates to less total interactions captured. Especially at resolutions of <10 kb, we can only faintly observe interactions on the corresponding matrices. This contrasts with library NS_rep1 where the maps appear significantly more detailed with discernible interaction patterns even at <5kb resolution. These results are not surprising given the metrics previously explored in **Fig. 1A-B**. For comparison, **Fig. 4C** shows one of the first HeLa Hi-C maps to be generated through *in-situ* Hi-C, achieving single kb resolution (Rao and Huntely, 2014).

To investigate how the biological replicates of Datasets A and B correlate with each other, pairwise scatterplots comparing interactions at different resolutions were produced (**Fig. 5A-B**). The smaller the bin size of the matrix, the finer differences can be scored. HeLa samples (A) show a higher degree of correlation between them in comparison to the HEK-293 replicates (B), indicating higher library quality. We observe that at 250 kb resolution the HeLa replicates show an almost perfect positive linear relationship ($r \geq 0.98$). Even when considering smaller bins of 5 kb, the relationship between the replicates is still very strong ($r = 0.85$). As for the HEK-293 samples, they already display a weaker correlation between replicates at 250 kb ($r \geq 0.87$), and at 100 kb the discrepancies appear more prominent, with $r = 0.67$ for untreated samples, and $r = 0.69$ for FP-treated replicates.

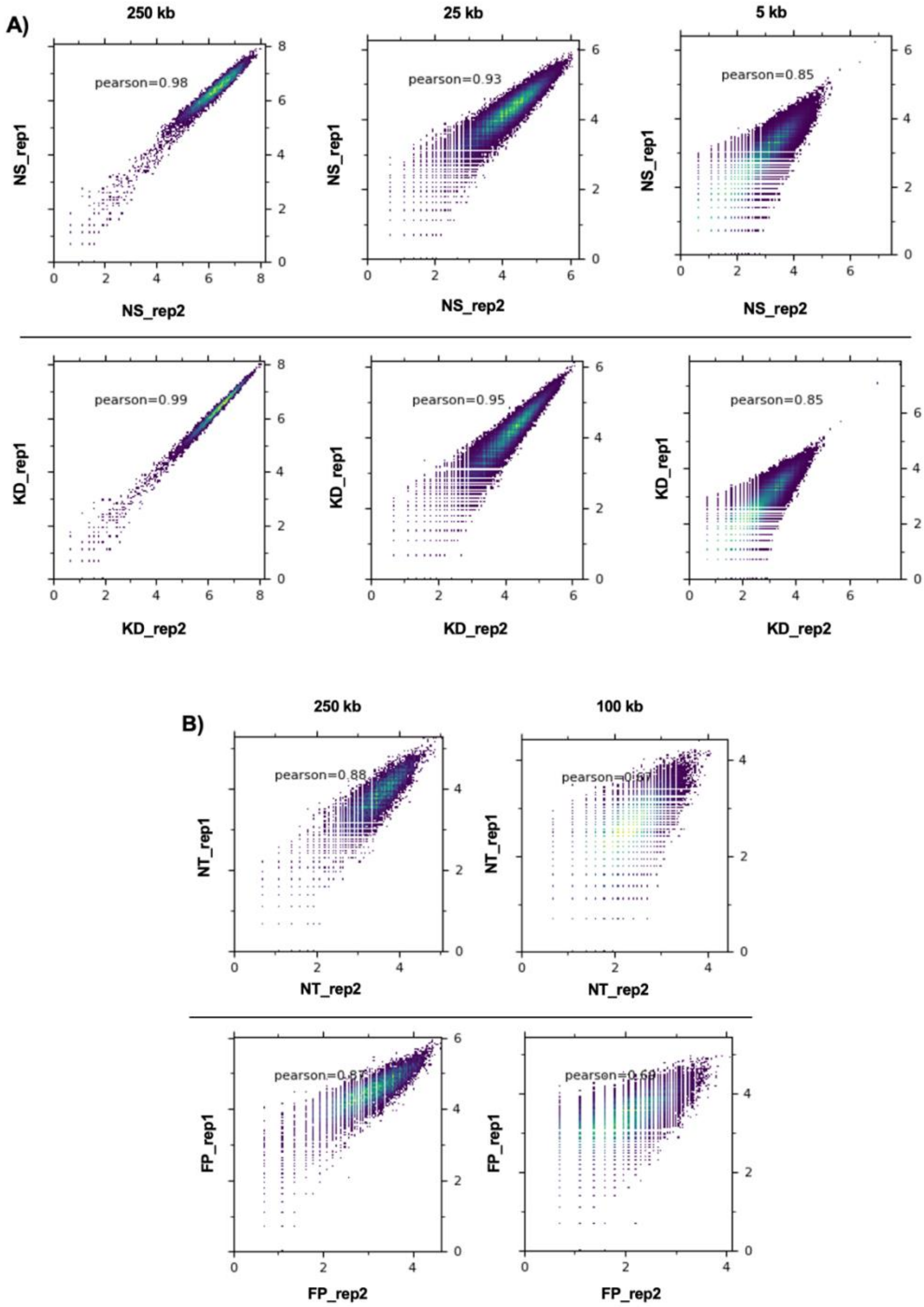


Fig. 5 Pairwise comparison of Pcf11 ChIA-PET replicates in HeLa and HEK-293 cells. A) Correlation of HeLa ChIA-PET interaction matrices. Top panel corresponds to N.S siRNA samples (NS) and bottom panel to siPcf11 replicates (KD). Strong positive linear correlation can be observed at all resolutions (250, 25 and 5 kb) between both replicates of the two conditions. **B)** Same as A, but for HEK-293 ChIA-PET samples. A weaker correlation is observed between replicates overall when compared to A). Even at a relatively high resolution (100 kb), the replicates of both conditions show a positive, yet weak correlation with a Person's correlation coefficient of $r = 0.67$ for non-treated (NT) replicates and $r = 0.67$ for Flavopiridol treated (FP) samples.

Dataset A (HEK-293) is of lower complexity compared to Dataset B (HeLa), which suggests that downstream analyses might be hindered. In Hi-C type experiments, the maximal effective resolution is based on a number of different factors, with the most important one being coverage (Lajoie et al., 2016). Dataset A was sequenced at a lower depth (~200M reads per library) compared to Dataset B (~300M reads per library). This alone can in part explain why we observed fewer interactions in Dataset A. Library complexity is another factor that strongly impacts the resolution in the present maps. In the case of Hi-C and ChIA-PET experiments, library complexity can be defined as the total number of unique PETs present in each sequenced sample. Number of unique PETs is a factor of both starting material (e.g., total number of cells), and overall library quality. Libraries with low complexity saturate faster as sequencing depth increases, e.g., additional sequencing does not significantly increase gain of information which holds true for Dataset A as due to lower starting material more PCR cycles were needed to amplify the library to reach DNA levels suitable for sequencing. Therefore, a fair assumption here is that even with increasing read depth, the number of cumulative unique interactions did not increase beyond a certain level and saturated at a lower depth compared to Dataset B.

Literature suggests that for *in-situ* ChIA-PET libraries of the human genome, a large enough number of PETs by high-throughput DNA sequencing is needed to generate enough data for genome-wide coverage. This can range from 100M-1B PETs. The authors of the *in-situ* ChIA-PET protocol report on routinely producing 200-500M PETs per library (Wang et al., 2021). This range provides sufficient coverage for chromatin loops and protein binding peaks. In absolute numbers, the total number of PETs for Dataset A libraries is 50-60M, whereas for Dataset B it is 160M, and therefore a lot closer to the authors' recommended range.

For downstream analyses in Chapter 5, I will also utilise Pol II ChIA-PET libraries from U2OS cells. These libraries were generated through the same protocol as Datasets A and B and each of them (n=3) has been immunoprecipitated with a different clone of Pol II antibody (Abcam's ab252854, ab252853, and ab252852). Pol II ChIA-PET data have been produced before, although not in U2OS cells. **Fig. 6** shows interaction matrices between the three libraries immunoprecipitated with different Pol II antibodies in U2OS cell lines, as well as publicly available ChIA-PET data for Pol II (Lee et al., 2020) in HFFc6 cells for comparison. Through visual comparison we observe that different cell lines seem to have conserved interactive regions. We are also able to observe that some larger interactive regions appear to visually resemble those of HFFc6 cells (**Fig. 5** top panel).

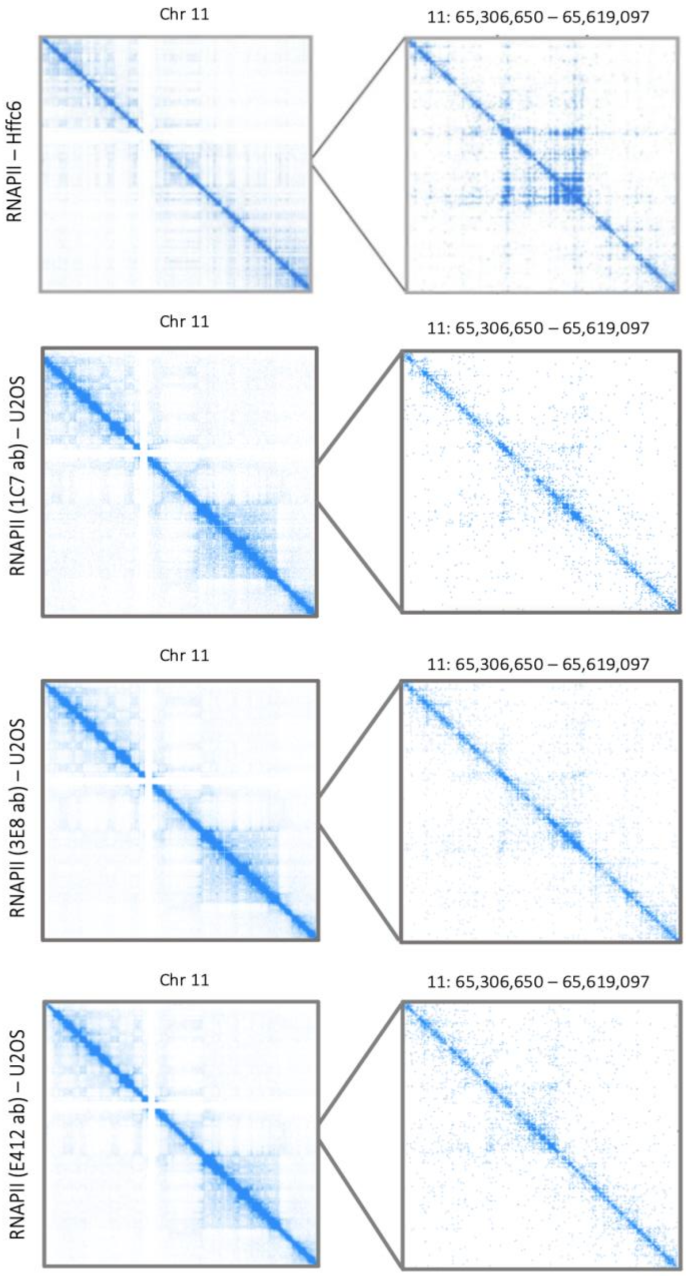


Fig. 6 RNAPII interaction matrices in Hffc6 and U2OS cell lines. The top panel shows the interaction matrix for chromosome 11 in Hffc6 cells (left) and the corresponding matrix for position 65,306,650 - 65,619,907 on the same chromosome at 1kb resolution (right), as published in Lee et al., 2020. The second, third, and bottom panels show the equivalent positions for U2OS cells, immunoprecipitated with the 1C7, 3E8, and 4E12 RNAPII antibodies, respectively (for specific antibody product numbers see Chapter 2: Materials and Methods).

Loop and domain annotations is another important aspect of ChIA-PET data that should be checked before performing more in-depth and specific analyses relevant to our research question. Although not a lot of biological information can be observed from merely overlaying these annotations on our matrices, it is still important to perform these quality checks to capture any potential issues with the data processing before continuing with further analysis. In the case where publicly available data exists, it is also useful to perform side-by-side comparisons to check for consistency. **Fig. 7** shows an example of one of our samples with the relevant loops and domains overlaid on the interaction matrix, along with the genome coverage and peaks tracks. The bottom panels (**Fig. 7B-C**) indicate different, more informative ways of exploring loops and contact maps.

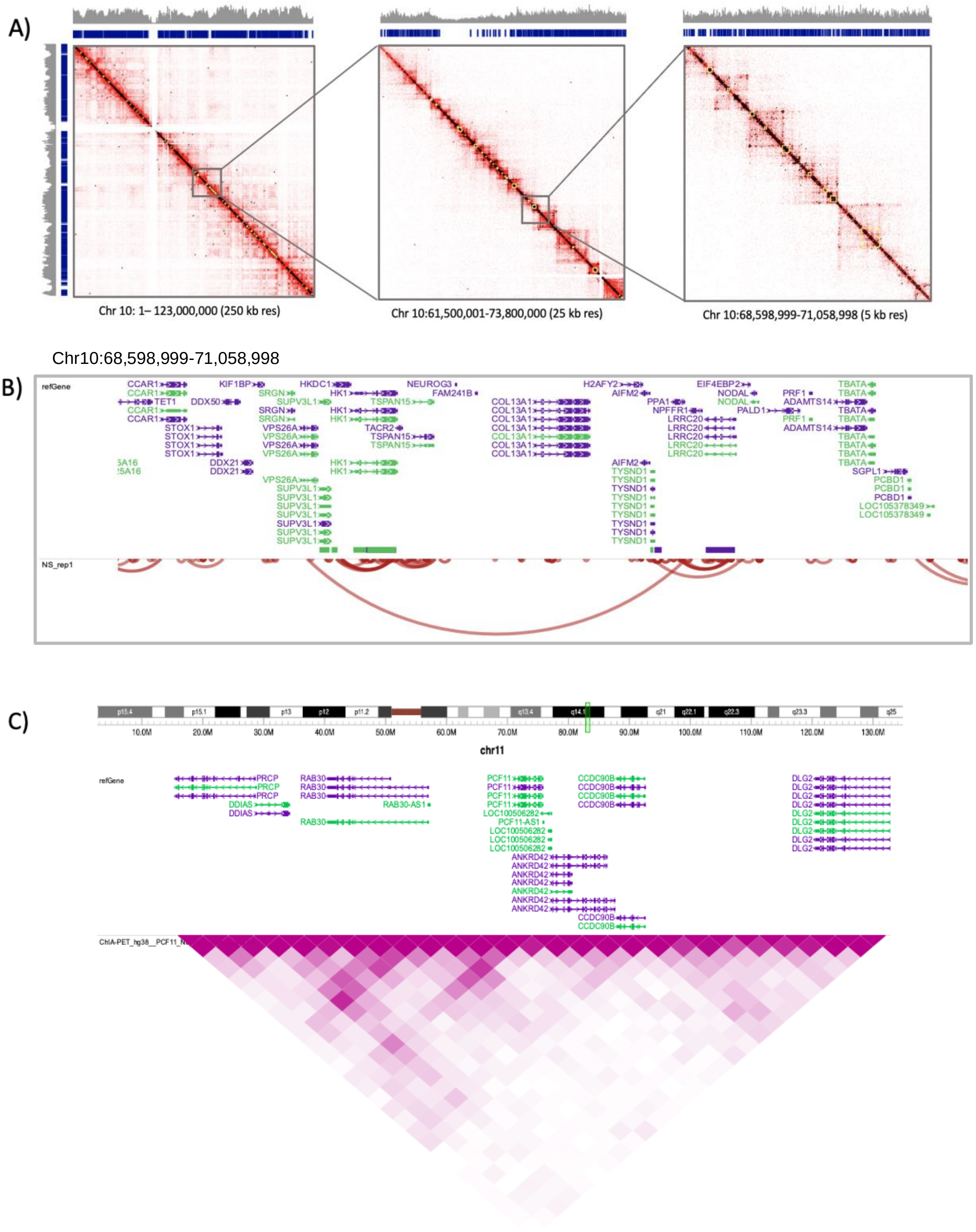


Fig. 7 Exploring loops and domains on HeLa Pcf11 ChIA-PET data. A) The top panels indicate matrices for sample NS_rep1 for different positions on chromosome 10. Top left panel shows loops (black) and domains (yellow) overlaid over the entirety of chromosome 10, along with the genome coverage (grey) and peak (blue) tracks, at 250 kb resolution. The middle panel provides a zoomed-in image of a specific position on chromosome 10 at 25 kb resolution, and the right panel indicates a more zoomed-in image on the same chromosome

at 5 kb resolution. **B)** The bottom panel shows the corresponding loops region from the top right matrix (chr 10: 68,598,999 - 71,058, 998) on the WashU Epigenome Browser (Zhou et al., 2011) along with the refGene gene annotations for hg38. **C)** Visualisation of contact matrix in the form of a heatmap on WashU Epigenome Browser (Zhou et al., 2011) for a location on chromosome 11, near the Pcf11 gene.

As there are available ENCODE ChIP-seq datasets for HeLa cells, I explored whether signals for histone marks match what we see in other Hi-C/ChIA-PET datasets. We expect that ChIP-seq signals from certain histone marks will be consistent with chromatin domains. **Fig. 8** shows an example of one of the HeLa ChIA-PET libraries (NS rep 1). We observe that the ChIP-seq track for H3K36me3 (Accession: ENCFF533LRJ) appears to match TAD locations.

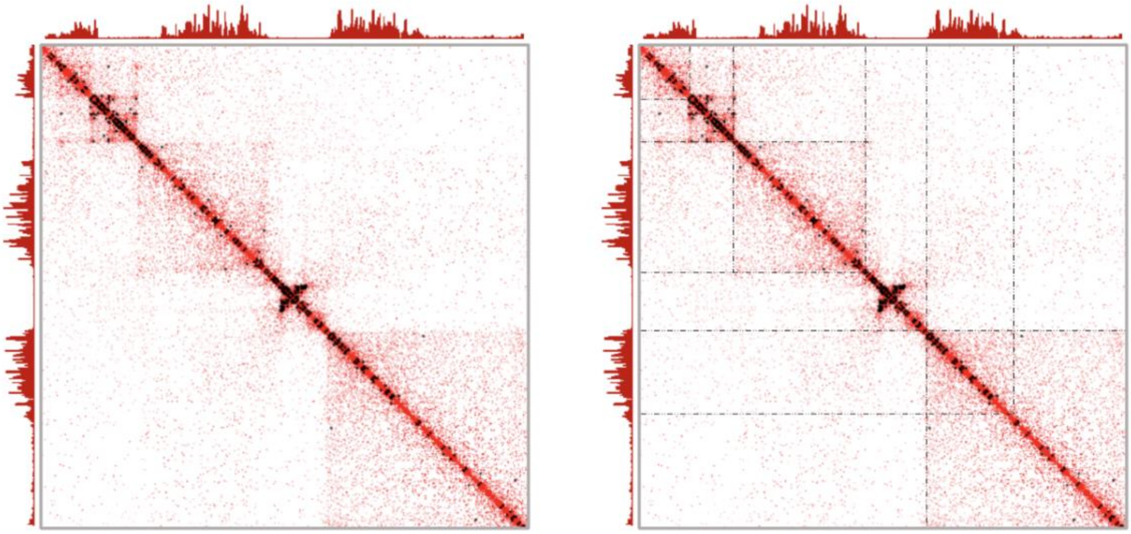


Fig. 8 H3K6me3 ChIP-seq signal on NS_rep1 ChIA-PET contact matrix. H3K36me3 (Accession: ENCFF533LRJ) indicates actively transcribing chromatin and is consistent with highly interactive regions. Left panel shows the raw matrix and ChIP-seq signal track and right panel shows the same area with dotted lines overlayed over TAD regions, both at 5 kb resolution. (Location: Chr15:91,578,979-94,038,978).

ChIA-PET libraries processed through the ChIA-PIPE pipeline provide vast amounts of data. To manually navigate through the contents of all the output files would be an impossible task. However, the above analyses are necessary for determining data quality and detecting patterns between datasets, as well as comparing with published data where available. These preliminary quality-check type of analyses are widely used to determine how well the technique itself has worked and the level of information I should be expecting to extract downstream (Rao and Huntely 2014, Li et al 2017, Wang et al., 2021). In order to be able to extract biologically meaningful information from the annotation files (e.g., loops files), custom tools are required, something that will be explored in the next Chapter.

Next, I assess the immunoprecipitation (ChIP-seq) aspect of the data.

4.2.3 Assessing Immunoprecipitation

Since ChIA-PET data also contain information equivalent to that of ChIP-seq libraries, it is useful to assess the quality of the data in that regard as well. As there are published ChIP-seq data of Pcf11 using the same antibody on the same cell line (Kamieniarz-Gdula et al., 2019), this dataset was used as a benchmark for comparison. The publication of these ChIP-seq libraries revealed the binding profile of Pcf11, which formed the basis of the present work. Although human Pcf11 is considered to be a termination factor and therefore expected to be present at the 5' ends of genes, ChIP-seq data surprisingly revealed its localisation on the TSS as well (Kamieniarz-Gdula et al., 2019). Given these findings, I was expecting to see a similar binding profile for our Pcf11 ChIA-PET samples. **Fig. 9A-B** shows heatmaps around the peak region, as well as coverage plots for the Pcf11 ChIA-PET samples in HeLa cells, after normalisation by the corresponding input control. The coverage plot of publicly available Pcf11 ChIP-seq data on HeLa cells using the same antibody for immunoprecipitation is also shown for comparison (GEO Accession: GSE127256).

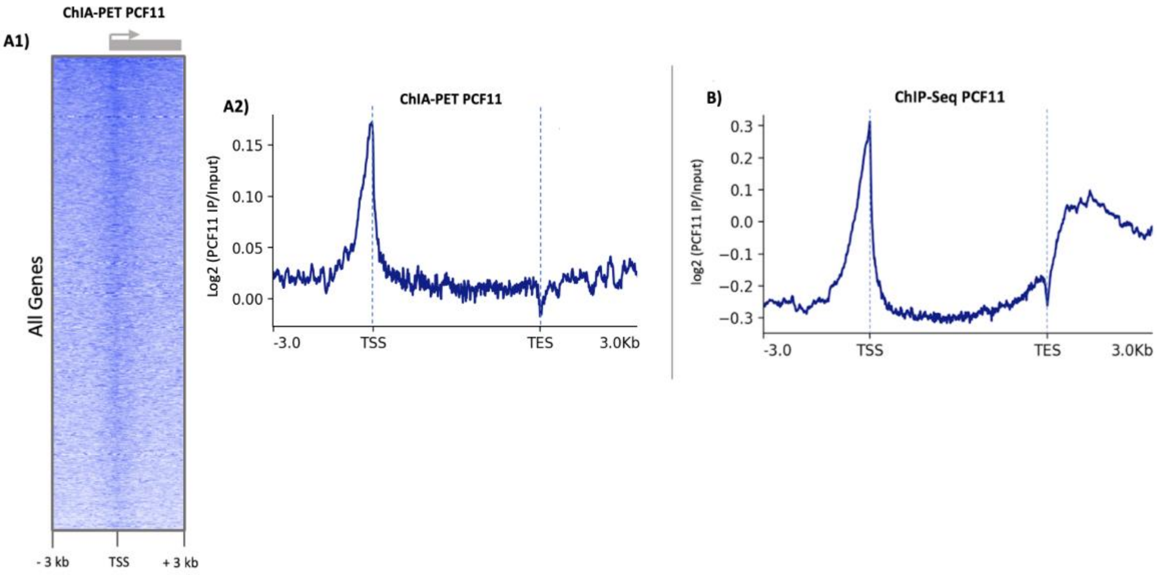


Fig. 9 Metagene (coverage) plots of Pcf11 ChIA-PET and ChIP-seq data. A1) Signal intensity heatmap around the peak region (TSS) **A2)** Coverage plots showing the average signal across all genes (hg38 annotation). High signal can be observed at the TSS. **B)** Coverage plot of publicly available ChIP-seq Pcf11 library on HeLa cells across all genes on the human genome (hg38 annotation). Peaks can be observed near the TSS and TES, suggesting that Pcf11 localises close to the promoter and polyadenylation sites (PAS) of genes.

The coverage plot corresponding to the publicly available Pcf11 ChIP-Seq data (**Fig. 9B)** reveals the occupancy profile of the protein of interest on HeLa cells. On average, Pcf11 appears to be enriched both nearby the TSS and TES of genes. TES enrichment is not surprising as we already know the factor to be involved in polyadenylation. Its high enrichment on the TSS creates a number of possible hypotheses. For example, Pcf11 could have additional functional roles in transcription initiation, or its presence at the start of genes might indicate that it gets recruited by Pol II as soon as its CTD becomes phosphorylated after transcription initiation (Guo et al., 2019).

An equivalent plot for ChIA-PET Pcf11 in HeLa cells (**Fig. 9A1-2**) reveals a different occupancy pattern as significant enrichment on the TES/PAS cannot be observed. However, this does not necessarily suggest an unsuccessful ChIP reaction given that coverage of ChIA-PET and ChIP-seq differ. In a ChIA-

PET experiment, only interactive regions will be pulled-down, something that inevitably affects the overall coverage. Since Pcf11 ChIA-PET data have not been generated before and comparison with a pre-existing dataset was not possible, I compared the coverage of my own Pol II ChIA-PET libraries with that of publicly available Pol II ChIP-seq and Pol II ChIA-PET samples. **Fig.10** shows the corresponding coverage for these libraries, highlighting differences between Pol II ChIP-seq and ChIA-PET. We observe close resemblance of our Pol II samples to that of previously published ChIA-PET data and notice that the coverage for the ChIP-seq dataset differs notably. This indicates that the two techniques are not directly comparable when compared on a single dimension.

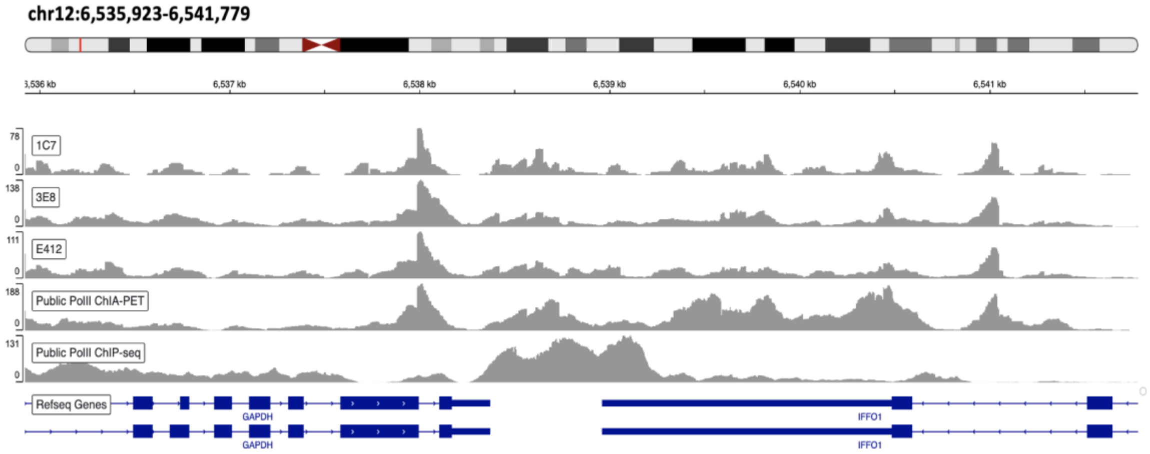


Fig. 10 Coverage plots of Pol II ChIA-PET and ChIP-seq samples (hg38) in region chr12:6,535,923-6,541,779. Top three panels correspond to the three Pol II ChIA-PET datasets (1C7, 3E8, E412) in U2OS cells, fourth panel corresponds to a publicly available Pol II ChIA-PET sample generated in GM12878 cells via the same protocol (4DN Portal Accession: 4DNESZ25MOZV), and the bottom panel is from Pol II ChIP-seq data in HeLa cells (Encode Accession: ENCSR000BGO). Very similar coverage is observed for all Pol II ChIA-PET libraries, with the publicly available library showing higher signal across the last exon of IFFO1. The Pol II ChIP-seq sample reveals a notably different coverage profile.

