THE UNIVERSITY
of EDINBURGH

UNIVERSITY OF EDINBURGH

DOCTORAL THESIS

Analysis, interpretation, and visualisation of DamID-seq experiments

---

Author:
James ASHMORE

Supervisor:
Prof. Keisuke KAJI
Dr. Simon TOMLINSON

A thesis submitted in fulfilment of the requirements for the degree of Doctor of Philosophy in the Biology of Reprogramming

May 13, 2019

Declaration of Authorship

I, James ASHMORE, declare that this thesis titled, "Analysis, interpretation, and visualisation of DamID-seq experiments" and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Signed:

Date:

Abstract

DNA adenine methyltransferase identification with sequencing (abbreviated DamID-seq) is a technique that can measure protein-DNA interactions in the genome. Unlike chromatin immunoprecipitation with sequencing (abbreviated ChIP-seq), this technique does not require validated antibodies, precipitation steps, or chemical cross-linking, and can be used with minimal numbers of cells. Although the technique was first developed in *Drosophila* nearly two decades ago, due to technical limitations only a handful of experiments using mammalian cells have been published. The optimisation of mammalian DamID-seq in our lab has highlighted the need to survey potential sources of bias, develop accurate analysis methods, and investigate the similarities and differences with ChIP-seq for detecting protein-DNA interactions. Here, I describe several variables that influence the accuracy of DamID-seq experiments, present the Daim software package (pronounced "Dime") for the comprehensive analysis of DamID-seq data, and assess the sensitivity and specificity of DamID-seq compared with competing techniques. In particular, I show that differences in the experimental procedure (polymerase usage and restriction digest) and features in the sequencing data (fragment length and nucleotide content) generate systematic bias and technical variation. I also demonstrate that DamID-seq data can be re-purposed to measure Dam-accessible DNA in the genome, comparable with other chromatin accessibility techniques (ATAC-seq, DNase-seq, and FAIRE-seq). To analyse DamID-seq data, I

developed the Daim software package which incorporates methods for preprocessing, normalisation, and identification of DNA binding and accessibility sites. Several options for functional and sequence analysis of results are also included. The use of Daim was demonstrated using data for transcription factors Oct4 and Sox2 in mouse embryonic stem cells, embryonic fibroblast cells, and neural stem cells from a range of cell numbers. Finally, I show that DNA binding and accessibility sites vary substantially between and within techniques, yet no clear reason for these differences has been detected, prompting careful consideration of any biological conclusions. These results show that Daim can be successfully used for the analysis, interpretation, and visualisation of DamID-seq experiments, and that to achieve comprehensive results, different techniques should be treated as complementary rather than competing.

Lay Summary

Cells are the building blocks of life. They form the structure of all animals and plants, from organs to tissues and everything within. There are different types of cells, each of which play a specialised role in life, such as producing energy or fighting disease. They contain a substance called DNA which provides a blueprint for building and maintaining the cell. All of the DNA in a cell is referred to as the genome. Genes are regions of the genome that are responsible for different characteristics, such as cell shape and size. Different types of cells are created by proteins binding to the genome and activating or deactivating specific genes. The key to understanding developmental biology is determining how and which genes are regulated by protein binding. To detect where proteins bind, a technique called Dam identification (abbreviated DamID) was developed. The protein of interest is tied to a protein known as Dam produced by bacteria. When the protein binds, Dam leaves a recognisable mark on the DNA nearby the binding site. The DNA is then fragmented and only the marked fragments are measured. To identify genome-wide binding sites, I have written a computer program called Daim (pronounced "Dime") which determines the location of the protein-bound fragments in the genome, and measures how often the DNA was marked. It then uses statistics to test the probability that this measurement represents actual binding and not just random variation. This information can then be used to check which genes are bound by the protein, providing a greater understanding of its role in defining a cell's characteristics.

Acknowledgements

I would like to thank my supervisors Prof Keisuke Kaji and Dr Simon Tomlinson for their support and advice during my project. Additionally, I would like to thank my committee members Prof Steven Pollard and Dr Abdenour Soufi for their comments and suggestions regarding my work. I would also like to thank my family and friends for their encouragement and reassurance amid stressful periods. Lastly, I would like to thank my fiancé and soon to be wife Alice for supporting me these nine years we have been together and for being my best friend.

Contents

List of Figures

List of Tables

List of Publications

Tosti, Luca et al. (Mar. 2018). "Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo". en. In: *Genome Res.*

Chapter 1

Introduction

1.1 Transcriptional regulation of gene expression

Recent initiatives such as the mouse cell atlas have demonstrated there are around 800 major cell types and potentially more than 1,000 subtypes in mammalian organisms (Han et al., 2018). Each of these cell types exhibit divergent gene expression profiles which drive the morphological and physiological characteristics underlying their functionality. The central dogma of molecular biology is that genes encode for functional products called proteins which perform specific roles within the cell such as carrying out chemical reactions or transmitting signals. Gene expression profiles therefore comprise large networks of genes and proteins which are co-regulated by deterministic mechanisms to achieve the cellular diversity and functionality requisite of mammalian complexity. The amazing repertoire on display is even more impressive when one considers that all of these cells contain the same genetic information, only it is read and parsed in a completely different manner depending on environmental and developmental cues. Gene expression itself can be controlled at the level of transcription (DNA into mRNA) and translation (mRNA into protein) including other downstream modifications (Cooper, 2000). Transcriptional regulation is the primary

form of control in mammalian systems and is the combination of multiple protein-protein and protein-DNA interactions, most notably transcription factor binding and chromatin structure remodelling (Wilkinson, Nakauchi, and Göttgens, 2017).

1.1.1 Transcription factor binding

Transcription factors are DNA-binding proteins involved in controlling the rate at which DNA is transcribed into mRNA via the enzyme RNA polymerase (Latchman, 1993). They bind to specific DNA sequences along the genome, often within promoters and enhancers which modulate transcription of the corresponding gene. Binding specifically leads to recruitment of activator or repressor proteins which influence DNA folding at the promoter making it easier or harder for RNA polymerase to bind and form a pre-initiation complex leading to transcription (Haberle and Stark, 2018). Genome sequence analyses have suggested there are roughly 1,300 to 1,500 transcriptions factors encoded in the mammalian genome (Vaquerizas et al., 2009). These have been grouped into families characterised by shared DNA-binding and regulatory domains which underlie their sequence specificity and transcriptional influence, respectively.

There are a number of ways in which transcription factors can be classified. One way is by grouping them on the basis of their DNA binding domains which differ in their sequence and structure (Latchman, 1997). Examples of these include Homeobox, POU, Paired Box, Zinc Finger, Basic Element, and Ets domains. Those which share the 60 amino acid domain called the homeodomain are aptly referred to as homeobox proteins. The most studied of these are the hox genes, a cluster of eight genes which code for transcription factors involved in patterning of the principal body axis during embryonic development. They are particularly intriguing because they appear

one after another in the order in which they are expressed which matches their function along the anterior-posterior body (Garcia-Fernàndez, 2005). The POU family of transcription factors share a 150-160 amino acid domain first defined in Pit-1, Oct-1, and Unc-86 proteins, hence the abbreviation. Collectively these members have been shown to play an important role in development and functioning of the nervous system (Latchman, 1999). Another class of transcription factors are those which contain a zinc-finger domain, of which there are many different structures (coordinated by a zinc ion) depending on what combination of cysteine and histidine amino acids are incorporated. Members of this family can bind to DNA, RNA, and even other proteins. As such, they are involved in regulation of multiple processes, such as cell proliferation in skin, muscle differentiation, and even adipose generation (Cassandri et al., 2017).