In the next chapter, peak-annotation will be performed to determine which of the captured interactions are enriched for Pcf11. Additionally, the loop annotation analysis will be repeated, and instead, the peak-called regions of the publicly available ChIP-seq data will be used to annotate loops. Using

ChIP-seq data is a common way of annotating loops identified from Hi-ChIP experiments since the peaks called from a separate ChIP-seq experiment might be more reliable (Dori and Forcato, 2021). It is worth noting that the original authors of the Pcf11 ChIP-seq data, for their analyses, only considered regions to be Pcf11-enriched if they overlapped between two distinct Pcf11 antibodies- a Pcf11 antibody targeting an internal epitope (Pcf11-Int), and a Pcf11 antibody targeting a C-terminus one (Pcf11-Ct) (Kamieniarz-Gdula, 2019). However, although through my analysis the combined Pcf11-Ct antibody replicates call for 73,143 peak regions, and the Pcf11-Int libraries for 29,877, only 13,381 of those regions overlap between them. This further adds onto the previously observed concern that Pcf11 antibodies can be inconsistent. To check whether the replicates from our datasets overlap in terms of ChIP peaks, Venn diagrams were generated. **Fig. 11** shows Venn diagrams of peak overlaps between Pol II and Pcf11 ChIA-PET libraries.

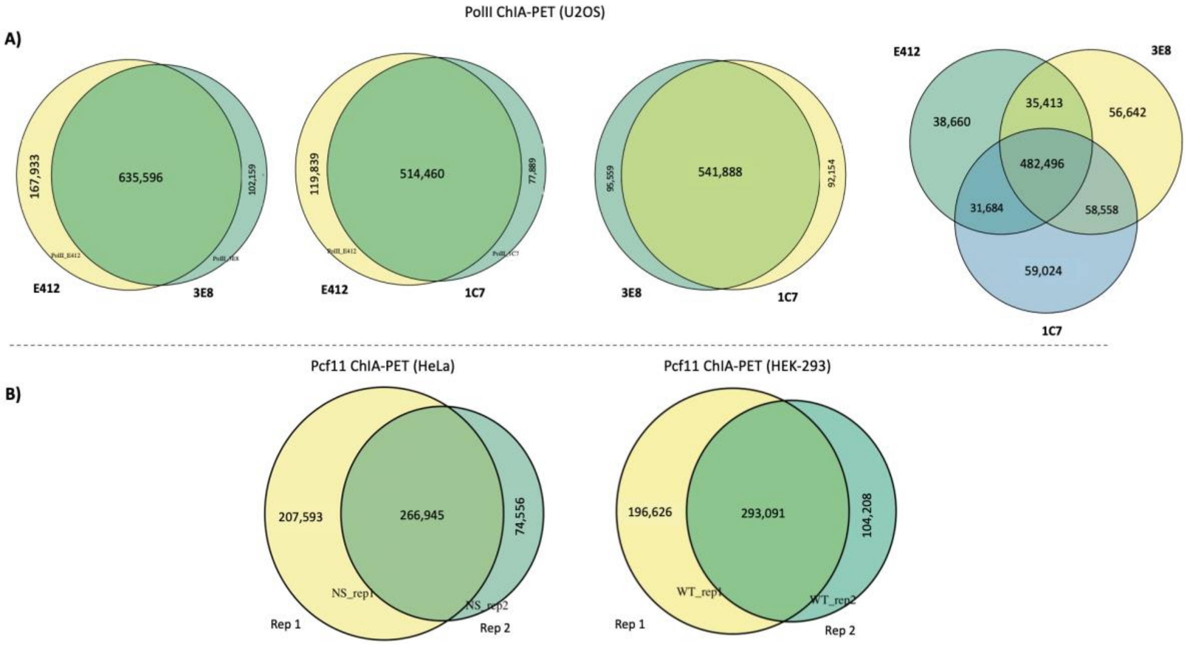


Fig. 11 ChIP peaks overlap across samples and replicates. A) Venn diagrams of all three Pol II libraries. Diagrams show the overlap between each of the libraries amongst themselves, as well as together. **B)** Venn diagrams showing the ChIP peak overlap between replicates of the Pcf11 HEK-293 (right) and Pcf11 HeLa (left).

Although different antibodies were used for the three Pol II libraries, the peaks overlap well between samples (>80% overlap, $p < 0.01$). The two Pcf11 replicates in the HeLa ChIA-PET dataset also show significant overlap ($p < 0.01$). However, replicate 2 has generated less total peaks which is most likely a result of weaker ChIP signal overall. A similar overlap profile can be observed for the HEK-293 dataset.

Overall, these analyses revealed the strengths and weaknesses of each of the datasets and established the essential foundation for downstream analyses in order to address the main research question. The HEK-293 samples represent libraries with lower complexity which might impact further analyses and limit the biological information that can be inferred. The HeLa and U2OS samples meet the QC standards and have an equivalent QC profile to that of published ChIA-PET datasets.

4.3 Discussion

4.3.1 Quality Control During Library Preparation

The validity of next generation sequencing (NGS) data and the amount of information that can be extracted downstream is often heavily dependent on factors influencing data quality. On the most basic level, this can simply refer to adapter contamination and GC content. Generally, when it comes to NGS, there are three main areas where QC can be applied: (i) starting material (e.g., DNA), (ii) after PCR library preparation, and (iii) after sequencing. Ultimately, the most relevant and accurate QC will be derived from the raw sequencing data.

In the context of ChIA-PET data, in-depth examination of data quality is of high importance due to the multi-level nature of the dataset. As it has been mentioned in the Introduction section of this chapter, ChIA-PET libraries are generated through a lengthy experimental process that requires a large number of cells and reactions. Ensuring library quality is therefore important, especially when repeating experiments or re-sequencing might not be feasible. For these reasons, the original authors of the protocol recommend six stages for QC during the experimental process; post (i) restriction digestion, (ii) bridge linker ligation, (iii) immunoprecipitation, (iii) transposase tagmentation, (iv) PCR amplification of the library, and (v) after size selection (Li et al., 2017, Wang et al., 2021). Primarily, QC steps are performed through quantifying DNA and determining overall fragment distribution which indicates whether certain reactions were successful and to what extent. As for ChIP assessment, ChIP-qPCR can be used. Due to limited amounts of DNA and several Western blot experiments having confirmed antibody validity to a certain degree, this step was omitted. All other steps were individually optimised during test-runs of ChIA-PET. Example QC profiles after restriction digestion with *Alu*I, after bridge linker ligation, and after transposase tagmentation are shown in Chapter 2: Materials and Methods. These match the profiles indicated in the ChIA-PET protocols both by Li et al., (2017) and Wang et al., (2021). The authors also recommend the use of a BluePippin (Sage Science) machine to size-select fragments between 300-600 bp prior to sequencing. However, as the facilities I had access to lack such a machine and size selection through a DNA gel would result in significant loss of library sample, I tried to ensure a fragment range as close to the recommended one as possible through the use of PCR clean-up beads. The authors also recommend that prior to total sequencing, a test sequencing run be completed. They suggest that a small library sample is sequenced to a depth of ~ 20-30M reads using MiSeq. As we outsource all the sequencing, extra MiSeq runs would significantly prolong the turnaround time of the complete dataset by several weeks. Therefore, it was decided to bypass this step and proceed with complete sequencing.

Overall, several smaller experiments were carried out to ensure that aspects of the *in-situ* ChIA-PET protocol were optimised according to the authors' notes (e.g., cell and nuclear lysis, sonication, bridge linker ligation, etc). Given that the protocol was only formally published very recently (Wang et al., 2021), datasets generated using the same technique are scarce. Hence, optimisation experiments and extensive QC constituted important aspects of this research work.

4.3.2 Quality Control After Sequencing

Since ChIA-PET is a method that combines 3C and ChIP, the way to assess library quality differs from other NGS experiments. As the present samples are some of the first ChIA-PET libraries to be generated using the *in-situ* ChIA-PET protocol, and the first Pcf11 ChIA-PET libraries in general, it was important to use publicly available datasets as benchmarks for comparison and thoroughly assess quality before trying to address deeper biological questions.

The first dataset (Pcf11 ChIA-PET in HEK-293) was generated as a back-up to the main dataset (Pcf11 ChIA-PET in HeLa) after repeated attempts to confirm a Pcf11 knockdown were unsuccessful. However, once the knockdown was eventually validated, the original dataset was generated as well. The Pol II U2OS libraries were initially generated for different research purposes, but they are incorporated into this work to facilitate QC since several Pol II ChIA-PET libraries exist for direct comparison, and to also explore 3' - 5' end crosstalk bound by Pol II in Chapter 5. Although having a knock-down as control is not essential to uncovering the looping profile of a factor of interest, both knock-down and Hi-C libraries were generated to act as appropriate controls for downstream analyses (Chapter 5). Publicly available datasets do not appear to be using these controls, however I wanted to make sure to cover most aspects of normalisation and experimental control (e.g., generating input Hi-C libraries for peak-calling control). Particularly, I considered it important to have input controls for the ChIP aspect of the dataset in order to eliminate as much of the noise as possible, something that we do not observe in other publicly available ChIA-PET data.

Due to significant time spent on optimisation experiments it was necessary to generate a complete dataset to address the research question. Therefore, although the technique would have benefited from further optimisation prior to producing the HEK-293 dataset, it was important to generate data. The rationale for the FP-treated HEK-293 dataset lies within the fact that Pcf11 is recruited by Pol II's phosphorylated CTD tail to assist in termination (Kuehner et al., 2011), and that the transcription inhibitor FP prevents p-TEFb from phosphorylating Pol II (Chao and Price, 2011). Although FP is a potent inhibitor that would be expected to cause several genome-wide effects, it would still act as a form of control for the observed Pcf11 looping profile in untreated cells, potentially also giving us insights into how Pcf11-bound loops are altered after transcriptional arrest. However, QC has revealed that this dataset falls below the standards of high-quality ChIA-PET libraries as indicated by Wang et al (2021). A low number of unique PETs reveals a dataset that is less complex from what is expected. Both metrics (**Fig. 2A**) and contact maps (**Fig. 4A**) highlight the impact of a low number of unique interactions. This does not necessarily suggest that the dataset cannot be used for further analyses, but it does indicate that a degree of biological information might be lost. This is in contrast to the other two datasets (Pcf11 ChIA-PET in HeLa and Pol II in U2OS) that were generated after further experimental optimisation was completed, prior to sequencing. These samples benefited from higher number of starting material, smaller loss of DNA sample after reactions, and fewer PCR cycles before sequencing. Both HeLa and Pol II datasets meet the QC standards and closely resemble those of the benchmark dataset (Lee et al., 2020), therefore I expect to extract more information with a higher degree of confidence downstream.

In regards to the ChIP aspect of the dataset, the Pcf11 ChIA-PET libraries do not appear to fully resemble the coverage profile of publicly available Pcf11 ChIP-seq libraries, as occupancy downstream of the TES cannot be observed. Generally, the ChIP signal from ChIA-PET datasets is weaker than that of ChIP-seq data near the relevant binding sites (Fullwood et al., 2009), however, we assume that the pull-down of only interactive regions is affecting the observed coverage. Since these libraries are the first ChIA-PET samples for the factor Pcf11, comparing their coverage profile to identical datasets was not possible. However, for downstream analyses, the ChIP peaks identified from publicly available Pcf11 ChIP-seq libraries on HeLa cells will also be used for annotating loops, allowing for comparison.

4.3.2.1 Methods of QC in 3C Datasets

The most common way of determining the quality of data generated through Hi-C techniques, at least in terms of resolution, is through the construction of contact maps. Contact matrices are able to provide a “zoomed-out” and overall image of all interactions identified in a dataset. A key feature that emerges in the contact matrix is that of topologically associated domains (TADs), i.e., regions that interact with each other at a much higher frequency than they do with other genomic locations (high intradomain and low interdomain contact frequency) (Sexton et al., 2012). On a finer scale, *cis* contacts, or chromatin loops, can also be identified and visualised (**Fig. 7A-B**). The resolution of these maps and the number of TADs and loops that can be identified will primarily be determined during the experimental procedure, as well as sequencing depth. The resolution of the maps shown in **Fig. 4A-B** and **Fig. 6** highlight this, as 1 kb resolution was achieved for the Pcf11 HeLa and Pol II U2OS libraries. This is both a result of deeper sequencing, as well as further experimental optimisation. Additional ways to improve resolution would be to use restriction enzymes that recognise more frequent restriction sites and generate smaller fragments and improve resolution. For example, a variation of the Hi-C protocol, the COLA (Concatemer Ligation Assay) protocol, uses a restriction enzyme that is able to recognize an RCGY region where R can be either A or G, and Y can be C or G. This protocol yields an even smaller average fragment size and increases the proportion of reads containing three or more fragments of close proximity (Darrow et al., 2016). However, in the case of ChIA-PET, it is possible to achieve 1 kb resolution at a much lower sequencing depth due to the fact that the ChIP step will mostly retain fragments that are bound by the factor of interest, and therefore there is no need to generate billions of reads, something that is common of more recent Hi-C datasets (Belaghzal et al., 2017, Rowley et al., 2017).

Several efforts have been made to build databases that store publicly available chromatin architecture data. These databases include large consortia such as ENCODE, Roadmap epigenomics, and the 4D Nucleome consortium (Pal et al., 2018). With the rapid generation of this type of datasets, the importance of establishing standardised procedures to assess reproducibility of replicates as well as data quality is prominent. Since the *in-situ* ChIA-PET protocol was only very recently made publicly available (Wang et al., 2021), the present datasets represent some of the first libraries to have been generated with this protocol.

As a result, all software tools and quality control standards applied to the present work made use of all the recommendations, analysis pipelines and datasets found in the latest relevant publications from Wang et al. (2021) and Lee et al. (2020).

Interaction-calling and data visualisation is another important aspect of QC in the context of interactome data analysis. Several tools and algorithms exist for the identification of interactions and TADs; however, these have a high level of variation between them (Forcato et al., 2017, Zufferey et al., 2018). High levels of variation between tools is a bigger issue when analysing ChIA-PET data since common Hi-C algorithms are not suitable for more sparse matrices like those produced from ChIA-PET experiments. However, a number of tools specifically designed for the analysis of ChIA-PET data exist. In the present work, all pre-processing and interaction-calling was performed using ChIA-PIPE which is the latest available tool for the analysis of ChIA-PET data (Lee et al., 2020). Although an established critical issue in the field of Hi-C is the lack of common standard data formats, the outputs from the ChIA-PIPE pipeline are compatible with several visualisation software, such as Juicebox, HiGlass, and commonly used genome browsers. This allowed us to compare our datasets to other publicly available maps stored on the visualisation interfaces themselves (**Fig. 4C**). It is worth noting that the visualisation tool developed and recommended by the developers of the ChIA-PIPE pipeline (Lee et al., 2020), known as BASIC Browser, has not been kept up to date, and although it provides some additional benefits for the visualisation of ChIA-PET data, it could not be utilised.

Overall, through visualisation of 2D contact maps, I was able to conclude that the Pcf11 HeLa and Pol II U2OS datasets meet the criteria for complexity and QC for ChIA-PET samples.

4.3.3 Downstream Analysis

Downstream analysis includes all the methods that can be applied for the extraction of biologically meaningful information. A degree of this has already been explored in this chapter by plotting matrices at multiple levels of resolution, overlaying loops and CCDs, and calling for ChIP peaks. To investigate whether Pcf11 and Pol II are implicated in gene looping and to explore how this relates to transcriptional activity, tailored analyses are required.

As the present preliminary analyses have identified the strengths and weaknesses of each dataset and facilitated QC, I can now begin to explore more specific questions. In the next Chapter, I will be exploring gene looping by further analysing the identified ChIP peak-supported *cis* interactions of each dataset. Our aim is to identify whether gene loops bound by Pcf11 and/or Pol II occur in protein-coding, transcriptionally active genes, as well as whether Pcf11's TSS enrichment is the result of looping. I will also perform suitable statistical tests to establish if this is a phenomenon that occurs by chance, or whether it holds functional and regulatory relevance.

Chapter 5: 3' - 5' end Crosstalk and Gene Expression

5.1. Introduction

5.1.1. Chromatin and Gene Looping

The looping of chromatin bound by proteins and protein complexes is a common phenomenon that has implications in fundamental cellular processes, such as transcription and DNA replication (Saiz and Vilar, 2006). Advances in the field of chromatin conformation capture (3C) have aided our understanding of the properties governing chromatin looping and how these properties relate to the regulation of genes and gene clusters. Two of the most prominent examples of far-acting (*trans*) gene regulation are those of the human *Hox* clusters and the *β-globin* locus control region (LCR).

Hox genes code for developmental regulators that are evolutionarily conserved. Their transcriptional silencing is a key feature since *Hox* gene expression at the wrong time can result in disease states (Ferraiuolo et al., 2010). Ferraiuolo and colleagues (2010), through the use of 3C technology, showed that transcription activity-dependent long-range chromatin contacts were a feature of all four *Hox* silent clusters. The β-globin locus control region (LCR) is another fascinating example of how eukaryotic transcription and gene expression could be regulated through long-range interactions. The LCR is located at a linearly distant position to the β-globin gene on human chr11, but during active transcription these two regions come into close proximity. This phenomenon suggests that a looped structure could be bridging the gap between the LCR and the β-globin gene (Kim and Dean, 2012). The formation of chromatin loops appears to be a critical event in transcription, both in activation and repression, although the underlying reason as to why that is remains unclear mainly because we do not yet fully understand how loops are initially formed.

A subset of such *cis* interactions is what appear to be gene loops. Gene loops differ from chromatin loops as they refer to the juxtaposition of the starts and ends of genes. Although the phenomenon of gene looping is a relatively recent finding, it has been observed in several organisms; from prokaryotes such as *E. coli* (Cournac and Plumbridge, 2013) and lower eukaryotes such as yeast (O’Sullivan et al., 2004, Singh et al., 2009), to higher eukaryotes like *D. melanogaster* (Rowley et al., 2017) and human cells (Tang et al., 2015). Gene loop formation has been suggested to be a prevalent occurrence in protein coding genes in eukaryotes, something that is corroborated by Pol II’s functional roles in transcription (Hebenstreit, 2013). Mapendano and colleagues (2010) were able to demonstrate that inducing mutations to the polyadenylation site of a specific gene decreased the initiation of transcription of that gene. These findings could suggest that a number of factors implicated in the transcriptional cycle - from initiation to termination - could be working synergistically, and looping could be acting as a means of facilitation so that these factors can interact with one another.

5.1.2. Looping and the Role of Pol II and Pcf11

It has been suggested that a proportion (< 5%) of actively transcribing genes are predicted to be looped (Grosso et al., 2012), and previous work from our lab utilising publicly available Pol II ChIA-PET data from K-526 cells has revealed a positive correlation between transcriptional bursting and level of gene looping (Cavallaro et al., 2021). The observed occupancy profile of Pcf11 in human cells at the TSS of genes (Kamieniarz-Gdula et al., 2019) provides an indication that Pcf11 might be present at the 5’ ends of genes through the formation of loops. Given that Pcf11 is not known to have a functional role in transcription initiation or elongation, its strong enrichment at the TSS could be an indication that the factor’s presence close to the promoter region is the result of gene looping, or that perhaps, the factor could indeed be directly involved in transcription initiation.

Pol II’s involvement in chromatin looping has long been speculated (O'Reilly et al., 2007, Shandilya et al., 2012, Allepuz-Fuster et al., 2019) due to its significant role in transcription. Loops can facilitate distant communication of enhancers and promoters and allow Pol II to come into close proximity with the TSS. Enhancers, although short DNA sequences spanning only 20-400 bp, can activate transcription by targeting *trans* promoters over long genomic distances (Bondarenko et al., 2003). In most cases, it appears that the way enhancers operate is through interaction with a promoter. This can be achieved through the formation of loops by protein-bound enhancers and promoters (Bondarenko et al., 2003, Krivega and Dean, 2012). For a long time, the dominant concept of distance activation of genes by enhancers was that of ‘recruiting’ which was first established to explain transcriptional activation over short distances in prokaryotic organisms (Ptashne, 1986). However, the concept of looping and far-acting interactions between enhancers and promoters could explain long distance transcriptional regulation. **Fig. 1A-B** shows a schematic of how transcription by Pol II could be activated over short (A) and long (B) distances, through recruiting and DNA looping, respectively.

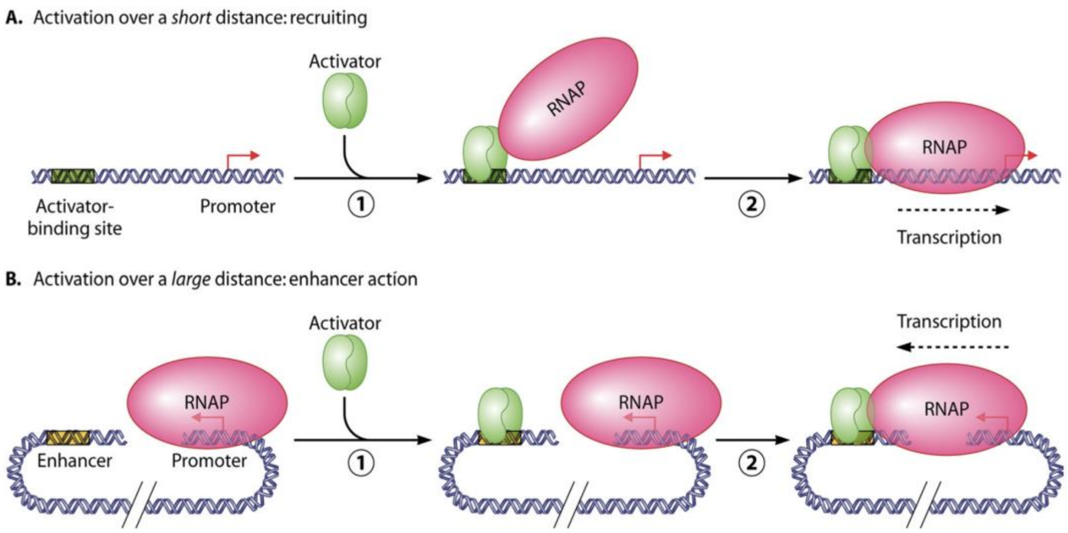


Fig. 1 Transcriptional activation mechanisms (recruiting and looping). **A)** Activation over a short distance. An activator binds to the chromatin upstream of the promoter site. Pol II is recruited by the activator and transcription starts. **B)** Activation over long distances through the action of enhancers. A distantly located enhancer identifies and comes into close proximity with the promoter region (marked by the presence of Pol II) through looping. An activator binds

to the enhancer region and transcription by Pol II begins (Schematic taken from Kriverga and Dean, 2012).

Most ChIA-PET studies so far have focused around the “key” candidates responsible for chromatin interactions, such as Pol II and CTCF. Although there is evidence for 3’ - 5’ end crosstalk (Mapendano et al., 2010, Cavallaro et al., 2021), termination factors like Pcf11 have not been investigated before for potential involvement in chromatin and gene looping. To be able to explore the question of whether 3’ - 5’ end crosstalk is a result of looping, a sequencing technique such as ChIA-PET or Hi-ChIP would be the only way of precisely capturing and quantifying Pcf11-bound loops. Ultimately, it could be the case that Pcf11 is not involved in chromatin or gene looping and that Pcf11’s enrichment at the TSS is the result of other interactions. To the best of my knowledge, this is the first chromatin conformation capture dataset generated to study the relationship between human Pcf11 and gene looping. The potential discovery of Pcf11’s involvement in gene looping would indicate that the transcriptional cycle might be regulated through the formation of such loops, allowing for the interplay between the different complexes governing transcription initiation and termination.

In the following section, I will primarily present results from analyses performed to identify gene loops bound by Pol II and Pcf11 in human cell lines. Publicly available RNA-seq data will also be used to determine if there is any correlation between gene looping and expression levels. Additionally, I will compare and characterise looped genes through GO term analysis.

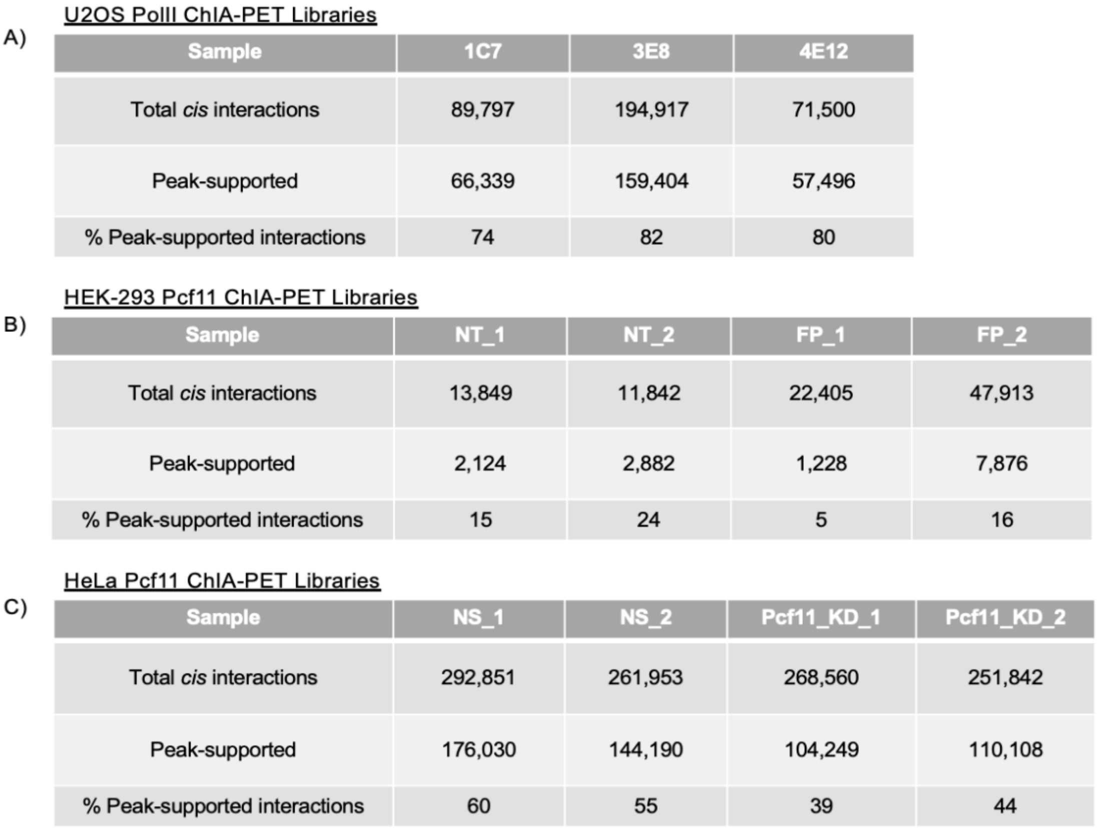
5.2. Results

5.2.1. Chromatin Interactions Bound by Pol II and Pcf11

For simplicity, in the context of ChIA-PET data, we refer to all intrachromosomal interactions as *cis*, regardless of whether they might be several kb away from each other, and all interchromosomal interactions as *trans*. For the purposes of this research, I only considered *cis* interactions and filtered out all *trans* ones. Although it would be interesting to investigate the effect of interchromosomal interactions bound by Pol II or Pcf11, a lot of these contacts are considered to be noise captured from random ligation events during the experimental procedure and are generally excluded for most studies utilising data from ChIA-PET experiments.

As a first step, I looked at raw numbers of peak-annotated *cis* interactions between the samples within the different datasets. **Table 1(A-C)** summarises the number of interactions identified in each sample, as well as how many of those were supported by ChIP peaks. The ChIA-PIPE analysis pipeline annotates interaction coordinates with values/scores of 0, 1, and 2 corresponding to whether there is no ChIP peak (0), peak on just one anchor (1), or peak at both anchors (2). For this work, I considered interactions as Pol II or Pcf11 enriched only if both anchors were peak supported, therefore omitting 0 and 1 scored loops. This was done as a way of ensuring that the downstream analysis of gene looping identification is only making use of interactions where the factor of interest is peak supported at both anchors.

Table 1 Summary of interactions identified in each of the datasets, including proportion of peak-supported contacts. Peak-supported interactions represent those with a value of 2. A) U2OS Pol II libraries and their respective numbers for total interactions. Sample names correspond to antibody clones. **B)** Same as A, but for the HEK-293 Pcf11 dataset. NT refers to not treated cells and FP refers to Flavopiridol-treated samples. **C)** Same as previous, but for HeLa Pcf11 libraries. NS corresponds to non-silencing control siRNA, and KD corresponds to Pcf11 siRNA samples.



A) U2OS PolII ChIA-PET Libraries

Sample	1C7	3E8	4E12
Total *cis* interactions	89,797	194,917	71,500
Peak-supported	66,339	159,404	57,496
% Peak-supported interactions	74	82	80

B) HEK-293 Pcf11 ChIA-PET Libraries

Sample	NT_1	NT_2	FP_1	FP_2
Total *cis* interactions	13,849	11,842	22,405	47,913
Peak-supported	2,124	2,882	1,228	7,876
% Peak-supported interactions	15	24	5	16

C) HeLa Pcf11 ChIA-PET Libraries

Sample	NS_1	NS_2	Pcf11_KD_1	Pcf11_KD_2
Total *cis* interactions	292,851	261,953	268,560	251,842
Peak-supported	176,030	144,190	104,249	110,108
% Peak-supported interactions	60	55	39	44

Overall, the Pol II libraries (**Table 1A**) display lower levels of noise in comparison to the other two datasets since a large percentage (74-82%) of all interactions captured are also supported by significant peaks at both anchors (score of 2). Pcf11 libraries in HEK-293 cells (**Table 1B**) have captured a much lower number of interactions, and only a small percentage of those are supported by peaks. This is not surprising given the quality control metrics presented in Chapter 4 revealing low numbers of unique PETs. Total numbers of interactions positively correlate with the number of unique PETs. As for the

HeLa Pcf11 dataset (**Table 1C**), we observe a large number of interactions for all samples, and as expected, a smaller percentage of the knock-down replicates are supported by peaks. The samples treated with non-silencing siRNA (NS_rep 1 and 2) display a higher proportion of peak-supported loops, although it is worth noting that when using a more stringent peak-calling algorithm (MACS2's narrowPeaks), only ~ 20% and ~ 1% of NS and Pcf11_KD replicates are supported by peaks, respectively. These libraries were also normalised against corresponding Input controls to account for background noise.

Since ChIP-seq data for Pcf11 in HeLa cells are available (Kamieniarz-Gdula et al., 2019), these were also utilised to annotate peaks for the interactions identified in the NS HeLa samples. This was done to allow for further comparisons and to determine whether gene loops identified through ChIA-PET peak annotations and through ChIP-seq would yield different results, something that will be further explored in the following sections. Using the ChIP-seq Pcf11 peak files to annotate chromatin interactions of samples NS_rep1 and NS_rep2, yielded 74,858 and 54,258 peak-supported interactions, respectively.

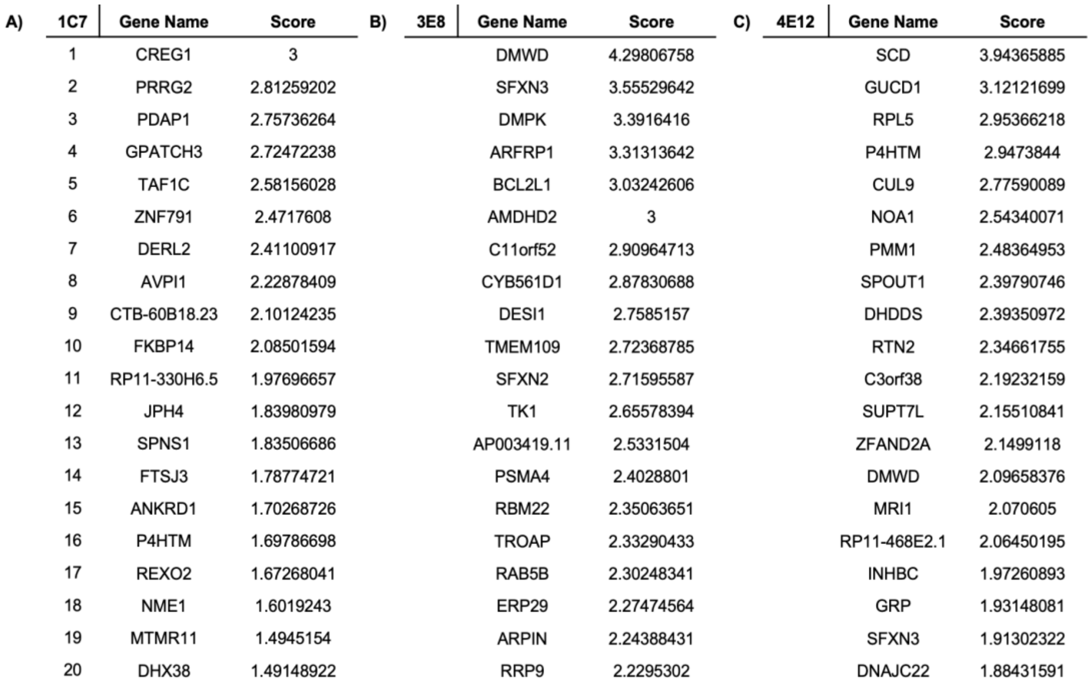
5.2.2. Identifying Looped Genes

Although thousands of peak-supported *cis* interactions were identified in each dataset (**Table 1**), to address the question of whether pcf11 and Pol II are involved in gene looping, I isolated interactions corresponding to contacts ranging from the TSS to the TES of genes. For this purpose, I only considered genes that are protein-coding (n = 19,941) and only used the peak-supported interaction coordinates of each sample. Briefly, I consider a gene to be looped if the span of anchor A and anchor B of each of the peak-supported interaction coordinates overlapped within +/- 500 bp of the TSS and TES of a gene. These are stringent parameters as they only allow for the identification of "very" looped genes that suggest clear contacts between the TSS and TES and exclude "less" looped genes. Even though this approach provides high confidence that the resulting gene list will only contain looped genes, we still assigned a scoring method for each entry. Each gene identified as looped was assigned a score which was calculated based on (i) the overlapping proximity of anchors to the TSS/TES (e.g., < +/- 500 bp away from the TSS/TES would give a higher score), and (ii) how many PETs were mapped to that specific interaction region. The following results sections will be split into two parts; the first will be identifying and characterising looped genes in Pol II U2OS datasets and the second part will be utilising the identical method to do the same in Pcf11 ChIA-PET libraries.

5.2.2.1 Looped Genes Bound by Pol II in U2OS Cells

The three Pol II libraries correspond to U2OS cells, with the only difference between the samples being the Pol II antibody they were immunoprecipitated with. All three antibodies targeted the CTD domain of Pol II, with antibody 1C7 raised against total Pol II, 3E18 raised against Ser5P, and 4E12 raised against Ser7P. As it was shown in Chapter 4, all three libraries showed a high degree of overlap in terms of the identified ChIP peaks. As such, downstream analyses for these samples should be comparable to a degree. The library IDs correspond to the antibody clones used - 1C7, 3E8, and 4E12. After running the analysis for gene looping identification as described in **5.2.2** above, 74 genes were identified to be looped in library 1C7, 315 in library 3E8, and 142 in 4E12. **Table. 2A-C** lists the top 20 protein-coding looped genes in each of the datasets, sorted by score (highest to lowest).

Table 2. Top 20 looped genes in U2OS Pol II ChIA-PET libraries sorted by score (highest to lowest). A) Looped genes in sample 1C7 (total = 74). **B)** Looped genes in sample 3E8 (total = 315). **C)** Looped genes in sample 4E12 (total = 142).



A) 1C7	Gene Name	Score	B) 3E8	Gene Name	Score	C) 4E12	Gene Name	Score
1	CREG1	3		DMWD	4.29806758		SCD	3.94365885
2	PRRG2	2.81259202		SFXN3	3.55529642		GUCD1	3.12121699
3	PDAP1	2.75736264		DMPK	3.3916416		RPL5	2.95366218
4	GPATCH3	2.72472238		ARFRP1	3.31313642		P4HTM	2.9473844
5	TAF1C	2.58156028		BCL2L1	3.03242606		CUL9	2.77590089
6	ZNF791	2.4717608		AMDHD2	3		NOA1	2.54340071
7	DERL2	2.41100917		C11orf52	2.90964713		PMM1	2.48364953
8	AVPI1	2.22878409		CYB561D1	2.87830688		SPOUT1	2.39790746
9	CTB-60B18.23	2.10124235		DESI1	2.7585157		DHDDS	2.39350972
10	FKBP14	2.08501594		TMEM109	2.72368785		RTN2	2.34661755
11	RP11-330H6.5	1.97696657		SFXN2	2.71595587		C3orf38	2.19232159
12	JPH4	1.83980979		TK1	2.65578394		SUPT7L	2.15510841
13	SPNS1	1.83506686		AP003419.11	2.5331504		ZFAND2A	2.1499118
14	FTSJ3	1.78774721		PSMA4	2.4028801		DMWD	2.09658376
15	ANKRD1	1.70268726		RBM22	2.35063651		MRI1	2.070605
16	P4HTM	1.69786698		TROAP	2.33290433		RP11-468E2.1	2.06450195
17	REXO2	1.67268041		RAB5B	2.30248341		INHBC	1.97260893
18	NME1	1.6019243		ERP29	2.27474564		GRP	1.93148081
19	MTMR11	1.4945154		ARPIN	2.24388431		SFXN3	1.91302322
20	DHX38	1.49148922		RRP9	2.2295302		DNAJC22	1.88431591

The Pol II ChIA-PET samples represent libraries constructed from the same cell line, however, the number of total interactions for each file differs. This is possibly due to a combination of factors, such as antibody differences and total starting material, total number of sequenced reads, and biological noise. Although all three antibodies (1C7, 3E8, and 4E12) target the heptad repeat YSPTSPS of Pol II's CTD domain, antibody 3E8 is raised against Ser5P. Ser5 phosphorylation occurs during the start of transcription and early elongation, therefore the antibody is enriching for actively transcribing Pol II proteins found on the promoter sites of genes. Antibody 4E12 detects Ser7P whose phosphorylation pattern is less clear in mammals, but it appears to be phosphorylated throughout the transcribed region. In yeast, Ser7 phosphorylation shows higher peaks during TSS and early elongation (Egloff, 2012; Mayfield et al., 2016). Antibody 1C7 captures total Pol II. When normalised by total peak-supported interactions in each sample, antibody 4E12 appears to be identifying the most looped genes and 1C7 the least.

Nonetheless, all of the identified looped genes from these samples will be considered for further analyses after their statistical significance is confirmed. To do this, an approach of random permutation will be applied.

It is possible that the number of gene loops identified in each library (i.e., total genes whose interactions span +/- 500 bp from their TSS/TES coordinates) are the result of chance when given a large enough interactions file. If this is the case, then we would expect looped genes to not be characterised by specific features or functions. In order to determine the probability of identifying n number of looped genes or more in a dataset with x number of interactions, we performed 10,000 random permutations on the original interactions files. Briefly, we generated 10,000 shuffled versions of the original interactions file, keeping the total number of interactions, number of exonic interactions, their span, and the chromosome on which they originally occurred on consistent. We then process these files in the same way as the original to identify looped genes. **Fig. 2** shows the density plot of the distribution of numbers of looped genes from 10,000 permutations of the 1C7 library, with the red line indicating where the true number lies. The probability of identifying 74 or more looped genes from this distribution is extremely low $(p < 5.2e - 08)$ and therefore confirms that numbers of total gene loops identified from this library are highly statistically significant and unlikely to be the result of chance. This analysis was repeated for the other two libraries, yielding even more extreme probability values with the real number of looped genes lying far outside the right tail of the observed distribution and hence not able to be visualised on the distribution plot.

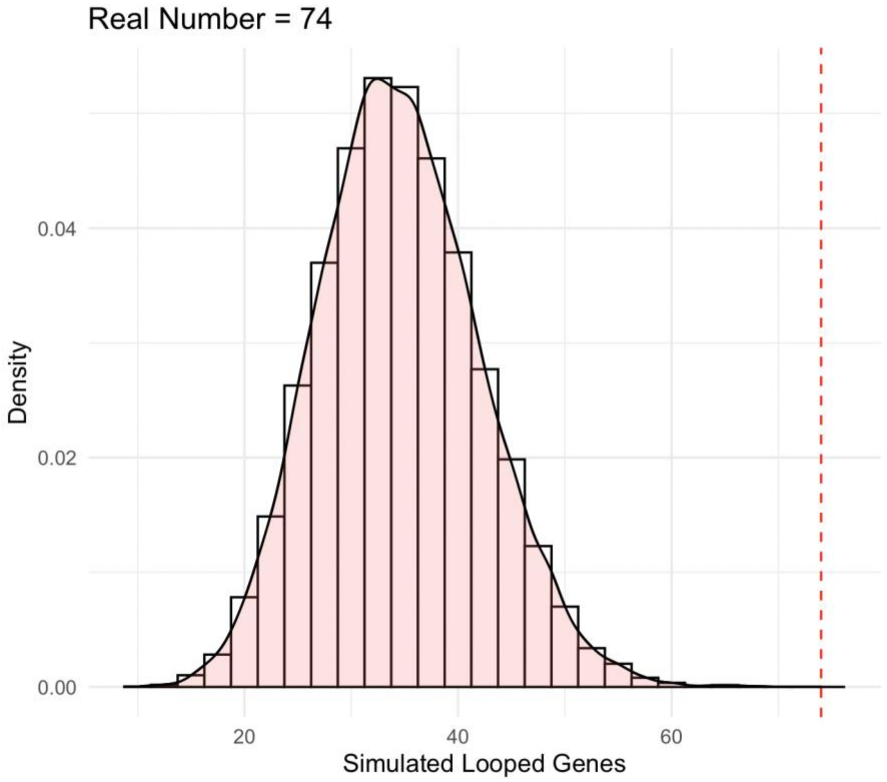


Fig. 2 Density plot of number of looped genes identified from 10,000 simulated shuffles for sample 1C7. In 10,000 shuffles (to allow for a p-value as small as 0.0001), the mean number of looped genes captured in a file of 66,339 unique interaction coordinates is 34.6 (SD = 7.5). Red dashed line represents the real number of looped genes. The probability of identifying 74 or more looped genes from this distribution is extremely unlikely, with $p < 5.2e - 08$.

The full lists of looped genes were used to construct a Venn diagram (**Fig. 3**) of the common looped genes. 8 genes appear to be common among all three samples, 31 genes among library 4E12 and 3E8, 14 genes among 1C7 and 3E8, and 6 genes among 4E12 and IC7.

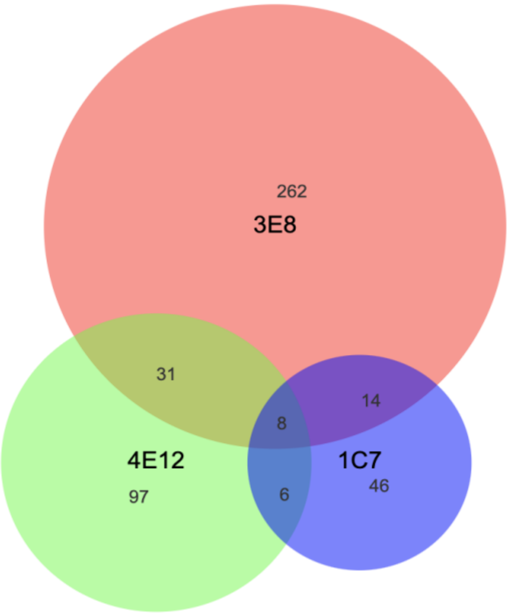


Fig. 3 Venn Diagram of looped genes common among Pol II ChIA-PET libraries, $(\boldsymbol{p} < \mathbf{0.0001})$. A total of 8 common genes were identified in all three samples. Some of which include SRRD, ODC1, TIGD4, P4HTM and a number of MT genes. P-value was calculated based on a 10,000-trial simulation. Venn diagram produced with BioVenn (Hulsen et al., 2008).

Although the numbers of looped genes in each sample vary between 74 and 315, this is attributed to the stringent, binary threshold we have chosen to apply in determining what constitutes a looped gene. Genes that are “less” looped than others will not be part of the output and therefore will be counted as not looped and omitted from the list. As such, even minor differences in the interaction coordinates between these files would alter the contents of the output list. Had we allowed for a spectrum of “loopness” to exist, this would have resulted in all genes being assigned a looping score, with higher scores signifying very looped genes and lower scores signifying less looped ones. In the present case where the parameters have strong cutoffs, a gene that is less looped in one of the samples due to its anchors being, for example, 510 bp away from the TSS/TES instead of the 500 bp cutoff point, will be excluded from the list entirely. When I repeated the analysis changing the anchor cutoff to +/- 1kb from the TSS/TES, more genes were identified in each dataset, as expected. For the purposes of this work, I am only interested in genes that show high confidence in 3' - 5' end crosstalk and will only consider genes identified to be looped using the 500 bp threshold for further analyses.

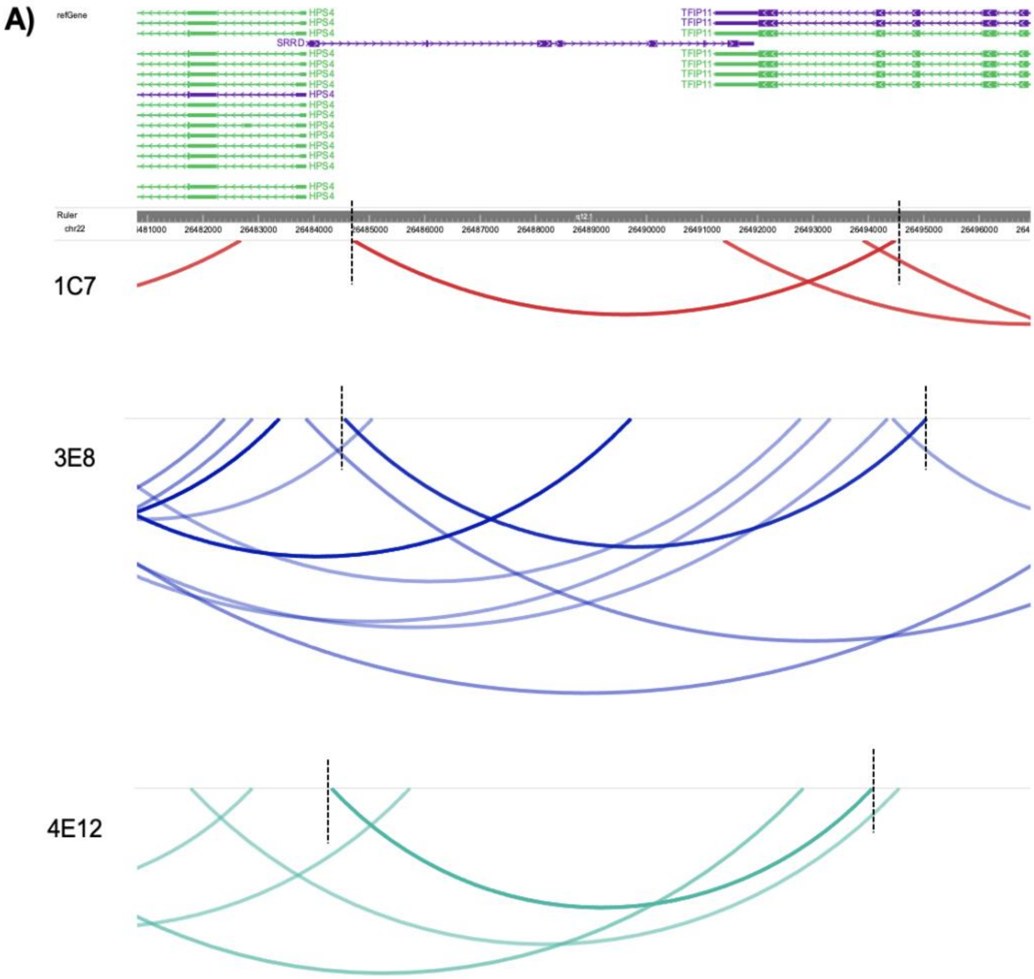
However, it was still important to determine whether the observed overlap between the samples is statistically significant or an event expected to occur by chance. To examine this, I performed an *in-silico* simulation to derive a probability value, similar to a hypergeometric test.

To simulate this event, I defined the following:

$$N = number\ of\ genes\ for\ selection$$
$$\alpha = number\ of\ looped\ genes\ in\ set\ 1$$
$$\beta = number\ of\ looped\ genes\ in\ set\ 2$$
$$\gamma = number\ of\ looped\ genes\ in\ set\ 3$$
$$\kappa = number\ of\ common\ genes\ in\ all\ sets$$

I then simulated this situation 10,000 times to derive a P value for the probability of $\kappa = 8$ occurring, when $N = 19{,}941$, $\alpha = 315$, $\beta = 142$, and $\gamma = 74$. Indeed, the likelihood of observing 8 or more overlapping genes when all the previous parameters held true returned a significance value of $p << 0.0001$. To test whether the large sample size $(N = 19{,}941)$ was overpowering the result, I repeated the simulation by assigning a much smaller pool of genes for selection $(N = 1000)$, which returned a P value of $p = 0.0154$. These tests confirm that the overlap between the three samples is highly statistically significant and extremely unlikely to have occurred by chance.

To further confirm that these interactions truly span between the TSS to the TES of genes and that our method for gene loop identification is performing as intended, I have produced plots based on the original interaction files and the looped genes identified after the analysis. **Fig. 4A-B** shows genes SRRD and ODC1, common looped genes in all three sets, along with the corresponding interaction profile for each of the three samples in the region.



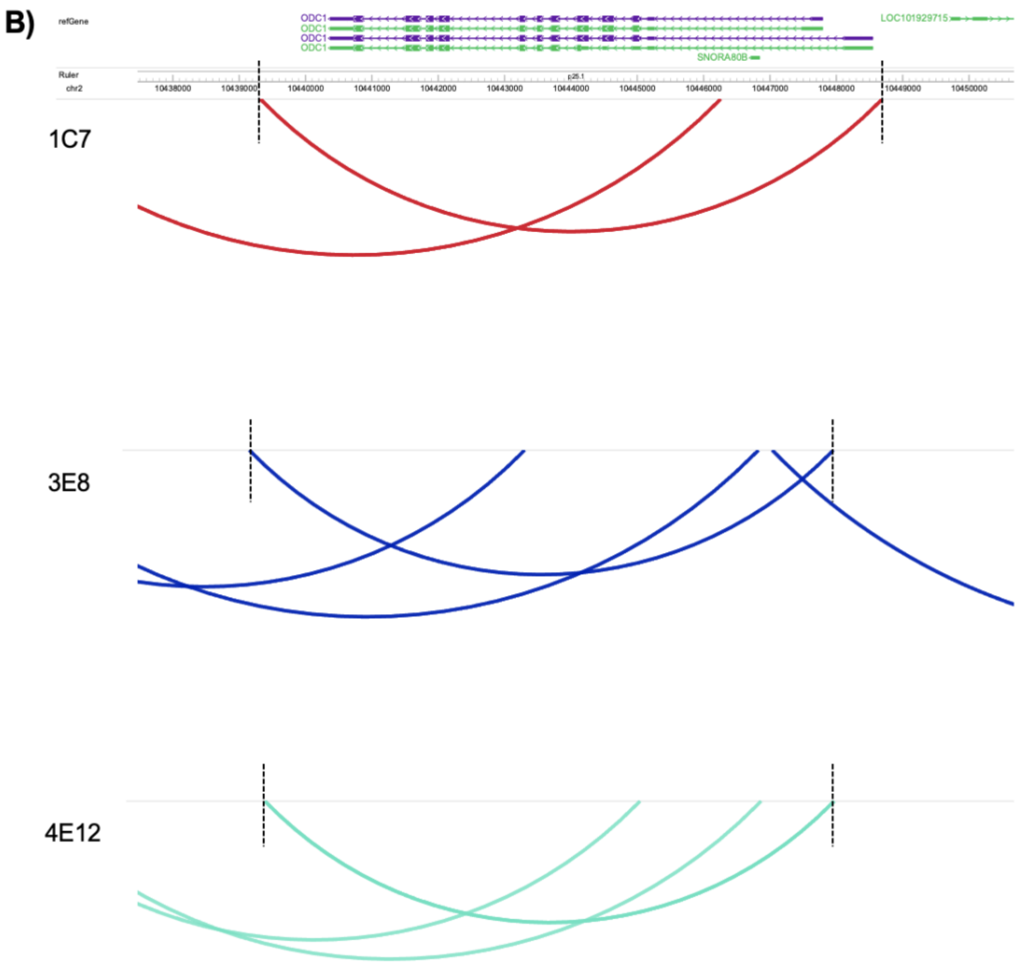


Fig. 4 SRRD and OCD1 genes are looped in all three Pol II libraries. A) A snapshot from the Epigenome WashU Genome Browser indicating the interactions in the SRRD gene, chr22:26,483,867-26,494,658 (GRCh38/hg38). SRRD is a protein coding gene involved in the regulation of expression of core clock genes and heme biosynthesis (NCBI, 2022). We observe a loop in all three panels (1C7, 3E8, 4E12) which falls within the cutoff point of +/- 500 bp from the TSS/TES. As expected, sample 3E8 shows a higher number of interactions in the region as the sample with the highest number of total peak-supported contacts. Vertical dotted lines show the starts and ends of the anchors. **B)** Similar to A), but for gene ODC1, chr2:10,439,968-10,448,327 (GRCh38/hg38). ODC1 codes for the rate-limiting enzyme of the polyamine biosynthesis pathway (NCBI, 2022) and appears to overlap with SNORA80B (chr2:10,446,714-10,446,849).

To determine whether there is a relationship between looped genes and particular biological processes, and to further examine whether the phenomenon of looping is a random event, I performed Gene Ontology (GO) analysis using all unique gene names between the three samples $(n = 465)$. Although a number of the identified looped genes are uncharacterised, GO analysis (**Fig. 5** and **Table 3**) reveals that characterised looped genes $(n = 333)$, appear to be associated with key cellular and metabolic processes $(p < 0.05)$.

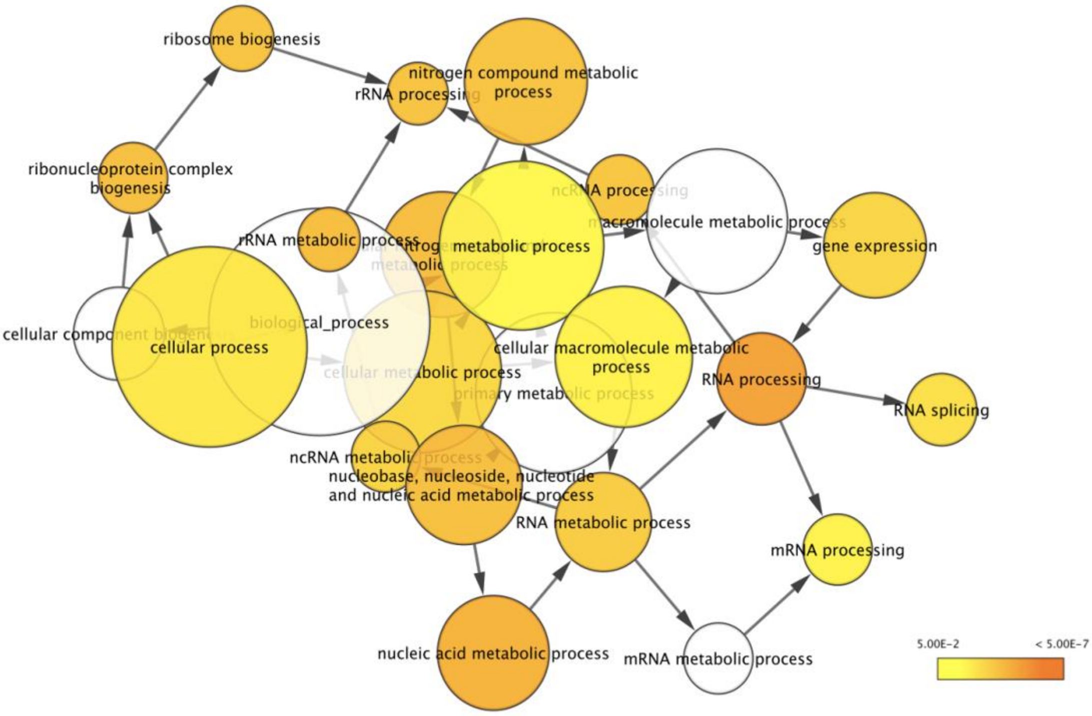
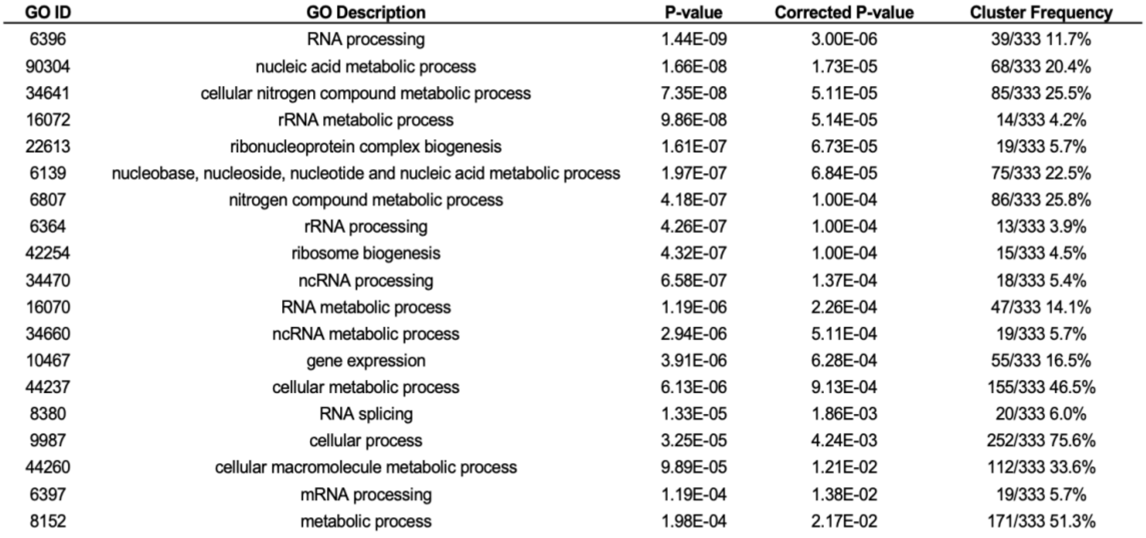

ribosome biogenesis
nitrogen compound metabolic process
rRNA processing
ribonucleoprotein complex biogenesis
ncRNA processing
macromolecule metabolic process
rRNA metabolic process
nitrogen compound metabolic process
metabolic process
gene expression
biological_process
cellular component biogenesis
cellular process
cellular macromolecule metabolic process
cellular metabolic process
primary metabolic process
RNA processing
RNA splicing
ncRNA metabolic process
nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
RNA metabolic process
mRNA processing
nucleic acid metabolic process
mRNA metabolic process
5.00E-2
< 5.00E-7

Fig. 5 Gene Ontology Analysis (GO) of biological processes for looped genes identified in Pol II ChIA-PET samples. Looped genes from all three Pol II datasets are mainly associated with fundamental metabolic and cellular processes. These include RNA processing, metabolic processes involving nucleic acids and rRNA, as well as ribonucleoprotein complex biogenesis. Gene expression, mRNA processing, and RNA splicing are also significantly associated with the identified looped genes. Scale bar corresponds to $p - value$ (diagram produced using BiNGO on Cytoscape v3.9.1).

Table 3 Top GO terms for looped genes in Pol II ChIA-PET samples sorted by FDR-adjusted P-value $(p < 0.05)$. Results generated with BiNGO on Cytoscape v3.9.1.



GO ID	GO Description	P-value	Corrected P-value	Cluster Frequency
6396	RNA processing	1.44E-09	3.00E-06	39/333 11.7%
90304	nucleic acid metabolic process	1.66E-08	1.73E-05	68/333 20.4%
34641	cellular nitrogen compound metabolic process	7.35E-08	5.11E-05	85/333 25.5%
16072	rRNA metabolic process	9.86E-08	5.14E-05	14/333 4.2%
22613	ribonucleoprotein complex biogenesis	1.61E-07	6.73E-05	19/333 5.7%
6139	nucleobase, nucleoside, nucleotide and nucleic acid metabolic process	1.97E-07	6.84E-05	75/333 22.5%
6807	nitrogen compound metabolic process	4.18E-07	1.00E-04	86/333 25.8%
6364	rRNA processing	4.26E-07	1.00E-04	13/333 3.9%
42254	ribosome biogenesis	4.32E-07	1.00E-04	15/333 4.5%
34470	ncRNA processing	6.58E-07	1.37E-04	18/333 5.4%
16070	RNA metabolic process	1.19E-06	2.26E-04	47/333 14.1%
34660	ncRNA metabolic process	2.94E-06	5.11E-04	19/333 5.7%
10467	gene expression	3.91E-06	6.28E-04	55/333 16.5%
44237	cellular metabolic process	6.13E-06	9.13E-04	155/333 46.5%
8380	RNA splicing	1.33E-05	1.86E-03	20/333 6.0%
9987	cellular process	3.25E-05	4.24E-03	252/333 75.6%
44260	cellular macromolecule metabolic process	9.89E-05	1.21E-02	112/333 33.6%
6397	mRNA processing	1.19E-04	1.38E-02	19/333 5.7%
8152	metabolic process	1.98E-04	2.17E-02	171/333 51.3%

The GO analysis results reveal that looped genes appear to be associated with fundamental cellular processes, as shown in **Table 3** above. Housekeeping processes, like ribosome biogenesis, cellular metabolic processes, and mRNA processing are essential for the basic functionality and maintenance of all cell types. Based on the complete list of looped genes from all three Pol II ChIA-PET samples, I observe that there are clusters of genes of the same family. This further adds onto the assumption that looping is not a random occurrence and that there are aspects to this phenomenon. For example, ABHD1, ABHD4, and ABHD10 are all looped genes. Mammalian α/β hydrolase domain (ABHD) proteins are found in every reported genome and are involved in metabolic pathways and signal transduction (Lord et al., 2013), although more recent research suggests that they might also be associated with different types of disease (Xu et al., 2018). A number of Zinc-finger proteins, part of the ZNF superfamily, are also looped. These include ZNF35, ZNF260, ZNF384, ZNF397, ZNF488, ZNF526, ZNF740, and ZNF791. ZNFs are known to be involved in several cellular processes, such as gene expression, and do possess key roles in differentiation and development, along with being associated with cancer progression and tumorigenesis (Cassandri et al., 2017). Another example of looped genes identified in the present datasets are those of the transmembrane (TMEM) family. 8 TMEM genes appear to be looped, including TMEM43 which has been shown to have important roles in maintaining the structure of the nuclear envelope through protein complex organisation (Bengtsson and Otto, 2008).

As looped genes appear to be involved in housekeeping processes, gene looping could possibly be a cellular mechanism that enables efficient and rapid generation of transcripts, a concept that is also supported by work from our lab (Cavallaro et al., 2021). To further examine this and to determine whether looping correlates with mRNA expression levels, I analysed bulk RNA-seq data from U2OS cells (Accession: GSE162163). Indeed, looped genes appear to be more highly expressed than the rest of protein-coding genes (**Fig. 6A**). A Wilcoxon signed-rank test returns a significance value of $p < 2.2e - 16$, confirming that there is a large, statistically significant difference in transcript generation between looped and non-looped genes. The median expression for looped genes was 10,334 TPM, whereas for non-looped genes it was 147 TPM, a ~ 70-fold difference. To further investigate features of looped genes beyond GO analysis and expression levels, I looked at gene length. Looped genes appear to be shorter than non-looped genes (**Fig. 6B**) (Wilcoxon signed-rank significance value of $p < 2.2e - 16$). The length of looped genes is less than twice that of non-looped ones, with 12,366 bp and 28,428 bp median length spans, respectively.

Although we showed that looped genes are shorter than non-looped ones, and there is a positive correlation between gene size and UTR length (**Fig. 7A-B**), I still explored whether there are any differences in 3’ and 5’ UTR lengths between looped and non-looped genes. I considered a single 3’ and 5’ UTR for each gene, corresponding to the full-length transcript. Hence, UTR lengths of alternative polyadenylated transcripts and splice variants were not considered for this comparison. Looped genes appear to have shorter 5’ UTRs than non-looped genes ($p < 0.02118$). The median 5’ UTR length for looped genes is 114 bp and for non-looped genes it is 134 bp. However, there is no significant difference between 3' UTR sizes ($p < 0.6688$). The median 3' UTR is 1,096 bp for looped genes, and 1,020 bp for non-looped genes.

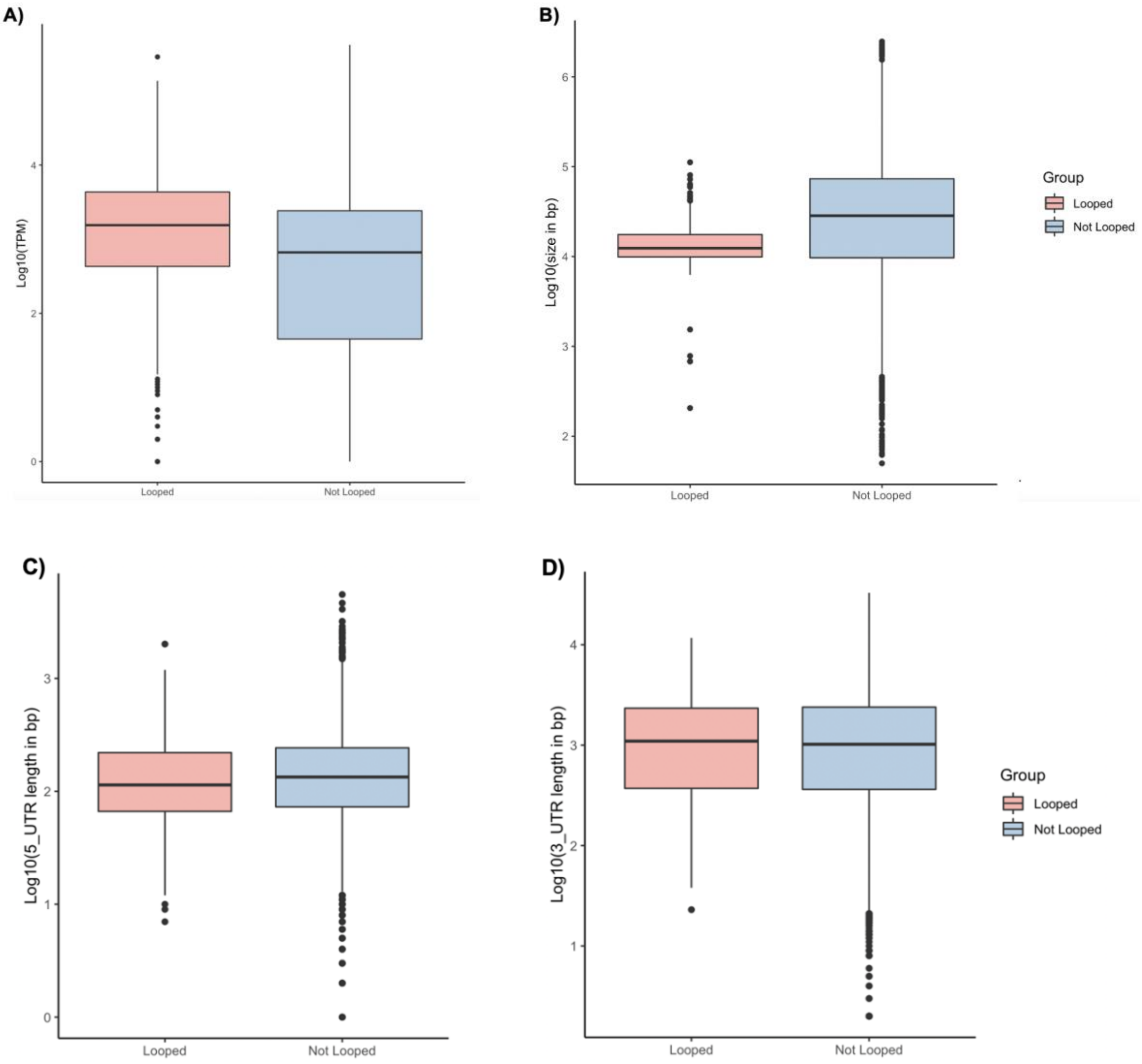


Fig. 6 Boxplot of Log-transformed mRNA expression levels in transcripts per million (TPM), gene, and UTR lengths for looped and not looped genes in U2OS cells. A) Protein-coding looped genes (pink) identified from Pol II ChIA-PET libraries are on average more highly expressed than the rest of the protein-coding genes (blue), according to bulk RNA-seq data from the same cell line. Wilcoxon signed-rank test $p < 2.2e - 16$. **B)** On average, looped genes (pink) are shorter than non-looped genes, $p < 2.2e - 16$ (RNA-seq data Accession: GSE162163). **C)** Looped genes (pink) appear to have shorted 5' UTRs compared to non-looped genes (blue). Wilcox signed-rank test $p = 0.02118$. **D)** There is no significant difference in 3' UTR size between looped (pink) and non-looped genes (blue). Wilcox signed-rank test $p = 0.6688$.

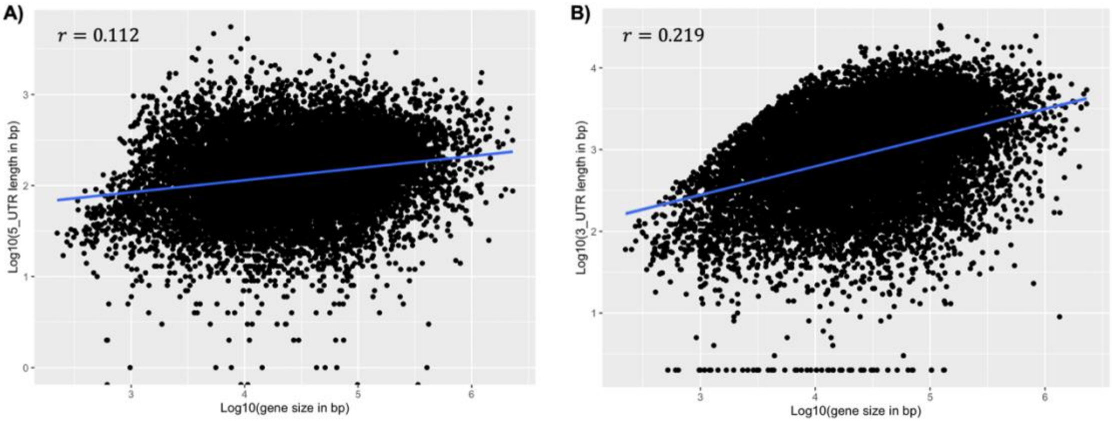


Fig. 7 Correlation between gene size and UTR length. A) Positive correlation of 5' UTR lengths and gene size. Pearson's $r = 0.112$. **B)** Positive correlation of 3' UTR lengths and gene size. Pearson's $r = 0.219$.

These analyses reveal that gene looping is not only a non-random phenomenon, but also one that appears to have specific characteristics associated with it. These results are in line with those of Grosso et al. (2012) where they utilised ChIP-seq data and observed similar features associated with genes with Pol II Ser5P enrichment on their 3' end, which was suggestive of gene looping. The present work provides further evidence that principles of gene looping might be conserved across human cell lines and that, perhaps, they represent an evolutionary mechanism for transcriptional regulation.

5.2.2.2 Looped Genes Bound by Pcf11

Applying the same methodology as with the Pol II ChIA-PET libraries, I also investigated whether Pcf11 is involved in gene looping. Although the Pol II samples were generated on a different cell line, it would still be interesting to determine whether Pcf11-bound gene loops are similar to those of Pol II and whether looped genes are consistent between cell lines. The metagene profile of Pcf11 showing significant occupancy at the TSS raises the question of whether the enrichment is a result of looping. However, for this to really be the case, a significant number of genes would need to be looped, as well as enriched for Pcf11.

Although the occupancy of Pcf11 at the TSS of genes is supported by both my own Pcf11 ChIA-PET data, as well as data from Pcf11 ChIP-seq samples generated by Kamieniarz-Gdula et al., (2019) and it would be very unlikely for the peak to be a technical error, I wanted to confirm this profile by knocking-down Pcf11. **Fig. 8** shows a metagene plot around the TSS region for Pcf11 ChIA-PET samples with Pcf11 siRNA and non-silencing (NS) siRNA for control. We observe a reduction of Pcf11 occupancy close to the promoter region of genes upon its depletion, and therefore confirm that the factor is indeed enriched at the start sites of genes.

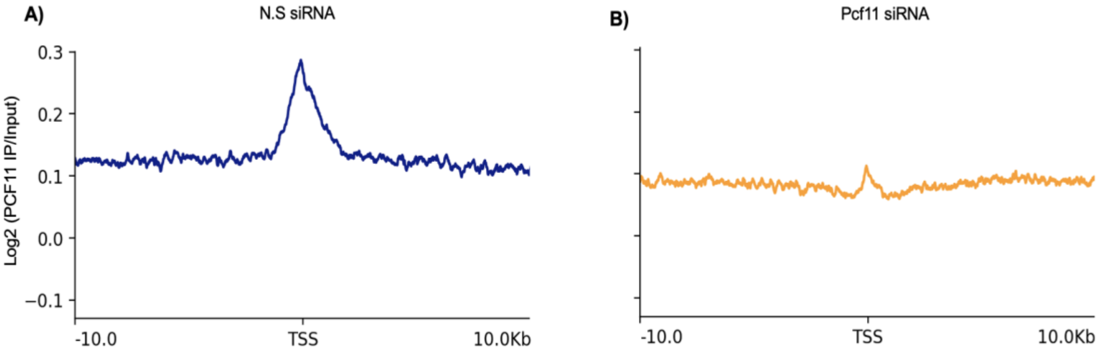


Fig. 8 Coverage of Pcf11 ChIA-PET samples across the TSS region. A) Metagene plot of non-silencing siRNA sample after Pcf11 immunoprecipitation. A peak at the TSS can be observed, indicating the localisation of Pcf11 close to the promoter site of genes. **B)** Same as A), but for Pcf11 siRNA sample, indicating reduction of Pcf11 occupancy at the TSS.

The Pcf11 HEK-293 libraries were not suitable for gene looping analysis due to low numbers of unique PETs. Apart from sample FP_rep2, all other samples generated only ~1,200-2,900 peak-supported interactions which is a very small number to return a list of looped genes. For this reason, the Pcf11 HEK-293 dataset will not be used in the gene looping analysis.

As discussed in Chapter 4, since published ChIP-seq data for Pcf11 in HeLa cells are available, these will also be used to annotate interactions and compare results. Using the peak annotations from the ChIA-PET data, sample NS_rep1 returns 579 looped genes, and sample NS_rep2 240. With the ChIP-seq peak annotations, sample NS_rep1 returned 456 genes, and sample NS_rep2 231. The lower number of identified Pcf11-enriched gene loops in the second replicate could be attributed to the ChIP reaction being weaker resulting in more unspecific pull-down of DNA. After normalising for total peak-supported interactions captured, Pcf11 samples appear to be enriched in ~ 13% more looped genes than Pol II. Although this could have biological significance, it could also be due to cell line differences. A direct comparison would have been more favourable if it was between Pol II ChIA-PET samples from HeLa cells.

The total number of looped genes identified using the peaks from the ChIP-seq and ChIA-PET data indicates a close overlap. **Fig. 9A-B** shows Venn diagrams of the overlap between looped genes identified with ChIP and ChIA-PET annotations for the two Pcf11 replicates. For further analyses, I considered gene loops to be enriched for Pcf11 if the gene was identified as looped using both peak annotation methods. Unique looped genes from both replicates were considered as Pcf11-bound looped genes ($n = 606$). **Fig. 9C** shows the overlap between the replicates after the above consideration was applied. Similar to the Pol II analysis, the significance of the overlap between the two replicates is highly significant ($p < 0.0001$). **Table 4** shows the top 20 looped genes in the two replicates.

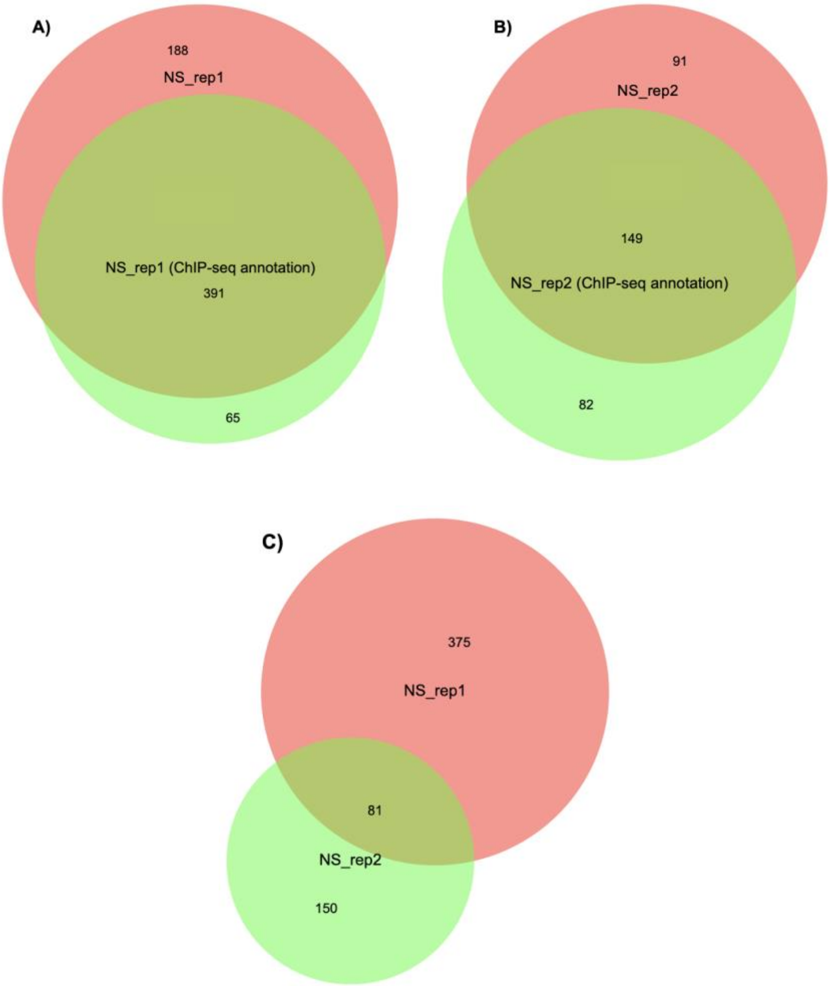
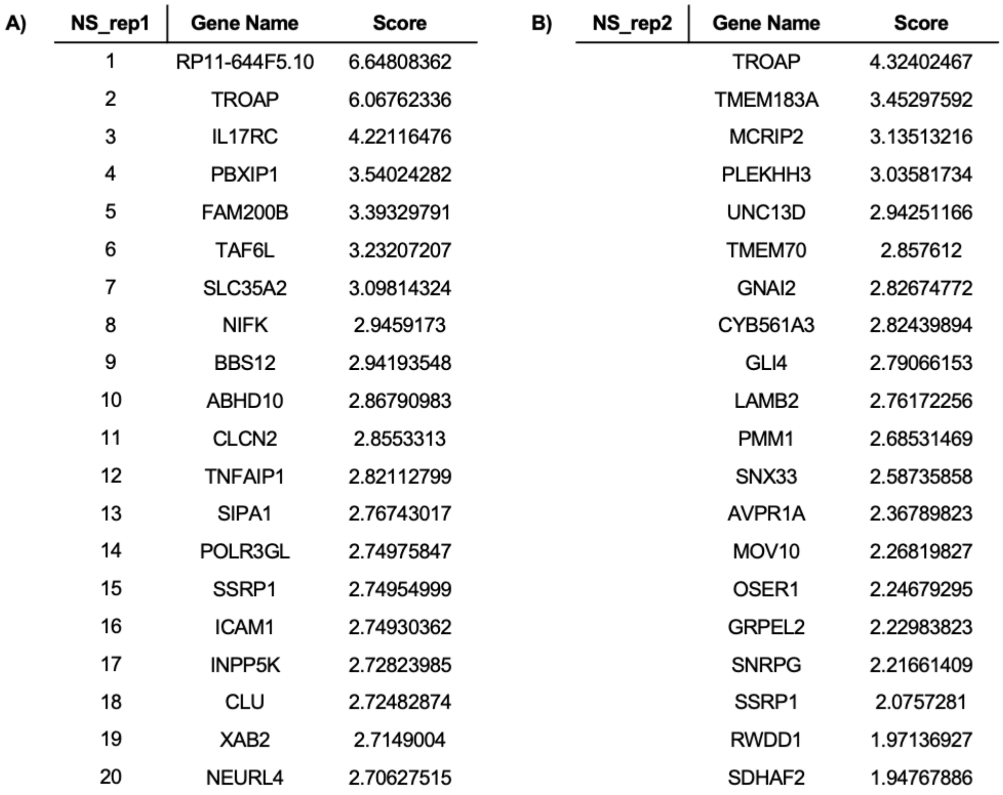


Fig. 9 Venn Diagrams of Pcf11 ChIA-PET replicates. A-B) Venn diagrams of replicates indicating the overlap between peak annotations derived from both ChIA-PET and ChIP-seq experiments. **C)** Overlap between the two replicates after only considering genes to be looped if the gene overlapped between the ChIP and ChIA-PET annotations for each replicate independently ($p < 0.0001$).

Table 4 List of top 20 looped genes in the two Pcf11 ChIA-PET replicates. The list of genes for each replicate consists of looped genes found to overlap between the two peak-annotation methods (ChIP-seq and ChIA-PET).



A)

NS_rep1	Gene Name	Score
1	RP11-644F5.10	6.64808362
2	TROAP	6.06762336
3	IL17RC	4.22116476
4	PBXIP1	3.54024282
5	FAM200B	3.39329791
6	TAF6L	3.23207207
7	SLC35A2	3.09814324
8	NIFK	2.9459173
9	BBS12	2.94193548
10	ABHD10	2.86790983
11	CLCN2	2.8553313
12	TNFAIP1	2.82112799
13	SIPA1	2.76743017
14	POLR3GL	2.74975847
15	SSRP1	2.74954999
16	ICAM1	2.74930362
17	INPP5K	2.72823985
18	CLU	2.72482874
19	XAB2	2.7149004
20	NEURL4	2.70627515

B)

NS_rep2	Gene Name	Score
	TROAP	4.32402467
	TMEM183A	3.45297592
	MCRIP2	3.13513216
	PLEKHH3	3.03581734
	UNC13D	2.94251166
	TMEM70	2.857612
	GNAI2	2.82674772
	CYB561A3	2.82439894
	GLI4	2.79066153
	LAMB2	2.76172256
	PMM1	2.68531469
	SNX33	2.58735858
	AVPR1A	2.36789823
	MOV10	2.26819827
	OSER1	2.24679295
	GRPEL2	2.22983823
	SNRPG	2.21661409
	SSRP1	2.0757281
	RWDD1	1.97136927
	SDHAF2	1.94767886

Similar to Pol II gene loops, GO analysis was performed on the identified looped genes from both Pcf11 replicates. **Fig. 10A-B** shows examples of looped genes, and **Fig. 11** the GO analysis diagram highlighting the most significantly enriched GO terms associated with biological processes. **Table 5** lists the most enriched GO terms, sorted by FDR-adjusted P-value. Genes with Pcf11-bound loops appear to also be involved in key cellular processes, such as RNA processing and metabolism. Although it has been suggested that looping is cell-type dependent and that it is not an intrinsic function of a specific subset of genes (Grosso et al., 2012), in both Pol II U2OS and Pcf11 HeLa datasets, it appears that the biological function of looped genes serves similar purposes. Although the comparison is between datasets generated on different cell lines, ~14% of looped genes overlapped between the two datasets.

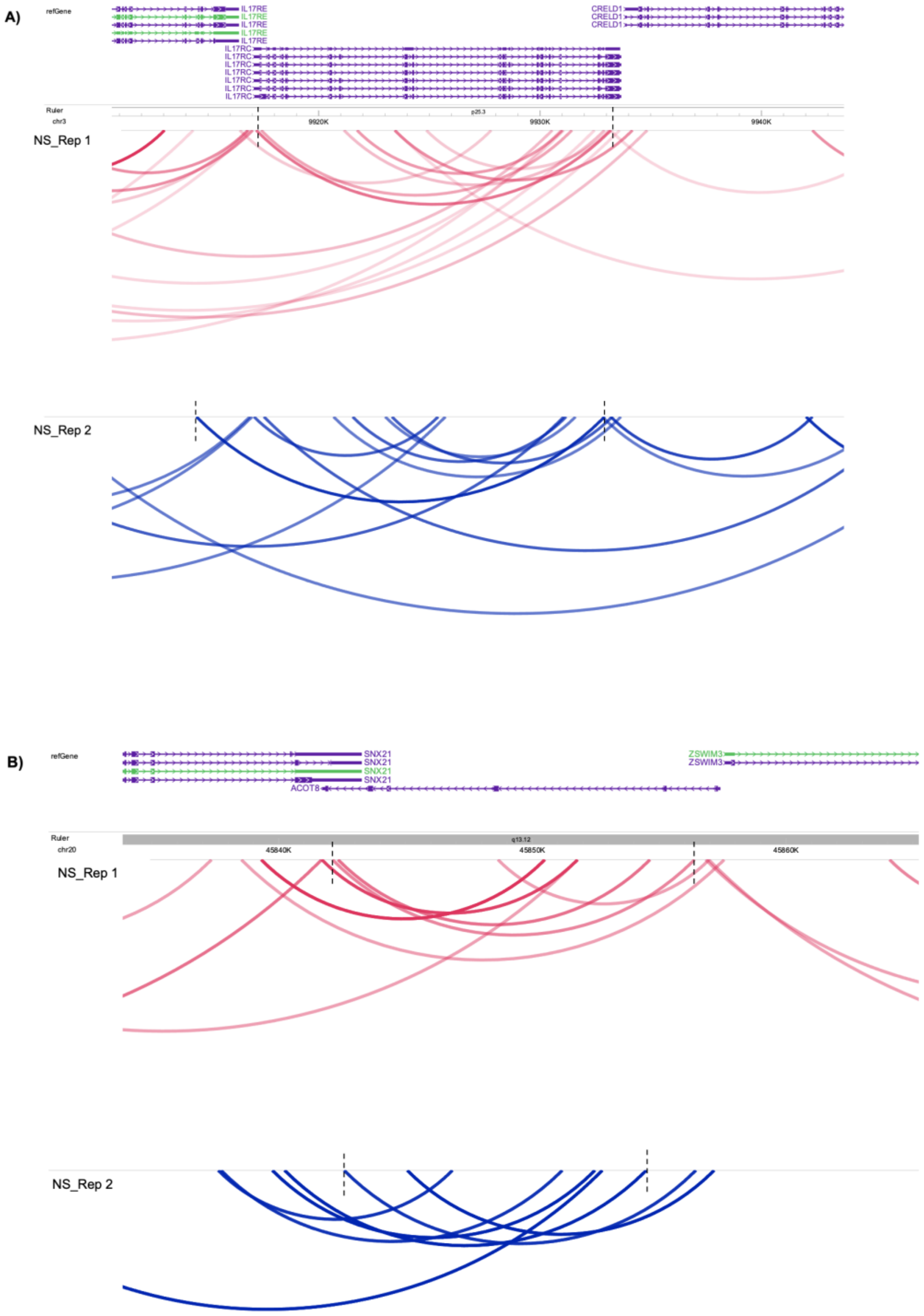


Fig. 10 IL17RC and ACOT8 are looped in Pcf11 ChIA-PET replicates. A) A snapshot from the Epigenome WashU Genome Browser indicating the interactions at IL17RC, chr3:9,917,074-9,933,630 (GRCh38/hg38). IL17RC encodes for a protein involved in the immune response. We observe a loop in both Pcf11 replicates within the cutoff point of +/- 500 bp from TSS/TES. Vertical dotted lines show the starts and ends of the anchors. **B)**

Similar to A), but for gene ACOT8, chr20:45,841,721-45,857,405 (GRCh38/hg38). ACOT8 codes for a protein involved in the oxidation of fatty acids.

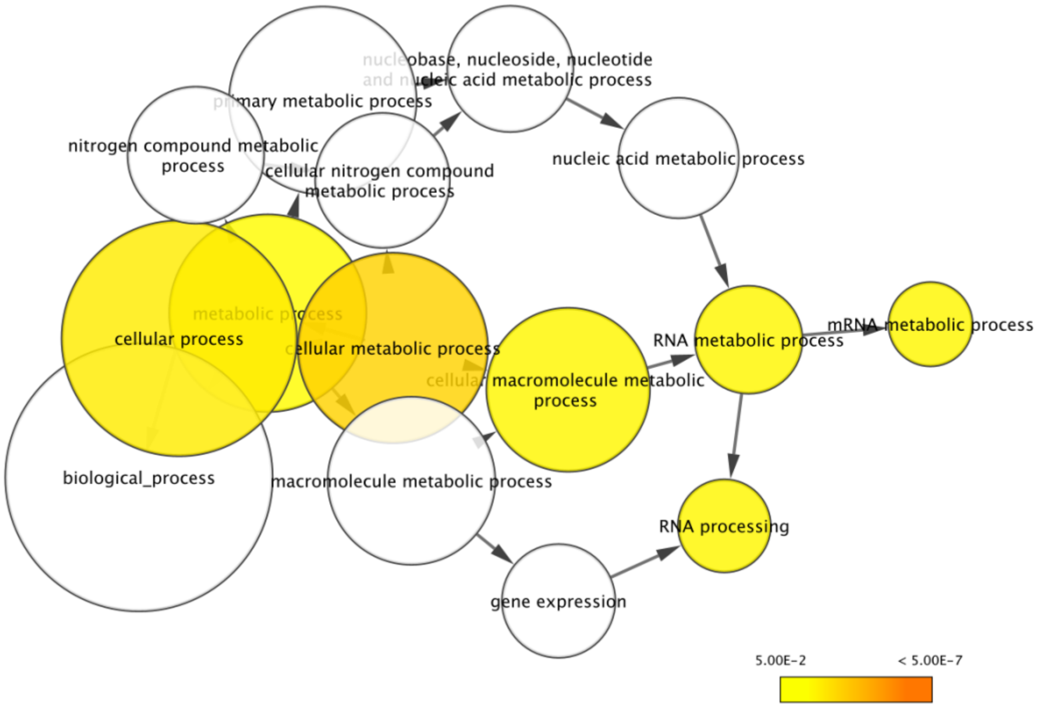
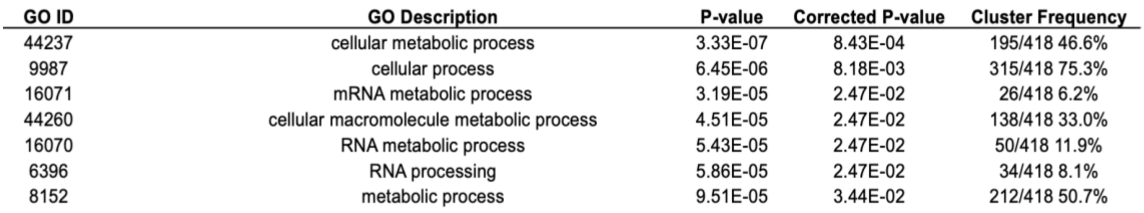

nucleobase, nucleoside, nucleotide and nucleic acid metabolic process
primary metabolic process
nitrogen compound metabolic process
cellular nitrogen compound metabolic process
nucleic acid metabolic process
metabolic process
cellular process
cellular metabolic process
cellular macromolecule metabolic process
RNA metabolic process
mRNA metabolic process
biological_process
macromolecule metabolic process
RNA processing
gene expression
5.00E−2
< 5.00E−7

Fig. 11 Gene Ontology Analysis (GO) of biological processes for looped genes identified in Pcf11 ChIA-PET samples. Looped genes from Pcf11 replicates are mainly associated with fundamental metabolic and cellular processes, similar to those of Pol II-bound looped genes in U2OS cells. Scale bar corresponds to $p - value$ (diagram produced using BiNGO on Cytoscape v3.9.1).

Table 5 Top GO terms for looped genes in Pcf11 ChIA-PET sorted by FDR-adjusted P-value $(p < 0.05)$. Results generated with BiNGO on Cytoscape v3.9.1.



GO ID	GO Description	P-value	Corrected P-value	Cluster Frequency
44237	cellular metabolic process	3.33E-07	8.43E-04	195/418 46.6%
9987	cellular process	6.45E-06	8.18E-03	315/418 75.3%
16071	mRNA metabolic process	3.19E-05	2.47E-02	26/418 6.2%
44260	cellular macromolecule metabolic process	4.51E-05	2.47E-02	138/418 33.0%
16070	RNA metabolic process	5.43E-05	2.47E-02	50/418 11.9%
6396	RNA processing	5.86E-05	2.47E-02	34/418 8.1%
8152	metabolic process	9.51E-05	3.44E-02	212/418 50.7%

To investigate whether Pcf11-bound looped genes in HeLa cells display similar features as with those found with Pol II in U2OS cells, I repeated the analysis of gene expression using HeLa bulk RNA-seq data. Looped genes appeared to be more expressed than non-looped genes, as well as shorter in size, which mirrors the features identified for looped genes in the Pol II libraries (**Fig. 12A-B**). When looking at 5’ and 3’ UTR lengths, similarly to Pol II-bound gene loops, we observe that looped genes have shorter 5’ UTRs (**Fig. 12C**). The median 5’ UTR length of looped genes is 124 bp, and 134 for non-looped genes ($p = 0.03075$). In contrast to Pol II-bound gene loops, Pcf11 loops show a significant difference in 3’ UTR length when compared to non-looped genes **(Fig. 12D)**. The median 3’ UTR size of Pcf11 looped genes is 782 bp and 1,032 for non-looped ones ($p = 9.874e - 06$). The 3’ UTR length of Pcf11 looped genes is also significantly shorter than that of Pol II looped genes ($p = 0.000953$), potentially suggesting that Pcf11 is more abundant at the 3’ ends of genes with shorter 3’ UTRs.

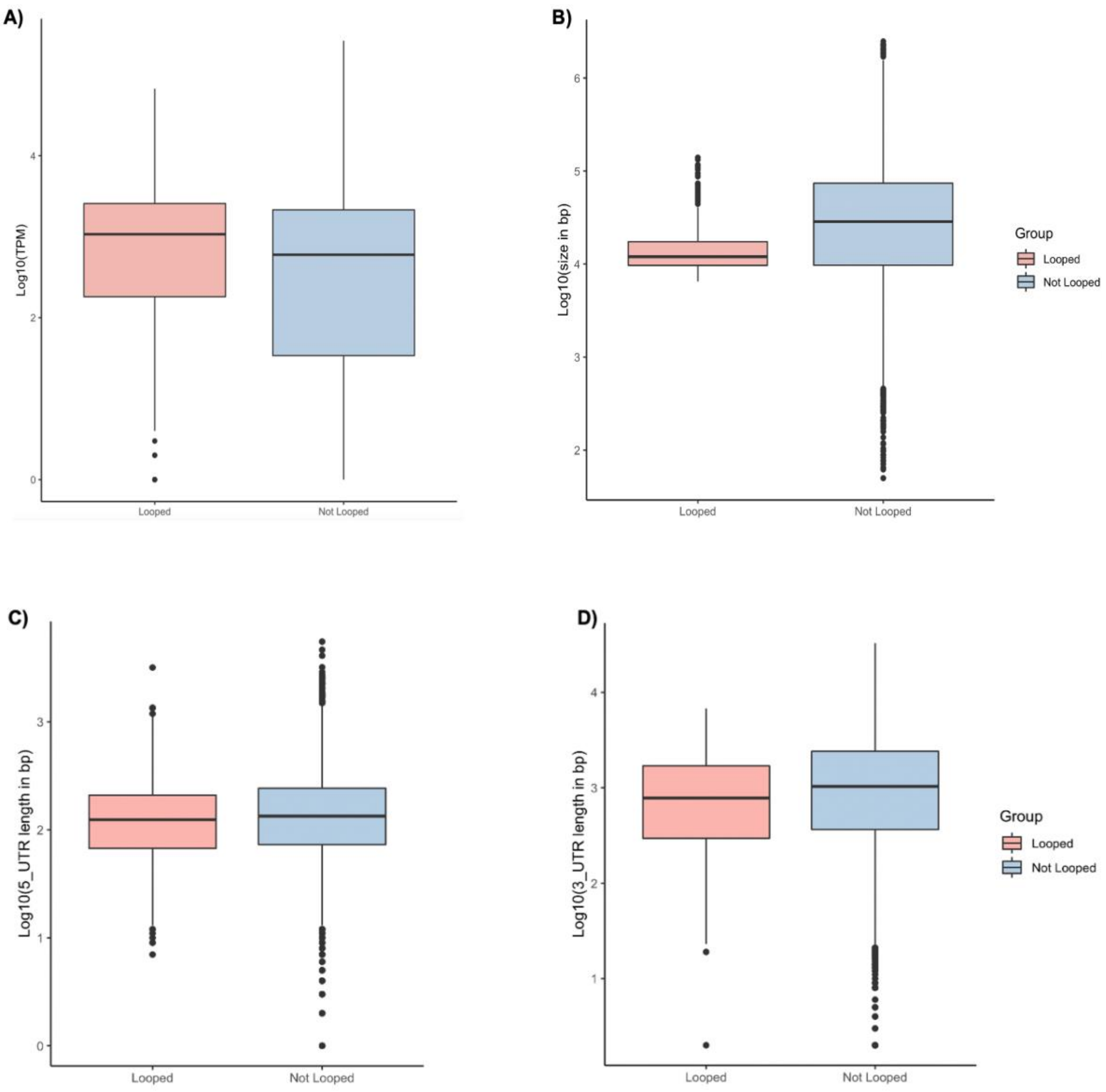


Fig. 12 Boxplot of Log-transformed mRNA expression levels in transcripts per million (TPM) for looped and not looped genes in HeLa cells. A) Protein-coding looped genes (pink) identified from Pcf11 ChIA-PET libraries are on average more highly expressed than the rest of the protein-coding genes (blue), according to bulk RNA-seq data from the same cell line. Wilcoxon signed-rank test $p < 2.2e - 16$. **B)** On average, looped genes (pink) are shorter than non-looped genes, $p < 2.2e - 16$ (RNA-seq data ENA Accession: PRJNA245463). **C)** Looped genes (pink) appear to have shorted 5' UTRs compared to non-looped genes (blue). Wilcox signed-rank test $p = 0.03075$. **D)** Looped genes (pink) have significantly shorter 3' UTR lengths than non-looped genes. Wilcox signed-rank test ($p = 9.874e - 06$).

It appears that looped genes are generally more expressed than non-looped ones, regardless of cell line (HeLa or U2OS). The fact that Pol II and Pcf11 are bound to the DNA of these genes could be due to the genes' increased transcriptional activity resulting in the accumulation of proteins involved in transcription, such as Pol II and Pcf11. Looping could be an intrinsic cellular function to enable rapid generation of transcripts depending on cellular necessities, hence why most looped genes appear to be enriched for essential cellular and metabolic processes. However, I wanted to further examine this hypothesis and determine if Pcf11 depletion would affect expression of looped genes.

5.2.2.2.1 Pcf11 and Gene Expression

To investigate whether Pcf11-bound gene loop formation could be related to the activity of Pcf11 itself, or if Pcf11 molecules are merely present where looping already occurs due to increased transcriptional activity, I used publicly available 3' mRNA-seq data of Pcf11 knock-down libraries from HeLa cells. I performed differential gene expression analysis using four Pcf11 knock-down biological replicates and four siRNA Luciferase (control) ones (Accession: GSE123105).

Depletion of Pcf11 appears to disrupt gene expression, with 1,739 genes showing significant differential expression $(padj < 0.05)$ between the two conditions, indicating that the function of Pcf11 is necessary for the normal regulation of several genes. Out of the genes that showed significant differential expression, 762 are downregulated and 977 are upregulated when Pcf11 is depleted (**Fig. 13A**). Of the upregulated genes, only 58% $(n = 570)$ are protein coding, whereas out of all significantly downregulated genes, 93% $(n = 707)$ are protein coding. The observed downregulatory effect on a large number of protein coding genes upon Pcf11 knock-down suggests that the protein's activity is essential for transcription initiation and is in line with previous findings showing that knock-down of Pcf11 results in reduced transcription initiation rates (Mapendano et al., 2010). This further suggests that effective termination is necessary for transcription initiation and that the proteins involved in the transcriptional cycle, regardless of stage, could perhaps be co-regulating each other's activity. GO analysis revealed that downregulated genes are associated with metabolic and cellular processes (**Table 6**), like some of the terms found to be associated with looped genes (**Table 5**). Upregulated genes were not found to be associated with any specific GO terms, indicating that although depletion of Pcf11 results in the upregulation of a number of genes, these do not seem to be involved in any particular biological processes.

Interestingly, when comparing expression of looped and non-looped genes upon Pcf11 depletion, looped genes appear to be significantly more downregulated (**Fig. 13B**). The median log_2FC of looped genes is -0.236, whereas for non-looped ones it is 0.003 (Wilcox test $p < 7.616e-14$), indicating that the expression of looped genes is overall significantly lower than that of non-looped genes. This indicates that Pcf11 depletion heavily impacts the expression of looped genes. Given that Pcf11-bound looped genes are generally more expressed (**Fig. 12A**), their downregulation upon Pcf11 depletion suggests that Pcf11 levels control aspects of transcription initiation and gene regulation.

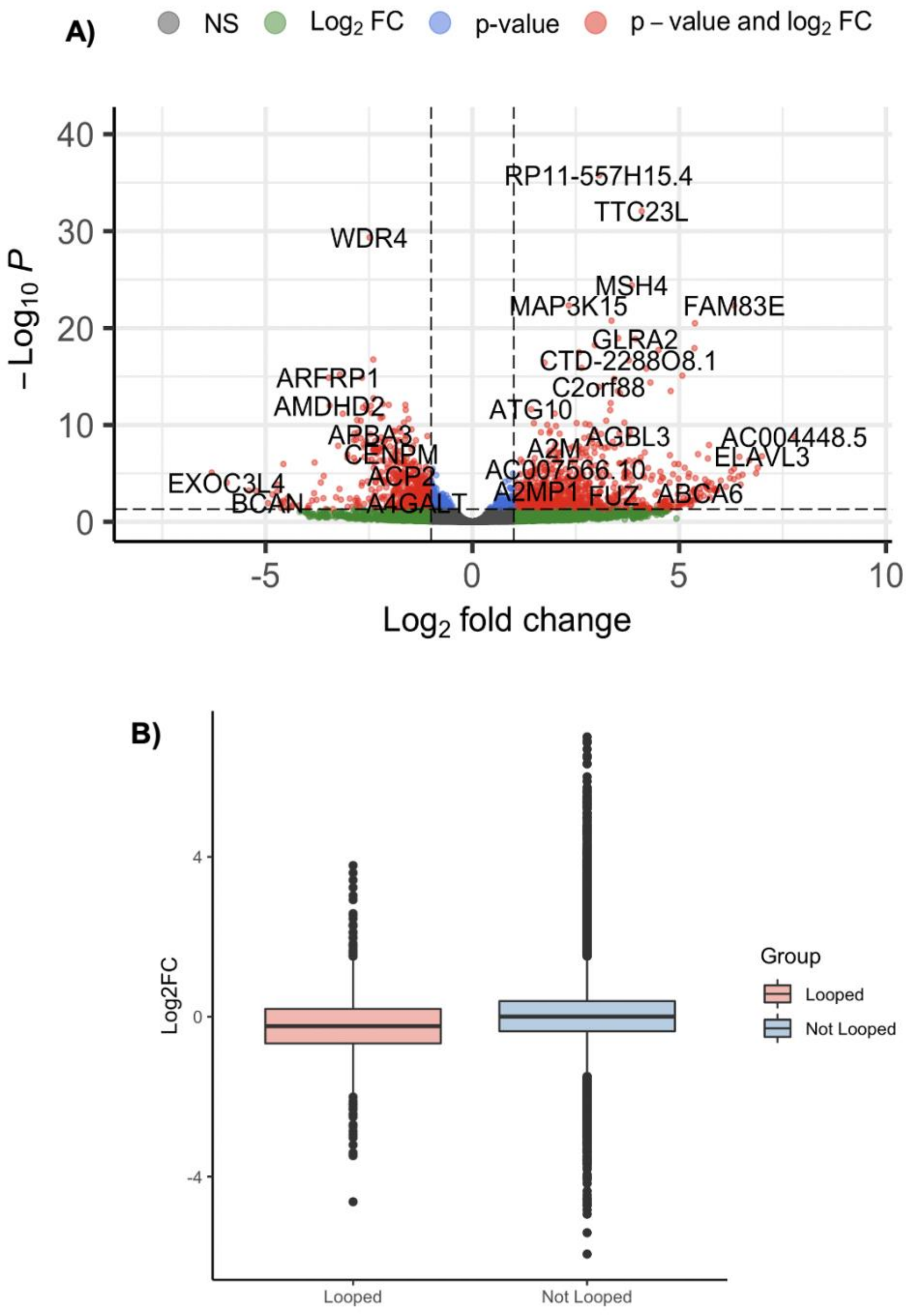
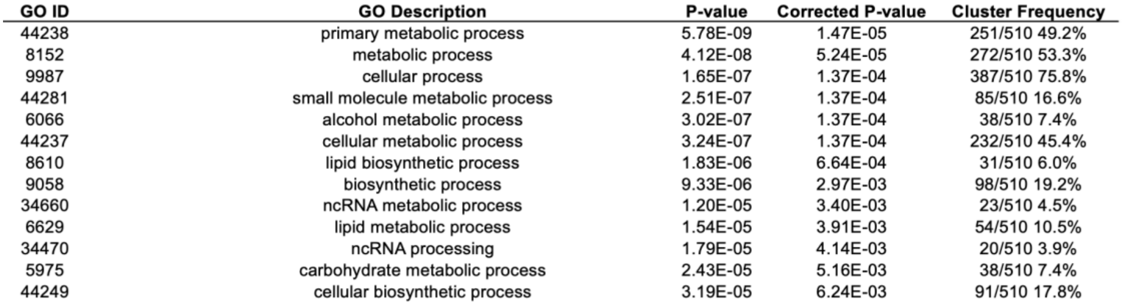


Fig. 13 Differential gene expression between siPcf11 and control samples. A) Volcano plot of differentially expressed genes between conditions (Pcf11 knock-down and control). Y-axis represents -Log10 of adjusted P-values, and x-axis represents log2FC. Right-hand side of the plot indicates genes that are upregulated in siPcf11, and the left-hand side indicates genes that are downregulated. Red dots indicate genes that are above the significance cutoff (P<0.05). **B)** Boxplot of Log2FC of looped and not looped genes upon Pcf11 depletion. Looped

genes appear to be significantly downregulated in comparison to non-looped ones ($p < 7.616e - 14$).

Table 6 Top GO terms for downregulated genes in siPcf11 vs Control samples sorted by FDR-adjusted P-value $(p < 0.05)$**.** Results generated with BiNGO on Cytoscape v3.9.1**.**



GO ID	GO Description	P-value	Corrected P-value	Cluster Frequency
44238	primary metabolic process	5.78E-09	1.47E-05	251/510 49.2%
8152	metabolic process	4.12E-08	5.24E-05	272/510 53.3%
9987	cellular process	1.65E-07	1.37E-04	387/510 75.8%
44281	small molecule metabolic process	2.51E-07	1.37E-04	85/510 16.6%
6066	alcohol metabolic process	3.02E-07	1.37E-04	38/510 7.4%
44237	cellular metabolic process	3.24E-07	1.37E-04	232/510 45.4%
8610	lipid biosynthetic process	1.83E-06	6.64E-04	31/510 6.0%
9058	biosynthetic process	9.33E-06	2.97E-03	98/510 19.2%
34660	ncRNA metabolic process	1.20E-05	3.40E-03	23/510 4.5%
6629	lipid metabolic process	1.54E-05	3.91E-03	54/510 10.5%
34470	ncRNA processing	1.79E-05	4.14E-03	20/510 3.9%
5975	carbohydrate metabolic process	2.43E-05	5.16E-03	38/510 7.4%
44249	cellular biosynthetic process	3.19E-05	6.24E-03	91/510 17.8%

5.2.2.2.2 Differential Looping

To investigate whether there is differential looping between the NS siRNA and siPcf11 samples, I performed differential looping analysis using the raw interaction coordinates for each sample (i.e., all coordinates before peak-support filtering). Ideally, this analysis would have been performed on replicates of siPcf11 and NS siRNA samples after a Hi-C experiment to avoid biasing the results from the ChIP pulldown. Although ChIA-PET interactions prior to peak-annotation are considered to be similar to Hi-C (Lee et al., 2020), a fair assumption is that the control samples would have enriched for pull-down of more Pcf11-associated regions compared to the knock-down replicates. Nonetheless, for this analysis we will treat these samples as representative of the global interactome.

Pcf11 is not a factor known to be involved in genome architecture, hence it would be interesting to potentially observe dissociation and/or formation of loops upon Pcf11 knock-down, which would indicate that the factor plays an active role in loop formation and is not merely present there. So far, most studies have investigated how interactions change upon knockdown of key factors, e.g., genes involved in methylation and loop formation, like the *Drosophila* heterochromatin protein (HP1) (Zenk et al., 2021), or by significantly disrupting the cells, e.g., through heat shock (Lyu et al., 2018). These are genomic and cellular perturbations that are expected to have prominent effects in global genome architecture, thus we expect knock-down of Pcf11 to have more subtle effects, if any.

The analysis identified 1,205 (~ 0.5 %) differential loops ($padj \leq 0.05$) out of a total of 229,877 interactions. We further filtered out this list to only consider interactions that are strongly present in one condition and entirely absent in the other (i.e., 0 PETs for both replicates). 496 loops are significantly present in the siPcf11 samples and absent in N.S. siRNA replicates, and 516 are significantly present in N.S. siRNA samples and absent from siPcf11. Knock-down of Pcf11 appears to have a small effect in global looping, however, this effect does not appear to be significantly different in terms of direction – i.e., we observe a similar proportion of overrepresented and underrepresented differential loops upon Pcf11 knock-down which could suggest that a larger sample size would be needed to draw conclusions, although most Hi-C/ChIA-PET experiments only produce two replicates. Of these statistically significant differential loops, only few correspond to gene loops and neither of these genes are Pcf11-enriched upon peak-annotation. **Fig. 14** shows a volcano plot of all differential loops (n=1,205), with each point corresponding to an interaction coordinate.

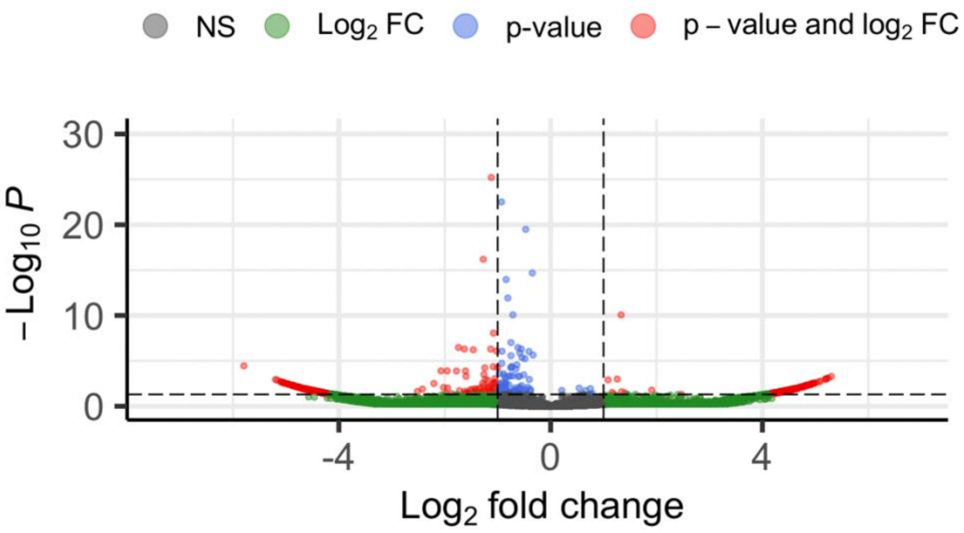


Fig 14. Volcano plot of differentially expressed loops between siPcf11 and N.S. siRNA. Each point on the plot corresponds to an interaction coordinate. Negative Log2FC indicates overrepresentation in siPcf11 replicates and positive Log2FC indicates overrepresentation in N.S. siRNA replicates.

However, when we simply look at looped genes in each sample, we find that 45 looped genes present in either of the two peak-supported NS siRNA replicates are absent from both siPcf11 replicates (0 PETs). This does not necessarily suggest that depletion of Pcf11 causes dissociation of loops or that the activity of Pcf11 is responsible for the formation of said loops, although this might be the case for some genes. Ideally, these analyses would be repeated through Hi-C experiments.

5.2.2.2.3 Is Pcf11's TSS Enrichment due to Gene Looping?

The original research question sought to determine whether Pcf11's global enrichment at the TSS of genes is due to gene looping. The findings presented in this chapter suggest that although Pcf11 is found to be bound on genes that are looped, the total number of looped genes is not high enough to explain the significant occupancy of Pcf11 on a metagene plot, which represents the average enrichment from all genes on the TSS region. To generate this profile, a significant number of genes would need to be looped and enriched for Pcf11 on their TSS.

Since non protein coding genes were excluded from the gene looping analysis, it was still possible that a large number of gene loops could have gone undetected, and that the enrichment at the TSS could be explained through significant numbers of Pcf11-bound gene loops in non-protein coding genes. To test this, I repeated the analysis of gene looping identification considering all genes. Although a higher number of loops was identified, this was still not high enough to explain genome wide TSS enrichment. Using the ChIA-PET peak-annotated interaction files as input and a +/- 500 bp cutoff threshold from the real TSS/TES, sample NS_1 returned 579 protein coding looped genes and 203 non protein coding ones. Sample NS_2 returned 240 protein coding and 115 non protein coding. Since non protein coding genes amount to 39,444 entries, this indicates that protein-coding genes $(n = 19{,}941)$ appear to be more likely to be looped, which again suggests that looping is a method of transcriptional regulation and hence less prominent in genes that do not encode for proteins. A chi-squared test was performed to test this, which returned $p < 2.2e - 16$. To confirm that this is not specific to Pcf11-bound gene loops and rather a general principle of looping, I performed the same analysis on the NS_Input sample which was not immunoprecipitated and corresponded to data equivalent to that of Hi-C. Indeed, out of all protein-coding genes ~1.2% were found to be looped, whereas out of all non-protein coding ones only ~0.3% were looped (chi-squared $p < 2.2e - 16$). Regardless, these results show that even when considering a wider pool of genes (both protein and non protein coding), Pcf11-bound gene loops are still not enough to cause a genome-wide TSS enrichment.

Another consideration that is worth mentioning is the possibility that genes with more than one promoter could also form loops spanning from their respective TSS to the TES. For example, a gene could have three transcription start sites, where the distal promoter is kilobases away from the start of the gene. For this work, each gene was assumed to have one TSS site corresponding to the start of the gene. Although this could be biassing against identifying gene loops in longer genes, or genes with more than one promoter site, the metagene plot is constructed by also assuming that there is only a single TSS corresponding to the start site of each gene. Therefore, although it would be interesting to investigate gene looping spanning from distal promoters to the TES, it is not something that would have helped us answer the original research question further.

The HEK-293 ChIA-PET libraries correspond to samples treated with the (FP) Flavopiridol inhibitor, as well as non-treated cells (NT). These samples are of lower quality in terms of interactome information (low number of unique PETs) but were still used to explore the effect of FP treatment on the occupancy of Pcf11. After normalising for each sample's read depth, a metagene plot was generated (**Fig. 15**). The metagene values were calculated by:

$$\frac{mean(FP\_1, FP\_2)/Input\_FP}{mean(NT\_1, NT\_2)/Input\_NT}$$

This is a way to look at the overall difference in Pcf11 binding across the TSS, gene body, and TES upon treatment with FP, compared to non-treated cells. The calculation takes into account input controls for each condition. We observe a relative reduction in the signal ratio around the TSS, where we would have expected Pcf11 to be enriched, indicating that treatment with FP reduces the total binding of Pcf11 around the TSS of genes. We also observe a peak at the TES, indicating an increase of Pcf11 at the end of genes upon FP treatment.

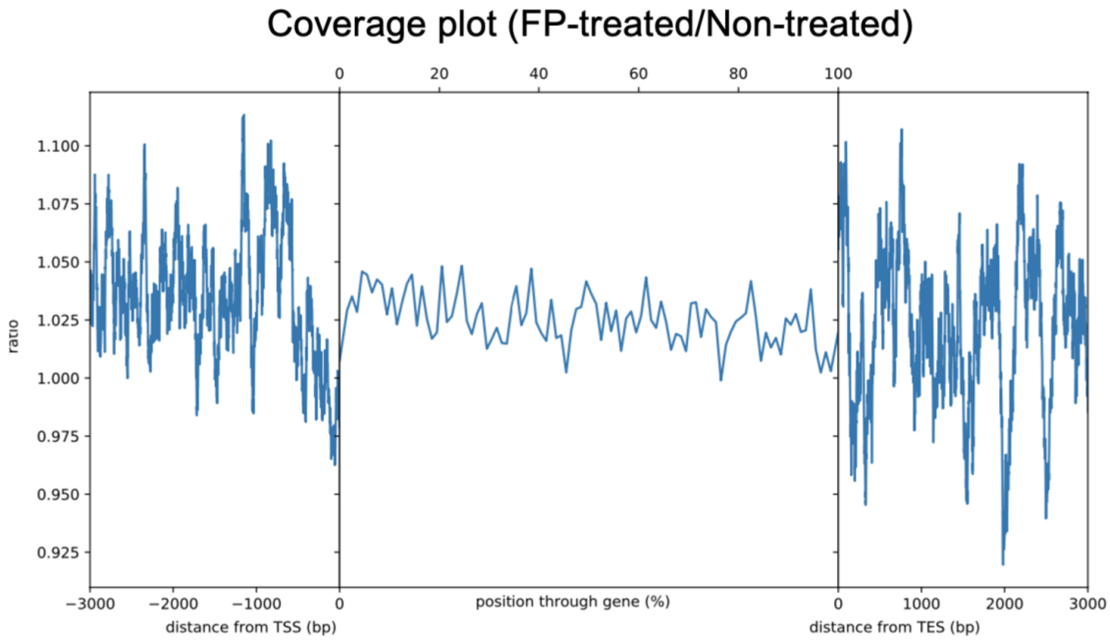


Fig. 15 Metagene plot of Pcf11 FP/Non-treated HEK-293 ChIA-PET samples. Average signal ratio across TSS, gene body, and TES. Flavopiridol treatment causes reduction of Pcf11 binding near the TSS of genes.

To statistically test the effect of FP treatment on Pcf11 binding, a bootstrapping method was applied. First, the mean signal ratio value +/- 200 bp of the TSS and TES on the metagene plot were calculated. These corresponded to 0.993 and 1.037, respectively. A bootstrapped distribution was generated, made of 1000 sampled points (**Fig. 16**). Each sample was generated by selecting a random 401 bp section from the gene body and/or TSS/TES and getting the mean signal across those bases for each dataset. This was done for every gene and then averaged across all genes for each dataset, before calculating the final value using the equation mentioned previously. This process produces one point in the distribution and was repeated 1000 times to generate the bootstrapped distribution. The minimum sampled value is 1.035, which is far higher than the original TSS value (0.993), indicating $p << 0.001$. The maximum sampled value is 1.015, which is only slightly lower than the value for TES Pcf11 enrichment, indicating $p <= 0.001$. We cannot quantify with more precision due to only having 1000 samples, however, the effects observed at the TSS and TES upon FP treatment are statistically significant.

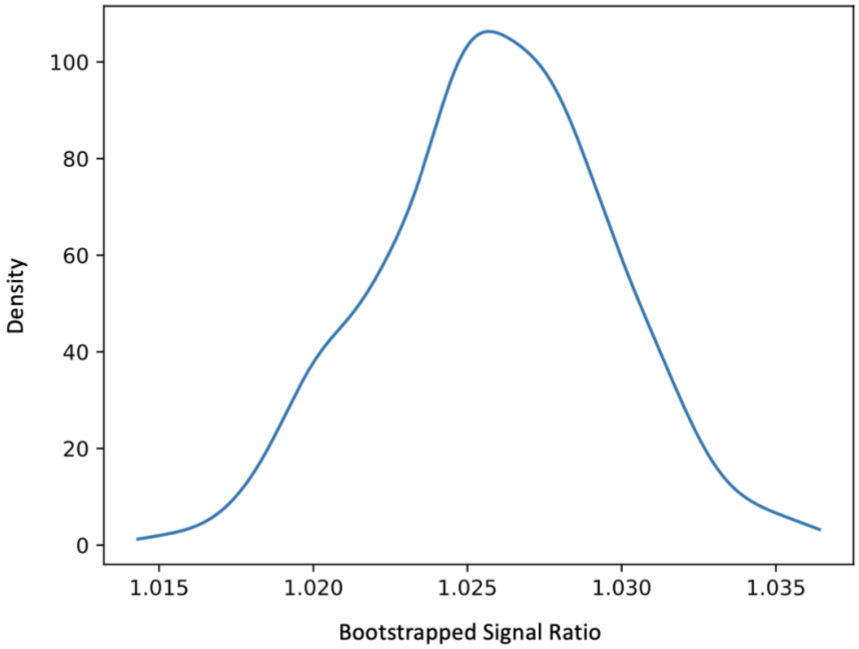


Fig. 16 Distribution of 1000 bootstrapped signal ratio values. To determine the significance of the observed TSS and TES profiles of Pcf11 enrichment upon FP treatment, a bootstrapped method was applied. Most signal ratio values range between 1.020 - 1.030, indicating that the observed dip at the TSS and enrichment at the TES are statistically significant.

As I have mentioned previously in this work, FP is a potent inhibitor of transcription which specifically inhibits the function of CDK9/p-TEFb which works by increasingly phosphorylating Pol II on Ser2. Phosphorylation by CDK9 leads to pause release of Pol II and allows Pol II to progress to the elongation stage. Inhibiting p-TEFb function with the use of FP has been shown to result in an increase of polymerases at the pausing site, close to the promoter (Jonkers et al., 2014). Assuming that Pcf11 stays bound to Pol II after termination and that Pol II is recycled back to nearby promoters (Kang et al., 2020), then possibly, reduction of Pcf11 enrichment at the TSS upon FP treatment could be due to polymerases being ‘stuck’ at the promoter site, preventing them from reaching the 3’ end of genes and transporting Pcf11 back to the promoter with them. Similarly, the increase of Pcf11 enrichment at the TES could be a result of Pcf11 build-up due to reduction of Pol II at the 3' end. Kamieniarz-Gdula et al (2019) report that Pcf11 appears to preferentially bind and enhance the expression of closely spaced genes. I hypothesise that this could be due to Pcf11 being bound to Pol II as it recycles back to nearby promoters. However, these experiments would need to be repeated to confirm the effect of FP treatment on Pcf11's occupancy.

5.3. Discussion

Gene looping is a phenomenon that has long been speculated, although its properties and functions remain unclear. In the present work, I generated Pcf11 data from ChIA-PET experiments in HeLa and HEK-293 cells, as well as equivalent Pol II libraries from U2OS cells. The primary aim of this work was to determine whether the significant enrichment of Pcf11 at the TSS of genes, as revealed from ChIP-seq data, was a result of gene looping.

Peak-supported interactions were isolated for each dataset using peak annotations generated via MACS2 on BroadPeak mode (method details available in Chapter 2). Although the immunoprecipitation step of the protocol should only capture interactions bound by the factor of interest, noise will inevitably be pulled down as well. For this reason, peak-support is an important step of the ChIA-PET data analysis protocol. ChIA-PIPE's default peak-calling algorithms include SPP (Kharchenko et al., 2008) and MACS2. MACS2 was preferred due to familiarity with the software, as well as due to the ChIP-seq Pcf11 data having been processed in the same way. MACS2 can be run in either BroadPeak or NarrowPeak (default) mode. NarrowPeak employs a more stringent algorithm and is appropriate for the identification of narrow peaks, e.g., for TFs who have specific binding regions. BroadPeak is used for broader peaks, which is suitable for factors that span wider ranges of genomic regions. Therefore, when investigating the genome-wide effects of a factor, BroadPeak mode is more suitable. As it was shown in **Table 1A-C**, the interaction coordinates that were retained for each sample corresponded to only those where both anchors were peak-supported. In chapter 4, the Pcf11 HEK-293 dataset was identified to be of poorer quality given the low number of unique interactions. This became more apparent after quantifying total and peak-supported interactions which revealed that the libraries would be unsuitable for gene loop identification analyses. Overall, the rest of the libraries (Pol II U2OS and Pcf11 HeLa) were deemed suitable for further analyses.

The Pol II samples all came from the same cell line and were generated as a single batch of samples. The difference between the libraries is the specificity of the Pol II antibody used. Library 1C7 is specific to the Pol II CTD heptad repeat and binds total Pol II, whereas libraries 3E8 and 4E12 are specific to the CTD's Ser5P and Ser7P, respectively. Differences in cell numbers and total DNA concentration will impact total numbers of interactions captured, however, after normalisation we observe that library 1C7 yields the lowest number of looped genes, possibly suggesting that enriching for actively transcribing Pol IIs can explain, to an extent, why the other two libraries identify more looped genes. It is also important to note that the method of gene loop identification is binary and sorts genes as looped and not looped. Minor differences in the interaction coordinates between samples would result in genes being filtered out if their anchors do not fall within the defined threshold of +/- 500 bp of TSS/TES. By repeating the analysis and allowing for a wider threshold of overlap (+/- 1 kb), I was able to identify a larger number of looped genes, but for the purposes of this work I wanted to isolate genes that show high confidence in promoter-termination sequence overlap.

The identification of looped genes revealed that there is a level of variation between the samples. Although the overlap between samples was statistically significant, only few genes overlapped between the three Pol II samples. In the two Pcf11 HeLa replicates, the overlap was ~ 13%. This is attributed to the gene loop identification method, antibody differences in the Pol II dataset, but also the biological variability found in interactome datasets. Lareau and Aryee (2018) report that there appears to be a possible artefact of ChIA-PET which results in extreme inconsistencies between replicates. This refers to the phenomenon where specific loops are overrepresented in one replicate (>= 5 PETs mapping to the region) but are entirely absent in the other replicate. It is not yet clear what such non-replicated loops between samples are attributed to.

Even though ChIA-PET/Hi-C datasets represent the average of all interactions between millions of cells, it is interesting to note that on the single-cell level, the interactome of individual cells varies greatly. Nagano and colleagues (2013) developed single-cell Hi-C and generated single-cell Hi-C libraries for individual cells. They found that structure domains appear to be highly variable between cells, but the analysis was not able to quantify variability in the high-resolution internal structure of domains. It would therefore be interesting to investigate the level of variation in single cells when it comes to gene looping. Additionally, ChIA-Drop (Zheng et al., 2019) could also be utilised to explore the effects of multiplexed interactions as it overcomes the limitation of relying on pairwise proximity ligation. ChIA-Drop is a droplet-based microfluidics assay with the ability to capture single-cell chromatin interactions, something that is not possible with methods that are based on population-level pairwise contacts (e.g., ChIA-PET) (Zheng et al., 2019).

Overall, it appears that looped genes are characterised by high expression and shorter length, as it was shown in **Fig.6 and 7**. These characteristics were also observed for genes with Ser5P enrichment at both their TSS and TES, as shown from analysis of ChIP-seq data (Grosso et al., 2012). High expression in looped genes is not surprising given the juxtaposition of the promoter and terminator sites allowing for efficient recycling of factors back to the promoter. It is unclear what exactly triggers looping and whether this is an intrinsic function of cells or if it is mediated by specific factors. The overlap between looped genes in U2OS and HeLa cells is ~14%, and although one dataset was IP'd for Pol II and the other for Pcf11, the overlap could suggest that looping is an intrinsic function of specific genes as well. Although, a more suitable comparison here would have been between the same cell lines. As it was shown in **Fig. 12B**, the expression of Pcf11-bound looped genes appears to be lower than that of non-looped genes when Pcf11 is depleted, suggesting that Pcf11's function contributes to transcriptional activity and perhaps regulates aspects of transcription initiation in looped genes.

It is also possible that loops are formed in a transient and asynchronous fashion. Perhaps, we observe high expression as a characteristic of looped genes because these are the genes that are actively transcribed by the majority of cells in a given cell population. The present work also identified that gene looping seems to occur more frequently in protein-coding genes, which also suggest that looping is a way of transcriptional regulation. The field would benefit from research investigating the path to looping and potentially identifying if loops are more likely to be formed when rapid generation of transcript is required. For example, if a population of cells was put under stress, would we observe different clusters of genes being looped between samples?

The hypothesis that termination factors piggy-back off Pol II and find their way back to the promoter region has been suggested before (Hsin and Manley, 2012) due to certain TFs being associated with factors involved in mRNA 3' formation. The authors proposed that factors required for 3' end formation potentially load onto the promoters of genes and then travel with elongating Pol II to the termination region. This however is unlikely to be the case, at least for most termination factors, since ChIP-seq data in yeast revealed that polyadenylation factors are only found at regions of termination (O'Sullivan et al., 2004). In the case of Pcf11 where significant enrichment is observed at the TSS, this hypothesis appears more likely. Although, when the authors compared ChIP-seq signals of Pol II and Pcf11, it became apparent that Pcf11 does not consistently travel with Pol II from the promoter to the PAS site, but they suggest it is possible that they interact across the gene body (Kamieniarz-Gdula et al., 2019). Possibly, Pcf11 remains bound onto Pol II and then dissociates from the complex when Ser2P levels drop. It is also worth considering the phenomenon of promoter proximal premature transcription termination (PTT). This is a process that appears to be widespread and can occur close to the TSS (Kamieniarz-Gdula and Proudfoot, 2019). These premature transcripts are alternatively polyadenylated (APA) and negatively regulate gene expression, since premature termination opposes full-length transcript formation. As it was shown from ChIP data, Pol II accumulates at the TSS, and is usually interpreted as stably paused Pol II. However, this might also be due to PTT as termination and RNA 3'-processing factors have been observed to accumulated at the 5' ends of genes (Brannan et al., 2012; Wagschal et al., 2012). Hence, Pcf11 accumulation at the promoter could be a result of PTT.

Inhibition of Pol II Ser5 phosphorylation through FP treatment (**Fig. 13**) resulted in reduction of Pcf11 enrichment at the TSS and increase at the TES. Although this could suggest interaction or co-regulation between Pol II and Pcf11, these experiments would need to be repeated to confirm these effects. The validity of these experiments hinge on the specificity of the inhibitor, however, FP appears to be quite specific (Jonkers et al., 2014). Cell treatment with Triptolide (Trp) could also provide insights regarding Pcf11's binding profile and its relationship with Pol II. Trp inhibits initiation of transcription (Jonkers et al., 2014) and it would be interesting to study how that affects Pcf11's occupancy at the TSS.

In summary, the present work shows that looping is a phenomenon that positively correlates with gene expression, indicating that gene looping is not a random occurrence and that it has significance in gene regulation. Analysis of Pcf11 3' mRNA data also revealed that Pcf11-bound looped genes are downregulated when Pcf11 is depleted, confirming previously observed findings that Pcf11 activity regulates transcription initiation.

Chapter 6: Conclusion

6.1 Summary of Main Findings

The juxtaposition of otherwise far-away DNA sequences is a phenomenon that has roles in the expression and regulation of genes. ChIA-PET provides a way to capture long-range chromatin interactions bound by a factor of interest. In the present work, I primarily sought to identify whether the terminator factor Pcf11 is implicated in gene looping, after previously published ChIP-seq data revealed high Pcf11 occupancy at the TSS of genes.

Given that the *in-situ* ChIA-PET protocol is a very recent development, extensive QC of libraries was required to determine data quality, suitability for downstream analyses, and to establish best laboratory practices for the generation of subsequent libraries. Using published datasets from the lab that developed the technique as benchmarks for comparison, I established data quality of my own samples. Since ChIA-PET libraries contain data comparable to both Hi-C and ChIP-seq, it was essential to cover all aspects of QC. It was determined that the Pcf11 HeLa and Pol II U2OS libraries met the quality standards, but that the Pcf11 HEK-293 samples had captured smaller numbers of unique interactions which could hinder downstream analyses. Library quality is an essential step in genomic analyses, especially when data are derived from techniques that are less widely used, such as *in-situ* ChIA-PET. The quality of these libraries is dependent on the success of several reactions, e.g., ligation, ChIP, and biotin pull-down which require checkpoints not only during the library production phase, but during the computational processing as well. Since all libraries were generated before the official publication of the protocol, and HEK-293 libraries were the first to be generated as part of this research, the present work highlights the improvement in subsequent libraries (HeLa and U2OS) after experimental optimisation was applied.

Additionally, we were able to identify looped genes bound by both Pol II and Pcf11. Through statistical and biological characterisation, we show that looping is not a random phenomenon, but rather a juxtaposition that appears to correlate with transcriptional activity. Generally, looping is more likely to be a feature of protein-coding genes. Through the utilisation of bulk RNA-seq data, we also observed that looped genes are more highly expressed than non-looped genes and are also involved in key cellular processes. High expression could be a result of active transcription, which leads to the assumption that looping could be occurring transiently in other genes, but only while they are in the process of being transcribed. The use of siPcf11 3' mRNA-seq data also revealed that Pcf11's activity appears to be essential to the expression of looped genes. In Pcf11-depleted mRNA samples, we observe a significant reduction in expression in looped genes compared to non-looped genes. This is in line with previous findings showing that Pcf11 knock-down leads to reduction in transcription initiation (Mapendano et al., 2010). We also observed that looped genes were shorter than non-looped genes, and although UTR length correlates with gene size, Pol II-bound looped genes do not appear to have significantly shorter 3' UTRs than non-looped genes. However, in the case of Pcf11-bound looped genes, we observed that looped genes had significantly shorted 3' UTRs on average – both compared to non-looped genes, as well as Pol II-bound looped genes.

Although Pcf11 finds its way to the TSS of genes through gene looping, as the present work has shown, looping alone cannot explain Pcf11's genome-wide enrichment at the TSS of genes. The metagene profile of untreated and FP-treated HEK-293 samples suggest that Pcf11's occupancy at the TSS depends on the function of Pol II. A possible hypothesis is that Pcf11 could be remaining bound onto Pol II and carried to the promoter as Pol II gets recycled back to nearby promoters. However, these experiments would need to be repeated through ChIP-seq and with more replicates.

6.2 Future Work

To strengthen the present work around gene looping and to also address further questions, it would be favourable to produce Pol II ChIA-PET libraries in HeLa cells. This would allow for a direct comparison between looped genes bound by both Pcf11 and Pol II and eliminate the bias of cross cell-line comparison. Few other factors besides Pol II, CTCF, and histones have been investigated for their involvement in long-range chromatin interactions. It would be of interest to the field of 3D genomics to explore the interactome of other factors, especially ones known to be involved in action-at-a-distance gene regulation. Although most genes are not known, or expected, to influence global chromatin conformation, it would still be interesting to understand the interaction network of different factors and how distant genomic loci come into contact with one another.

Additionally, little is known about transient looping and the path-to-looping. Studying interactions in real-time would provide valuable insights about how loops are formed and dissociate and what that means for gene expression and regulation. Understanding the mechanisms that drive looping would also shed light in the field of 3D genome structure. Even though we know that proteins involved in transcription, replication, and RNA processing assemble into membraneless compartments and that they have functional roles in genome regulation, the underlying mechanism driving their formation is not fully known. A possible hypothesis is that chromatin conformation is influenced by biophysical drivers, such as liquid-liquid phase separation (LLPS) (Rippe, 2022). Through *in vitro* induction of LLPS, we could apply high-throughput sequencing methods such as ChIA-PET to study looping and gene expression alterations, something that is actively investigated in our lab.

The enrichment of Pcf11 at the TSS of genes is still also not fully understood, although a combination of drivers could be contributing to this effect, with one of them being gene looping. To further explore this question, research should aim to improve our understanding of Pcf11's interactions with Pol II. Pcf11 interacts with Pol II through its CTD-interacting domain (CID), thus, future experiments could seek to explore how transcriptional activity and Pcf11 occupancy are impacted upon CID depletion. Until recently, we have mostly considered transcriptional factors to have separate and distinct roles during the process of transcription, but it is possible that factors implicated in different stages interact and co-regulate each other's activity throughout the transcriptional cycle. For example, it has been suggested that Pol II can 'sense' its passage through a functional polyadenylation site, likely through the interactions between the CPA complex with its CTD. This interaction can result in conformational changes in the active site of Pol II and result in pause, followed by release (Proudfoot, 2016). It is therefore important to understand the connection between all aspects of transcription and how they all orchestrate effective initiation and termination.

Bibliography

Achinger-Kawecka, J., Taberlay, P. and Clark, S., 2016. Alterations in Three-Dimensional Organization of the Cancer Genome and Epigenome. *Cold Spring Harbor Symposia on Quantitative Biology*, 81, pp.41-51.

Adelman, K. and Lis, J., 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews Genetics*, 13(10), pp.720-731.

Allepuz-Fuster, P., O'Brien, M., González-Polo, N., Pereira, B., Dhoondia, Z., Ansari, A. and Calvo, O., 2019. RNA polymerase II plays an active role in the formation of gene loops through the Rpb4 subunit. *Nucleic Acids Research*, 47(17), pp.8975-8987.

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H. and Shiroishi, T., 2009. Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Developmental Cell*, 16(1), pp.47-57.

Amrani, N., Minet, M., Wyers, F., Dufour, M., Aggerbeck, L. and Lacroute, F., 1997. PCF11 encodes a third protein component of yeast cleavage and polyadenylation factor I. *Molecular and Cellular Biology*, 17(3), pp.1102-1109.

Ansari, A. and Hampsey, M., 2005. A role for the CPF 3'-end processing machinery in RNAP II-dependent gene looping. *Genes & Development*, 19(24), pp.2969-2978.

Barilla, D., Lee, B. and Proudfoot, N., 2001. Cleavage/polyadenylation factor IA associates with the carboxyl-terminal domain of RNA polymerase II in Saccharomyces cerevisiae. *Proceedings of the National Academy of Sciences*, 98(2), pp.445-450.

Barutcu, A., Maass, P., Lewandowski, J., Weiner, C. and Rinn, J., 2018. A TAD boundary is preserved upon deletion of the CTCF-rich Firre locus. *Nature Communications*, 9(1).

Baumli, S., Lolli, G., Lowe, E., Troiani, S., Rusconi, L., Bullock, A., Debreczeni, J., Knapp, S. and Johnson, L., 2008. The structure of P-TEFb (CDK9/cyclin T1), its complex with flavopiridol and regulation by phosphorylation. *The EMBO Journal*, 27(13), pp.1907-1918.

Beagan, J. and Phillips-Cremins, J., 2020. On the existence and functionality of topologically associating domains. *Nature Genetics*, 52(1), pp.8-16.

Belaghzal, H., Dekker, J. and Gibcus, J., 2017. Hi-C 2.0: An optimized Hi-C procedure for high-resolution genome-wide mapping of chromosome conformation. *Methods*, 123, pp.56-65.

Bengtsson, L. and Otto, H., 2008. LUMA interacts with emerin and influences its distribution at the inner nuclear membrane. *Journal of Cell Science*, 121(4), pp.536-548.

Bolzer, A., Kreth, G., Solovei, I., Koehler, D., Saracoglu, K., Fauth, C., Müller, S., Eils, R., Cremer, C., Speicher, M. and Cremer, T., 2005. Three-Dimensional Maps of All Chromosomes in Human Male Fibroblast Nuclei and Prometaphase Rosettes. *PLoS Biology*, 3(5), p.e157.

Bondarenko, V., Liu, Y., Jiang, Y. and Studitsky, V., 2003. Communication over a large distance: enhancers and insulators. *Biochemistry and Cell Biology*, 81(3), pp.241-251.

Branco, M. and Pombo, A., 2006. Intermingling of Chromosome Territories in Interphase Suggests Role in Translocations and Transcription-Dependent Associations. *PLoS Biology*, 4(5), p.e138.

Brannan, K. 2012. mRNA decapping factors and the exonuclease Xrn2 function in widespread premature termination of RNA polymerase II transcription. *Mol. Cell*. (46) pp.311-324

Bridger, J., 2011. Chromobility: the rapid movement of chromosomes in interphase nuclei. *Biochemical Society Transactions*, 39(6), pp.1747-1751.

Cai, S., Lee, C. and Kohwi-Shigematsu, T., 2006. SATB1 packages densely looped, transcriptionally active chromatin for coordinated expression of cytokine genes. *Nature Genetics*, 38(11), pp.1278-1288.

Cassandri, M., Smirnov, A., Novelli, F., Pitolli, C., Agostini, M., Malewicz, M., Melino, G. and Raschellà, G., 2017. Zinc-finger proteins in health and disease. *Cell Death Discovery*, 3(1).

Cavallaro, M., Walsh, M., Jones, M., Teahan, J., Tiberi, S., Finkenstädt, B. and Hebenstreit, D., 2021. 3 ′-5 ′ crosstalk contributes to transcriptional bursting. *Genome Biology*, 22(1).

Cavalli, G. and Misteli, T., 2013. Functional implications of genome topology. *Nature Structural & Molecular Biology*, 20(3), pp.290-299.

Chao, S. and Price, D., 2001. Flavopiridol Inactivates P-TEFb and Blocks Most RNA Polymerase II Transcription in Vivo. *Journal of Biological Chemistry*, 276(34), pp.31793-31799.

Core, L., Waterfall, J. and Lis, J., 2008. Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters. *Science*, 322(5909), pp.1845-1848.

Cournac, A. and Plumbridge, J., 2013. DNA Looping in Prokaryotes: Experimental and Theoretical Approaches. *Journal of Bacteriology*, 195(6), pp.1109-1119.

Cremer, T. and Cremer, C., 2006. Rise, fall and resurrection of chromosome territories: a historical perspective. Part I. The rise of chromosome territories. *Eur J Histochem*, 50(3), pp.161-76.

Cremer, T., Cremer, M., Dietzel, S., Müller, S., Solovei, I. and Fakan, S., 2006. Chromosome territories – a functional nuclear landscape. *Current Opinion in Cell Biology*, 18(3), pp.307-316.

Darrow, E., Huntley, M., Dudchenko, O., Stamenova, E., Durand, N., Sun, Z., Huang, S., Sanborn, A., Machol, I., Shamim, M., Seberg, A., Lander, E., Chadwick, B. and Aiden, E., 2016. Deletion of <i>DXZ4</i> on the human inactive X chromosome alters higher-order genome architecture. *Proceedings of the National Academy of Sciences*, 113(31).

Dekker, J., Belmont, A., Guttman, M., Leshyk, V., Lis, J., Lomvardas, S., Mirny, L., O'Shea, C., Park, P., Ren, B., Politz, J., Shendure, J. and Zhong, S., 2017. The 4D nucleome project. *Nature*, 549(7671), pp.219-226.

DeMare, L., Leng, J., Cotney, J., Reilly, S., Yin, J., Sarro, R. and Noonan, J., 2013. The genomic landscape of cohesin-associated chromatin interactions. *Genome Research*, 23(8), pp.1224-1234.

Dixon, J., Jung, I., Selvaraj, S., Shen, Y., Antosiewicz-Bourget, J., Lee, A., Ye, Z., Kim, A., Rajagopal, N., Xie, W., Diao, Y., Liang, J., Zhao, H., Lobanenkov, V., Ecker, J., Thomson, J. and Ren, B., 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539), pp.331-336.

Dixon, J., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J. and Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398), pp.376-380.

Dixon, J., Xu, J., Dileep, V., Zhan, Y., Song, F., Le, V., Yardımcı, G., Chakraborty, A., Bann, D., Wang, Y., Clark, R., Zhang, L., Yang, H., Liu, T., Iyyanki, S., An, L., Pool, C., Sasaki, T., Rivera-Mulia, J., Ozadam, H., Lajoie, B., Kaul, R., Buckley, M., Lee, K., Diegel, M., Pezic, D., Ernst, C., Hadjur, S., Odom, D., Stamatoyannopoulos, J., Broach, J., Hardison, R., Ay, F., Noble, W., Dekker, J., Gilbert, D. and Yue, F., 2018. Integrative detection and analysis of structural variation in cancer genomes. *Nature Genetics*, 50(10), pp.1388-1398.

Dori, M. and Forcato, M., 2021. Analysis of HiChIP Data. *Methods in Molecular Biology*, pp.209-234.

Dowen, J., Fan, Z., Hnisz, D., Ren, G., Abraham, B., Zhang, L., Weintraub, A., Schuijers, J., Lee, T., Zhao, K. and Young, R., 2014.

Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. *Cell*, 159(2), pp.374-387.

Drissen, R., Palstra, R., Gillemans, N., Splinter, E., Grosveld, F., Philipsen, S. and de Laat, W., 2004. The active spatial organization of the β-globin locus requires the transcription factor EKLF. *Genes & Development*, 18(20), pp.2485-2490.

Dundr, M., Ospina, J., Sung, M., John, S., Upender, M., Ried, T., Hager, G. and Matera, A., 2007. Actin-dependent intranuclear repositioning of an active gene locus in vivo. *Journal of Cell Biology*, 179(6), pp.1095-1103.

Durand, N., Robinson, J., Shamim, M., Machol, I., Mesirov, J., Lander, E. and Aiden, E., 2016. Juicebox Provides a Visualization System for Hi-C Contact Maps with Unlimited Zoom. *Cell Systems*, 3(1), pp.99-101.

Egloff, S., 2012. Role of Ser7 phosphorylation of the CTD during transcription of snRNA genes. *RNA Biology*, 9(8), pp.1033-1038.

Eick, D. and Geyer, M., 2013. The RNA Polymerase II Carboxy-Terminal Domain (CTD) Code. *Chemical Reviews*, 113(11), pp.8456-8490.

Erdel, F. and Rippe, K., 2018. Formation of Chromatin Subcompartments by Phase Separation. *Biophysical Journal*, 114(10), pp.2262-2270.

Eskiw, C., Cope, N., Clay, I., Schoenfelder, S., Nagano, T. and Fraser, P., 2010. Transcription Factories and Nuclear Organization of the Genome. *Cold Spring Harbor Symposia on Quantitative Biology*, 75(0), pp.501-506.

Ferrai, C., Xie, S., Luraghi, P., Munari, D., Ramirez, F., Branco, M., Pombo, A. and Crippa, M., 2010. Poised Transcription Factories Prime Silent uPA Gene Prior to Activation. *PLoS Biology*, 8(1), p.e1000270.

Ferraiuolo, M.A., Rousseau, M., Miyamoto, C., Shenker, S., Qing, X., Wang., D., Nadler, M., Blanchette, M., Dostie., J. 2010. The three-dimensional architecture of *Hox* cluster silencing. *Nucleic Acids Research*, 38(21), pp. 7472-7484

Forcato, M., Nicoletti, C., Pal, K., Livi, C., Ferrari, F. and Bicciato, S., 2017. Comparison of computational methods for Hi-C data analysis. *Nature Methods*, 14(7), pp.679-685.

Franke, M., Ibrahim, D., Andrey, G., Schwarzer, W., Heinrich, V., Schöpflin, R., Kraft, K., Kempfer, R., Jerković, I., Chan, W., Spielmann, M., Timmermann, B., Wittler, L., Kurth, I., Cambiaso, P., Zuffardi, O., Houge, G., Lambie, L., Brancati, F., Pombo, A., Vingron, M., Spitz, F. and Mundlos, S., 2016. Formation of new chromatin domains determines pathogenicity of genomic duplications. *Nature*, 538(7624), pp.265-269.

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N. and Mirny, L., 2016. Formation of Chromosomal Domains by Loop Extrusion. *Cell Reports*, 15(9), pp.2038-2049.

Fullwood, M., Liu, M., You Fu, P., Liu, J., Xu, H., Mohamed, Y., Orlov, Y., Ho, S. and Andrea, M., 2009. An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, 462(7269), pp.58-64.

Ganji, M., Shaltiel, I., Bisht, S., Kim, E., Kalichava, A., Haering, C. and Dekker, C., 2018. Real-time imaging of DNA loop extrusion by condensin. *Science*, 360(6384), pp.102-105.

Gao, K. and Huang, L., 2008. Nonviral Methods for siRNA Delivery. *Molecular Pharmaceutics*, 6(3), pp.651-658.

Giorgetti, L., Galupa, R., Nora, E., Piolot, T., Lam, F., Dekker, J., Tiana, G. and Heard, E., 2014. Predictive Polymer Modeling Reveals Coupled Fluctuations in Chromosome Conformation and Transcription. *Cell*, 157(4), pp.950-963.

Goloborodko, A., Imakaev, M., Marko, J. and Mirny, L., 2016. Compaction and segregation of sister chromatids via active loop extrusion. *eLife*, 5.

Gonzalez-Sandoval, A., Towbin, B., Kalck, V., Cabianca, D., Gaidatzis, D., Hauer, M., Geng, L., Wang, L., Yang, T., Wang, X., Zhao, K. and Gasser, S., 2015. Perinuclear Anchoring of H3K9-Methylated Chromatin Stabilizes Induced Cell Fate in C. elegans Embryos. *Cell*, 163(6), pp.1333-1347.

Gross, S. and Moore, C., 2001. Five subunits are required for reconstitution of the cleavage and polyadenylation activities of Saccharomyces cerevisiae cleavage factor I. *Proceedings of the National Academy of Sciences*, 98(11), pp.6080-6085.

Grosso, A., de Almeida, S., Braga, J. and Carmo-Fonseca, M., 2012. Dynamic transitions in RNA polymerase II density profiles during transcription termination. *Genome Research*, 22(8), pp.1447-1456.

Gruber, A.R., Martin, G., Keller, W., Zavolan, M. 2013. Means to and end: mechanisms of alternative polyadenylation of messenger RNA precursors. *WIREs RNA*. 5(2):183-96.

Grzechnik, P., Gdula, M. and Proudfoot, N., 2015. Pcf11 orchestrates transcription termination pathways in yeast. *Genes & Development*, 29(8), pp.849-861.

Guo, Y., Manteiga, J., Henninger, J., Sabari, B., Dall'Agnese, A., Hannett, N., Spille, J., Afeyan, L., Zamudio, A., Shrinivas, K., Abraham, B., Boija, A., Decker, T., Rimel, J., Fant, C., Lee, T., Cisse, I., Sharp, P., Taatjes, D. and Young, R., 2019. Pol II phosphorylation regulates a switch between transcriptional and splicing condensates. *Nature*, 572(7770), pp.543-548.

Hacisuleyman, E., Goff, L., Trapnell, C., Williams, A., Henao-Mejia, J., Sun, L., McClanahan, P., Hendrickson, D., Sauvageau, M., Kelley, D., Morse, M., Engreitz, J., Lander, E., Guttman, M., Lodish, H., Flavell, R., Raj, A. and Rinn, J., 2014. Topological organization of multichromosomal regions by the long intergenic noncoding RNA Firre. *Nature Structural & Molecular Biology*, 21(2), pp.198-206.

Han, H., 2018. RNA Interference to Knock Down Gene Expression. *Methods in Molecular Biology*, pp.293-302.

Handoko, L., Xu, H., Li, G., Ngan, C., Chew, E., Schnapp, M., Lee, C., Ye, C., Ping, J., Mulawadi, F., Wong, E., Sheng, J., Zhang, Y., Poh, T., Chan, C., Kunarso, G., Shahab, A., Bourque, G., Cacheux-Rataboul, V., Sung, W., Ruan, Y. and Wei, C., 2011. CTCF-mediated functional chromatin interactome in pluripotent cells. *Nature Genetics*, 43(7), pp.630-638.

Harlen, K. and Churchman, L., 2017. The code and beyond: transcription regulation by the RNA polymerase II carboxy-terminal domain. *Nature Reviews Molecular Cell Biology*, 18(4), pp.263-273.

Hebenstreit, D., 2013. Are gene loops the cause of transcriptional noise?. *Trends in Genetics*, 29(6), pp.333-338.

Hollingworth, D., Noble, C., Taylor, I. and Ramos, A., 2006. RNA polymerase II CTD phosphopeptides compete with RNA for the interaction with Pcf11. *RNA*, 12(4), pp.555-560.

Holwerda, S. and Laat, W., 2012. Chromatin loops, gene positioning, and gene expression. *Frontiers in Genetics*, 3.

Hornshøj, H., Nielsen, M., Sinnott-Armstrong, N., Świtnicki, M., Juul, M., Madsen, T., Sallari, R., Kellis, M., Ørntoft, T., Hobolth, A. and Pedersen, J., 2018. Pan-cancer screen for mutations in non-coding elements with conservation and cancer specificity reveals correlations with expression and survival. *npj Genomic Medicine*, 3(1).

Hou, C., Li, L., Qin, Z. and Corces, V., 2012. Gene Density, Transcription, and Insulators Contribute to the Partition of the *Drosophila* Genome into Physical Domains. *Molecular Cell*, 48(3), pp.471-484.

Hulsen, T., de Vlieg, J. and Alkema, W., 2008. BioVenn – a web application for the comparison and visualization of biological lists using area-proportional Venn diagrams. *BMC Genomics*, 9(1), p.488.

Johnson, S., Cubberley, G. and Bentley, D., 2009. Cotranscriptional Recruitment of the mRNA Export Factor Yra1 by Direct Interaction with the 3′ End Processing Factor Pcf11. *Molecular Cell*, 33(2), pp.215-226.

Jonkers, I., Kwak, H. and Lis, J., 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *eLife*, 3.

Kadauke, S. and Blobel, G., 2009. Chromatin loops in gene regulation. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1789(1), pp.17-25.

Kamieniarz-Gdula, K., Gdula, M., Panser, K., Nojima, T., Monks, J., Wiśniewski, J., Riepsaame, J., Brockdorff, N., Pauli, A. and Proudfoot, N., 2019. Selective Roles of Vertebrate PCF11 in Premature and Full-Length Transcript Termination. *Molecular Cell*, 74(1), pp.158-172.e9.

Kamieniarz-Gdula, K., Proudfoot, N.J. 2019. Transcriptional Control by Premature Termination: A Forgotten Mechanism. *Trends Genet*. 35(8) pp. 553-564.

Kang, W., Ha, K., Uhm, H., Park, K., Lee, J., Hohng, S. and Kang, C., 2020. Transcription reinitiation by recycling RNA polymerase that diffuses on DNA after releasing terminated RNA. *Nature Communications*, 11(1).

Kerpedjiev, P., Abdennur, N., Lekschas, F., McCallum, C., Dinkla, K., Strobelt, H., Luber, J., Ouellette, S., Azhir, A., Kumar, N., Hwang, J., Lee, S., Alver, B., Pfister, H., Mirny, L., Park, P. and Gehlenborg, N., 2018. HiGlass: web-based visual exploration and analysis of genome interaction maps. *Genome Biology*, 19(1).

Kharchenko, P., Tolstorukov, M. and Park, P., 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nature Biotechnology*, 26(12), pp.1351-1359.

Kieffer-Kwon, K., Tang, Z., Mathe, E., Qian, J., Sung, M., Li, G., Resch, W., Baek, S., Pruett, N., Grøntved, L., Vian, L., Nelson, S., Zare, H., Hakim, O., Reyon, D., Yamane, A., Nakahashi, H., Kovalchuk, A., Zou, J., Joung, J., Sartorelli, V., Wei, C., Ruan, X., Hager, G., Ruan, Y. and Casellas, R., 2013. Interactome Maps of Mouse Gene Regulatory Domains Reveal Basic Principles of Transcriptional Regulation. *Cell*, 155(7), pp.1507-1520.

Kim, A. and Dean, A., 2012. Chromatin loop formation in the β-globin locus and its role in globin gene transcription. *Molecules and Cells*, 34(1), pp.1-5.

Kosak, S. and Groudine, M., 2004. Form follows function: the genomic organization of cellular differentiation. *Genes & Development*, 18(12), pp.1371-1384.

Krivega, I. and Dean, A., 2012. Enhancer and promoter interactions—long distance calls. *Current Opinion in Genetics & Development*, 22(2), pp.79-85.

Kuehner, J., Pearson, E. and Moore, C., 2011. Unravelling the means to an end: RNA polymerase II transcription termination. *Nature Reviews Molecular Cell Biology*, 12(5), pp.283-294.

Kuipers, J., Thurnherr, T., Moffa, G., Suter, P., Behr, J., Goosen, R., Christofori, G. and Beerenwinkel, N., 2018. Mutational interactions define novel cancer subgroups. *Nature Communications*, 9(1).

Kurukuti, S., Tiwari, V., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W. and Ohlsson, R., 2006. CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. *Proceedings of the National Academy of Sciences*, 103(28), pp.10684-10689.

Langdon, E., Qiu, Y., Ghanbari Niaki, A., McLaughlin, G., Weidmann, C., Gerbich, T., Smith, J., Crutchley, J., Termini, C., Weeks, K., Myong, S. and Gladfelter, A., 2018. mRNA structure determines specificity of a polyQ-driven phase separation. *Science*, 360(6391), pp.922-927.

Lanzuolo, C., Roure, V., Dekker, J., Bantignies, F. and Orlando, V., 2007. Polycomb response elements mediate the formation of chromosome higher-order structures in the bithorax complex. *Nature Cell Biology*, 9(10), pp.1167-1174.

Lareau, C. and Aryee, M., 2016. *diffloop: Identifying differential DNA loops from chromatin topology data*. [online] Rpubs.com. Available at: <https://rpubs.com/caleblareau/diffloop_vignette> [Accessed March 2022].

Lareau, C. and Aryee, M., 2017. diffloop: a computational framework for identifying and analyzing differential DNA loops from sequencing data. *Bioinformatics*, 34(4), pp.672-674.

Lareau, C. and Aryee, M., 2021. *diffloop: Identifying differential DNA loops from chromatin topology data*. https://github.com/aryeelab/diffloop.

Lee, B., Wang, J., Cai, L., Kim, M., Namburi, S., Tjong, H., Feng, Y., Wang, P., Tang, Z., Abbas, A., Wei, C., Ruan, Y. and Li, S., 2020. ChIA-PIPE: A fully automated pipeline for comprehensive ChIA-PET data analysis and visualization. *Science Advances*, 6(28).

Lettice, L., Heaney, S., Purdie, L., Li, L., de Beer, P. and Oostra, B., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, 12(14), pp.1725-1735.

Li, D., Hsu, S., Purushotham, D., Sears, R. and Wang, T., 2019. WashU Epigenome Browser update 2019. *Nucleic Acids Research*, 47(W1), pp.W158-W165.

Li, G., Ruan, X., Auerbach, R., Sandhu, K., Zheng, M., Wang, P., Poh, H., Goh, Y., Lim, J., Zhang, J., Sim, H., Peh, S., Mulawadi, F., Ong, C., Orlov, Y., Hong, S., Zhang, Z., Landt, S., Raha, D., Euskirchen, G., Wei, C., Ge, W., Wang, H., Davis, C., Fisher-Aylor, K., Mortazavi, A., Gerstein, M., Gingeras, T., Wold, B., Sun, Y., Fullwood, M., Cheung, E., Liu, E., Sung, W., Snyder, M. and Ruan, Y., 2012. Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. *Cell*, 148(1-2), pp.84-98.

Li, H. and Durbin, R., 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics*, 25(14), pp.1754-1760.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G. and Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078-2079.

Li, X., Luo, O., Wang, P., Zheng, M., Wang, D., Piecuch, E., Zhu, J., Tian, S., Tang, Z., Li, G. and Ruan, Y., 2017. Long-read ChIA-PET for base-pair-resolution mapping of haplotype-specific chromatin interactions. *Nature Protocols*, 12(5), pp.899-915.

Licatalosi, D., Geiger, G., Minet, M., Schroeder, S., Cilli, K., McNeil, J. and Bentley, D., 2002. Functional Interaction of Yeast Pre-mRNA 3′ End Processing Factors with RNA Polymerase II. *Molecular Cell*, 9(5), pp.1101-1111.

Lieberman-Aiden, E., van Berkum, N., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B., Sabo, P., Dorschner, M., Sandstrom, R., Bernstein, B., Bender, M., Groudine, M., Gnirke, A., Stamatoyannopoulos, J., Mirny, L., Lander, E. and Dekker, J., 2009. Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, 326(5950), pp.289-293.

Lomvardas, S., Barnea, G., Pisapia, D., Mendelsohn, M., Kirkland, J. and Axel, R., 2006. Interchromosomal Interactions and Olfactory Receptor Choice. *Cell*, 126(2), pp.403-413.

Lord, C., Thomas, G. and Brown, J., 2013. Mammalian alpha beta hydrolase domain (ABHD) proteins: Lipid metabolizing enzymes at the interface of cell signaling and energy metabolism. *Biochimica et Biophysica Acta (BBA) - Molecular and Cell Biology of Lipids*, 1831(4), pp.792-802.

Lyu, X., Rowley, M. and Corces, V., 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.

Maass, P., Rump, A., Schulz, H., Stricker, S., Schulze, L., Platzer, K., Aydin, A., Tinschert, S., Goldring, M., Luft, F. and Bähring, S., 2012. A misplaced lncRNA causes brachydactyly in humans. *Journal of Clinical Investigation*, 122(11), pp.3990-4002.

Madden, T., 2002. Chapter 16 The BLAST Sequence Analysis Tool. In: *The NCBI Handbook*. [online] Available at: <https://www.ncbi.nlm.nih.gov/books/NBK21101/> [Accessed March 2022].

Maharana, S., Iyer, K., Jain, N., Nagarajan, M., Wang, Y. and Shivashankar, G., 2016. Chromosome intermingling—the physical basis of chromosome organization in differentiated cells. *Nucleic Acids Research*, 44(11), pp.5148-5160.

Maharana, S., Wang, J., Papadopoulos, D., Richter, D., Pozniakovsky, A., Poser, I., Bickle, M., Rizk, S., Guillén-Boixet, J., Franzmann, T., Jahnel, M., Marrone, L., Chang, Y., Sterneckert, J., Tomancak, P., Hyman, A. and Alberti, S., 2018. RNA buffers the phase separation behavior of prion-like RNA binding proteins. *Science*, 360(6391), pp.918-921.

Mahmoudi, T., Katsani, K. and Verrijzer, C., 2002. GAGA can mediate enhancer function in trans by linking two separate DNA molecules. *The EMBO Journal*, 21(7), pp.1775-1781.

Mahy, N., Perry, P. and Bickmore, W., 2002. Gene density and transcription influence the localization of chromatin outside of chromosome territories detectable by FISH. *Journal of Cell Biology*, 159(5), pp.753-763.

Mahy, N., Perry, P., Gilchrist, S., Baldock, R. and Bickmore, W., 2002. Spatial organization of active and inactive genes and noncoding DNA within chromosome territories. *Journal of Cell Biology*, 157(4), pp.579-589.

Mapendano, C., Lykke-Andersen, S., Kjems, J., Bertrand, E. and Jensen, T., 2010. Crosstalk between mRNA 3′ End Processing and Transcription Initiation. *Molecular Cell*, 40(3), pp.410-422.

Marsden, M. and Laemmli, U., 1979. Metaphase chromosome structure: Evidence for a radial loop model. *Cell*, 17(4), pp.849-858.

Martelli, A., Falcieri, E., Zweyer, M., Bortul, R., Tabellini, G., Cappellini, A., Cocco, L. and Manzoli, L., 2002. The controversial nuclear matrix: a balanced point of view. *Histol Histopathol.*, 17(4), pp.1193-205.

Mayfield, J., Burkholder, N. and Zhang, Y., 2016. Dephosphorylating eukaryotic RNA polymerase II. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, 1864(4), pp.372-387.

McStay, B., 2016. Nucleolar organizer regions: genomic ‘dark matter’ requiring illumination. *Genes & Development*, 30(14), pp.1598-1610.

Meaburn, K., Gudla, P., Khan, S., Lockett, S. and Misteli, T., 2009. Disease-specific gene repositioning in breast cancer. *Journal of Cell Biology*, 187(6), pp.801-812.

Monahan, K., Horta, A., Mumbay-Wafula, A., Li, L., Zhao, Y., Love, P. and Lomvardas, S., 2018. Ldb1 mediates trans enhancement in mammals.

Monahan, K., Schieren, I., Cheung, J., Mumbey-Wafula, A., Monuki, E. and Lomvardas, S., 2017. Cooperative interactions enable singular olfactory receptor expression in mouse olfactory neurons. *eLife*, 6.

Muse, G., Gilchrist, D., Nechaev, S., Shah, R., Parker, J., Grissom, S., Zeitlinger, J. and Adelman, K., 2007. RNA polymerase is poised for activation across the genome. *Nature Genetics*, 39(12), pp.1507-1511.

Nagano, T., Lubling, Y., Stevens, T., Schoenfelder, S., Yaffe, E., Dean, W., Laue, E., Tanay, A. and Fraser, P., 2014. Single-cell Hi-C reveals cell-to-cell variability in chromosome structure. *Nature*, 502(7469), pp.59-64.

Nagano, T., Várnai, C., Schoenfelder, S., Javierre, B., Wingett, S. and Fraser, P., 2015. Comparison of Hi-C results using in-solution versus in-nucleus ligation. *Genome Biology*, 16(1).

Naumova, N., Imakaev, M., Fudenberg, G., Zhan, Y., Lajoie, B., Mirny, L. and Dekker, J., 2013. Organization of the Mitotic Chromosome. *Science*, 342(6161), pp.948-953.

Ni, Z., Saunders, A., Fuda, N., Yao, J., Suarez, J., Webb, W. and Lis, J., 2007. P-TEFb Is Critical for the Maturation of RNA Polymerase II into Productive Elongation In Vivo. *Molecular and Cellular Biology*, 28(3), pp.1161-1170.

Noble, C., Hollingworth, D., Martin, S., Ennis-Adeniran, V., Smerdon, S., Kelly, G., Taylor, I. and Ramos, A., 2005. Key features of the interaction between Pcf11 CID and RNA polymerase II CTD. *Nature Structural & Molecular Biology*, 12(2), pp.144-151.

Nora, E., Lajoie, B., Schulz, E., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J. and Heard, E., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature*, 485(7398), pp.381-385.

Norton, H., Emerson, D., Huang, H., Kim, J., Titus, K., Gu, S., Bassett, D. and Phillips-Cremins, J., 2018. Detecting hierarchical genome folding with network modularity. *Nature Methods*, 15(2), pp.119-122.

Nuebler, J., Fudenberg, G., Imakaev, M., Abdennur, N. and Mirny, L., 2018. Chromatin Organization by an Interplay of Loop Extrusion and Compartmental Segregation. *Biophysical Journal*, 114(3), p.30a.

Ogorodnikov, A., Levin, M., Tattikota, S., Tokalov, S., Hoque, M., Scherzinger, D., Marini, F., Poetsch, A., Binder, H., Macher-Göppinger, S., Probst, H., Tian, B., Schaefer, M., Lackner, K., Westermann, F. and Danckwardt, S., 2018. Transcriptome 3′end organization by PCF11 links alternative polyadenylation to formation and neuronal differentiation of neuroblastoma. *Nature Communications*, 9(1).

O'Sullivan, J., Tan-Wong, S., Morillon, A., Lee, B., Coles, J., Mellor, J. and Proudfoot, N., 2004. Gene loops juxtapose promoters and terminators in yeast. *Nature Genetics*, 36(9), pp.1014-1018.

Pal, K., Forcato, M. and Ferrari, F., 2018. Hi-C analysis: from data generation to integration. *Biophysical Reviews*, 11(1), pp.67-78.

Petrascheck, M., Escher, D., Mahmoudi, T., Verrijzer, C., Schaffner, W. and Barberis, A., 2005. DNA looping induced by a transcriptional enhancer in vivo. *Nucleic Acids Research*, 33(12), pp.3743-3750.

Phillips-Cremins, J., Sauria, M., Sanyal, A., Gerasimova, T., Lajoie, B., Bell, J., Ong, C., Hookway, T., Guo, C., Sun, Y., Bland, M., Wagstaff, W., Dalton, S., McDevitt, T., Sen, R., Dekker, J., Taylor, J. and Corces, V., 2013. Architectural Protein Subclasses Shape 3D Organization of Genomes during Lineage Commitment. *Cell*, 153(6), pp.1281-1295.

Porrua, O., Boudvillain, M. and Libri, D., 2016. Transcription Termination: Variations on Common Themes. *Trends in Genetics*, 32(8), pp.508-522.

Protter, D., Rao, B., Van Treeck, B., Lin, Y., Mizoue, L., Rosen, M. and Parker, R., 2018. Intrinsically Disordered Regions Can Contribute Promiscuous Interactions to RNP Granule Assembly. *Cell Reports*, 22(6), pp.1401-1412.

Proudfoot, N., 2016. Transcriptional termination in mammals: Stopping the RNA polymerase II juggernaut. *Science*, 352(6291), pp.aad9926-aad9926.

Ptashne, M., 1986. Gene regulation by proteins acting nearby and at a distance. *Nature*, 322(6081), pp.697-701.

Ramani, V., Deng, X., Qiu, R., Gunderson, K., Steemers, F., Disteche, C., Noble, W., Duan, Z. and Shendure, J., 2017. Massively multiplex single-cell Hi-C. *Nature Methods*, 14(3), pp.263-266.

Ramírez, F., Bhardwaj, V., Arrigoni, L., Lam, K., Grüning, B., Villaveces, J., Habermann, B., Akhtar, A. and Manke, T., 2018. High-resolution TADs reveal DNA sequences underlying genome organization in flies. *Nature Communications*, 9(1).

Ramírez, F., Ryan, D., Grüning, B., Bhardwaj, V., Kilpert, F., Richter, A., Heyne, S., Dündar, F. and Manke, T., 2016. deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Research*, 44(W1), pp.W160-W165.

Rao, S., Huntley, M., Durand, N., Stamenova, E., Bochkov, I., Robinson, J., Sanborn, A., Machol, I., Omer, A., Lander, E. and Aiden, E.,

2014. A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, 159(7), pp.1665-1680.

Reiff, S., Schroeder, A., Kirli, K., Cosolo, A., Bakker, C., Mercado, L., Lee, S., Veit, A., Balashov, A., Vitzthum, C., Ronchetti, W., Pitman, K., Johnson, J., Ehmsen, S., Kerpedjiev, P., Abdennur, N., Imakaev, M., Öztürk, S., Çamoğlu, U., Mirny, L., Gehlenborg, N., Alver, B. and Park, P., 2021. The 4D Nucleome Data Portal: a resource for searching and visualizing curated nucleomics data.

Rheinbay, E., Nielsen, M., Abascal, F., Tiao, G., Hornshøj, H., Hess, J., Pedersen, R., Feuerbach, L., Sabarinathan, R., Madsen, T., Kim, J., Mularoni, L., Shuai, S., Lanzós, A., Herrmann, C., Maruvka, Y., Shen, C., Amin, S., Bertl, J., Dhingra, P., Diamanti, K., Gonzalez-Perez, A., Guo, Q., Haradhvala, N., Isaev, K., Juul, M., Komorowski, J., Kumar, S., Lee, D., Lochovsky, L., Liu, E., Pich, O., Tamborero, D., Umer, H., Uusküla-Reimand, L., Wadelius, C., Wadi, L., Zhang, J., Boroevich, K., Carlevaro-Fita, J., Chakravarty, D., Chan, C., Fonseca, N., Hamilton, M., Hong, C., Kahles, A., Kim, Y., Lehmann, K., Johnson, T., Kahraman, A., Park, K., Saksena, G., Sieverling, L., Sinnott-Armstrong, N., Campbell, P., Hobolth, A., Kellis, M., Lawrence, M., Raphael, B., Rubin, M., Sander, C., Stein, L., Stuart, J., Tsunoda, T., Wheeler, D., Johnson, R., Reimand, J., Gerstein, M., Khurana, E., López-Bigas, N., Martincorena, I., Pedersen, J. and Getz, G., 2017. Discovery and characterization of coding and non-coding driver mutations in more than 2,500 whole cancer genomes.

Rippe, K., 2021. Liquid–Liquid Phase Separation in Chromatin. *Cold Spring Harbor Perspectives in Biology*, 14(2), p.a040683.

Robinson, J., Thorvaldsdóttir, H., Winckler, W., Guttman, M., Lander, E., Getz, G. and Mesirov, J., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24-26.

Robinson, J., Turner, D., Durand, N., Thorvaldsdóttir, H., Mesirov, J. and Aiden, E., 2018. Juicebox.js Provides a Cloud-Based Visualization System for Hi-C Data. *Cell Systems*, 6(2), pp.256-258.e1.

Rodríguez-Carballo, E., Lopez-Delisle, L., Zhan, Y., Fabre, P., Beccari, L., El-Idrissi, I., Huynh, T., Ozadam, H., Dekker, J. and Duboule, D., 2017. TheHoxDcluster is a dynamic and resilient TAD boundary controlling the segregation of antagonistic regulatory landscapes. *Genes & Development*, 31(22), pp.2264-2281.

Rowley, M., Nichols, M., Lyu, X., Ando-Kuri, M., Rivera, I., Hermetz, K., Wang, P., Ruan, Y. and Corces, V., 2017. Evolutionarily Conserved Principles Predict 3D Chromatin Organization. *Molecular Cell*, 67(5), pp.837-852.e7.

Sadowski, M., Dichtl, B., Hubner, W. and Keller, W., 2003. Independent functions of yeast Pcf11p in pre-mRNA 3' end processing and in transcription termination. *The EMBO Journal*, 22(9), pp.2167-2177.

Sainsbury, S., Bernecky, C. and Cramer, P., 2015. Structural basis of transcription initiation by RNA polymerase II. *Nature Reviews Molecular Cell Biology*, 16(3), pp.129-143.

Saiz, L. and Vilar, J., 2006. DNA looping: the consequences and its control. *Current Opinion in Structural Biology*, 16(3), pp.344-350.

Schindelin, J., Rueden, C., Hiner, M. and Eliceiri, K., 2015. The ImageJ ecosystem: An open platform for biomedical image analysis. *Molecular Reproduction and Development*, 82(7-8), pp.518-529.

Schuettengruber, B., Oded Elkayam, N., Sexton, T., Entrevan, M., Stern, S., Thomas, A., Yaffe, E., Parrinello, H., Tanay, A. and Cavalli, G., 2014. Cooperativity, Specificity, and Evolutionary Stability of Polycomb Targeting in *Drosophila*. *Cell Reports*, 9(1), pp.219-233.

Schwalb, B., Michel, M., Zacher, B., Frühauf, K., Demel, C., Tresch, A., Gagneur, J. and Cramer, P., 2016. TT-seq maps the human transient transcriptome. *Science*, 352(6290), pp.1225-1228.

Schwarzer, W., Abdennur, N., Goloborodko, A., Pekowska, A., Fudenberg, G., Loe-Mie, Y., Fonseca, N., Huber, W., Haering, C., Mirny, L. and Spitz, F., 2017. Two independent modes of chromatin organization revealed by cohesin removal. *Nature*, 551(7678), pp.51-56.

Segal, E., Fondufe-Mittendorf, Y., Chen, L., Thåström, A., Field, Y., Moore, I., Wang, J. and Widom, J., 2006. A genomic code for nucleosome positioning. *Nature*, 442(7104), pp.772-778.

Sexton, T., Bantignies, F. and Cavalli, G., 2009. Genomic interactions: Chromatin loops and gene meeting points in transcriptional regulation. *Seminars in Cell & Developmental Biology*, 20(7), pp.849-855.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A. and Cavalli, G., 2012. Three-Dimensional Folding and Functional Organization Principles of the *Drosophila* Genome. *Cell*, 148(3), pp.458-472.

Shandilya, J. and Roberts, S., 2012. The transcription cycle in eukaryotes: From productive initiation to RNA polymerase II recycling. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(5), pp.391-400.

Shen, Y., Yue, F., McCleary, D., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. and Ren, B., 2012. A map of the cis-regulatory sequences in the mouse genome. *Nature*, 488(7409), pp.116-120.

Shi, Y. and Manley, J., 2015. The end of the message: multiple protein–RNA interactions define the mRNA polyadenylation site. *Genes & Development*, 29(9), pp.889-897.

Shi, Y., Di Giammartino, D., Taylor, D., Sarkeshik, A., Rice, W., Yates, J., Frank, J. and Manley, J., 2009. Molecular Architecture of the Human Pre-mRNA 3′ Processing Complex. *Molecular Cell*, 33(3), pp.365-376.

Singh, B. and Hampsey, M., 2007. A Transcription-Independent Role for TFIIB in Gene Looping. *Molecular Cell*, 27(5), pp.806-816.

Singh, B., Ansari, A. and Hampsey, M., 2009. Detection of gene loops by 3C in yeast. *Methods*, 48(4), pp.361-367.

Song, S., Hou, C. and Dean, A., 2007. A Positive Role for NLI/Ldb1 in Long-Range β-Globin Locus Control Region Function. *Molecular Cell*, 28(5), pp.810-822.

Spilianakis, C. and Flavell, R., 2004. Long-range intrachromosomal interactions in the T helper type 2 cytokine locus. *Nature Immunology*, 5(10), pp.1017-1027.

Spilianakis, C., Lalioti, M., Town, T., Lee, G. and Flavell, R., 2005. Interchromosomal associations between alternatively expressed loci. *Nature*, 435(7042), pp.637-645.

Splinter, E., Heath, H., Kooren, J., Palstra, R., Klous, P., Grosveld, F., Galjart, N. and de Laat, W., 2006. CTCF mediates long-range chromatin looping and local histone modification in the beta-globin locus. *Blood Cells, Molecules, and Diseases*, 38(2), p.178.

Stevens, T., Lando, D., Basu, S., Atkinson, L., Cao, Y., Lee, S., Leeb, M., Wohlfahrt, K., Boucher, W., O’Shaughnessy-Kirwan, A., Cramard, J., Faure, A., Ralser, M., Blanco, E., Morey, L., Sansó, M., Palayret, M., Lehner, B., Di Croce, L., Wutz, A., Hendrich, B., Klenerman, D. and Laue, E., 2017. 3D structures of individual mammalian genomes studied by single-cell Hi-C. *Nature*, 544(7648), pp.59-64.

Sun, H., Shen, J. and Yokota, H., 2000. Size-Dependent Positioning of Human Chromosomes in Interphase Nuclei. *Biophysical Journal*, 79(1), pp.184-190.

Taberlay, P., Achinger-Kawecka, J., Lun, A., Buske, F., Sabir, K., Gould, C., Zotenko, E., Bert, S., Giles, K., Bauer, D., Smyth, G., Stirzaker, C., O'Donoghue, S. and Clark, S., 2016. Three-dimensional disorganization of the cancer genome occurs coincident with long-range genetic and epigenetic alterations. *Genome Research*, 26(6), pp.719-731.

Tang, Z., Luo, O., Li, X., Zheng, M., Zhu, J., Szalaj, P., Trzaskoma, P., Magalska, A., Wlodarczyk, J., Ruszczycki, B., Michalski, P., Piecuch, E., Wang, P., Wang, D., Tian, S., Penrad-Mobayed, M., Sachs, L., Ruan, X., Wei, C., Liu, E., Wilczynski, G., Plewczynski, D., Li, G. and Ruan, Y., 2015. CTCF-Mediated Human 3D Genome Architecture Reveals Chromatin Topology for Transcription. *Cell*, 163(7), pp.1611-1627.

Therizols, P., Illingworth, R., Courilleau, C., Boyle, S., Wood, A. and Bickmore, W., 2014. Chromatin decondensation is sufficient to alter nuclear organization in embryonic stem cells. *Science*, 346(6214), pp.1238-1242.

Tolhuis, B., Palstra, R., Splinter, E., Grosveld, F. and de Laat, W., 2002. Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell*, 10(6), pp.1453-1465.

Vakoc, C., Letting, D., Gheldof, N., Sawado, T., Bender, M., Groudine, M., Weiss, M., Dekker, J. and Blobel, G., 2005. Proximity among Distant Regulatory Elements at the β-Globin Locus Requires GATA-1 and FOG-1. *Molecular Cell*, 17(3), pp.453-462.

van der Lee, R., Buljan, M., Lang, B., Weatheritt, R., Daughdrill, G., Dunker, A., Fuxreiter, M., Gough, J., Gsponer, J., Jones, D., Kim, P., Kriwacki, R., Oldfield, C., Pappu, R., Tompa, P., Uversky, V., Wright, P. and Babu, M., 2014. Classification of Intrinsically Disordered Regions and Proteins. *Chemical Reviews*, 114(13), pp.6589-6631.

Van der Ploeg, L., Konings, A., Oort, M., Roos, D., Bernini, L. and Flavell, R., 1980. γ-β-Thalassaemia studies showing that deletion of the γ- and δ-genes influences β-globin gene expression in man. *Nature*, 283(5748), pp.637-642.

Volanakis, A., Kamieniarz-Gdula, K., Schlackow, M. and Proudfoot, N., 2017. WNK1 kinase and the termination factor PCF11 connect nuclear mRNA export with transcription. *Genes & Development*, 31(21), pp.2175-2185.

Volpi, E., Chevret, E., Jones, T., Vatcheva, R., Williamson, J., Beck, S., Campbell, R., Goldsworthy, M., Powis, S., Ragoussis, J., Trowsdale, J. and Sheer, D., 2000. Large-scale chromatin organization of the major histocompatibility complex and other regions of human chromosome 6 and its response to interferon in interphase nuclei. *Journal of Cell Science*, 113(9), pp.1565-1576.

Wagschal, A. 2012. Microprocessor, Setc, Xrn2, and Rrp6 co-operate to induce premature termination of transcription by RNAPII. Cell. 150 pp. 1147-1157.

Wang, P., Feng, Y., Zhu, K., Chai, H., Chang, Y., Yang, X., Liu, X., Shen, C., Gega, E., Lee, B., Kim, M., Ruan, X. and Ruan, Y., 2021. In situ Chromatin Interaction Analysis Using Paired-End Tag Sequencing. *Current Protocols*, 1(8).

West, S. and Proudfoot, N., 2007. Human Pcf11 enhances degradation of RNA polymerase II-associated nascent RNA and transcriptional termination. *Nucleic Acids Research*, 36(3), pp.905-914.

Williams, R., Broad, S., Sheer, D. and Ragoussis, J., 2002. Subchromosomal Positioning of the Epidermal Differentiation Complex (EDC) in Keratinocyte and Lymphoblast Interphase Nuclei. *Experimental Cell Research*, 272(2), pp.163-175.

Wolff, J., Bhardwaj, V., Nothjunge, S., Richard, G., Renschler, G., Gilsbach, R., Manke, T., Backofen, R., Ramírez, F. and Grüning, B., 2018. Galaxy HiCExplorer: a web server for reproducible Hi-C data analysis, quality control and visualization. *Nucleic Acids Research*, 46(W1), pp.W11-W16.

Wolff, J., Rabbani, L., Gilsbach, R., Richard, G., Manke, T., Backofen, R. and Grüning, B., 2020. Galaxy HiCExplorer 3: a web server for reproducible Hi-C, capture Hi-C and single-cell Hi-C data analysis, quality control and visualization. *Nucleic Acids Research*, 48(W1), pp.W177-W184.

Xu, J., Gu, W., Ji, K., Xu, Z., Zhu, H. and Zheng, W., 2018. Sequence analysis and structure prediction of ABHD16A and the roles of the ABHD family members in human disease. *Open Biology*, 8(5), p.180017.

Ye, C., Liu, G., Bremer, S. and Heng, H., 2007. The dynamics of cancer chromosomes and genomes. *Cytogenetic and Genome Research*, 118(2-4), pp.237-246.

Zenk, F., Zhan, Y., Kos, P., Löser, E., Atinbayeva, N., Schächtle, M., Tiana, G., Giorgetti, L. and Iovino, N., 2021. HP1 drives de novo 3D genome reorganization in early *Drosophila* embryos. *Nature*, 593(7858), pp.289-293.

Zhang, Y., Liu, T., Meyer, C., Eeckhoute, J., Johnson, D., Bernstein, B., Nusbaum, C., Myers, R., Brown, M., Li, W. and Liu, X., 2008. Model-based Analysis of ChIP-Seq (MACS). *Genome Biology*, 9(9).

Zhang, Z. and Gilmour, D., 2006. Pcf11 Is a Termination Factor in *Drosophila* that Dismantles the Elongation Complex by Bridging the CTD of RNA Polymerase II to the Nascent Transcript. *Molecular Cell*, 21(1), pp.65-74.

Zhang, Z., 2005. CTD-dependent dismantling of the RNA polymerase II elongation complex by the pre-mRNA 3'-end processing factor, Pcf11. *Genes & Development*, 19(13), pp.1572-1580.

Zheng, M., Tian, S., Capurso, D., Kim, M., Maurya, R., Lee, B., Piecuch, E., Gong, L., Zhu, J., Li, Z., Wong, C., Ngan, C., Wang, P., Ruan, X., Wei, C. and Ruan, Y., 2019. Multiplex chromatin interactions with single-molecule precision. *Nature*, 566(7745), pp.558-562.

Zhou, X., Maricque, B., Xie, M., Li, D., Sundaram, V., Martin, E., Koebbe, B., Nielsen, C., Hirst, M., Farnham, P., Kuhn, R., Zhu, J., Smirnov, I., Kent, W., Haussler, D., Madden, P., Costello, J. and Wang, T., 2011. The Human Epigenome Browser at Washington University. *Nature Methods*, 8(12), pp.989-990.

Zhu, L., Gazin, C., Lawson, N., Pagès, H., Lin, S., Lapointe, D. and Green, M., 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics*, 11(1).

Zufferey, M., Tavernari, D., Oricchio, E. and Ciriello, G., 2018. Comparison of computational methods for the identification of topologically associating domains. *Genome Biology*, 19(1).