Instead of classifying transcription factors by their structural domains, an alternative method is to classify them based upon four functional classes (Pope and Medzhitov, 2018): The first contains transcription factors which are expressed in all cell types and are responsible for regulating ubiquitously expressed genes. The enzyme GAPDH is required for glycolysis and its expression is continuously regulated by the transcription factor HIF-1 (Said et al., 2009). The second consists of pre-existing proteins (already translated) that are activated when induced. The most well-studied of these is the p53 tumour suppressor protein, which activates upon cellular stress and coordinates a stress response program through transcriptional activation (Zilfou and Lowe, 2009). By contrast, the third comprises of proteins that first require translation when induced. A prime example of this is the c-Myc oncogene which becomes expressed in response to certain receptor transduction pathways such as MAPK/ERK and WNT signalling (Dang, 2012).The last class of transcription factors are lineage-restricted and

responsible for regulating and maintaining cell-type specific morphology and function. Notable examples include MyoD which regulates muscle differentiation (Bar-Nur et al., 2018) and Sox2 which maintains neuronal progenitor identity (Kim et al., 2011). Given the complexity of transcription factor families and the potential for members to be reclassified based upon new evidence of novel function, classification by functional class may be rather vague relative to classification by domain architecture, which is generally more fixed. These different classifications are also liable to overlap in that a particular functional class of transcription factors may only be made up of proteins with a particular functional domain. For example, approximately 49% of transcription factor families in humans are tissue-specific, reflecting their specific physiological functions. However, C2HS zinc-finger proteins for example are much less tissue-specific, instead acting to repress transposable elements which occur in a broad range of tissues and cell types (Lambert et al., 2018). Overall, the role of transcription factors therefore is to make sure genes are expressed in the right cell at the right time in response to developmental and environmental signalling cues.

1.1.2 Chromatin structure remodelling

Eukaryotic DNA is wrapped around histone proteins to form a hierarchically packaged structure called chromatin. The basic unit of chromatin is the nucleosome. It consists of a 146 bp segment of DNA wrapped around a core of eight histone proteins (includes two copies of H2A, H2B, H3, and H4). Multiple nucleosomes are arranged along a DNA molecule (separated by 38-53 bp segments of linker DNA) to form a 10 nm chromatin fibre, otherwise known as a nucleosomal array. Under a microscope these fibres appear like "beads-on-a-string" and are often considered the 'primary' structure of chromatin (Luger, Dechassa, and Tremethick, 2012). Short-range interactions between nucleosomes in the same fibre create folds which generate a thicker 30 nm chromatin fibre. This high-order structure is considered the 'secondary' structure

of chromatin (Tremethick, 2007). The exact folding mechanism however is unclear, with two prevailing theories still under debate: The solenoid model predicts that consecutive nucleosomes interact and fold the fibre into a one-start helix structure (Kornberg, 1974). The zigzag model suggests that alternative nucleosomes interact and fold the fibre into a two-start helix structure (Woodcock, Frado, and Rattner, 1984). Whilst both models have considerable supporting evidence, a recent review has suggested that chromatin structure is more of a continuum of various states and that a model combining both mechanisms is required (Luger, Dechassa, and Tremethick, 2012). Long-range interactions between individual nucleosomes on different fibres provide additional folding to form aggregates of fibres, also defined as the 'tertiary' structure of chromatin. These aggregates then undergo successive rounds of coiling and looping (driven by DNA-binding proteins such as cohesion and CTCF) to eventually form condensed chromatin complexes called chromosomes which pair up and fit inside the nucleus of the cell (Merkenschlager and Nora, 2016). Both the 'secondary' and 'tertiary' structures are stabilised by architectural proteins (e.g. HP1, HMG1, and PARP1) and influenced by histone variants (e.g. H2A.X, HSB1, and H3.3) and post-translational modifications (e.g. acetylation, methylation, and phosphorylation) (Venkatesh and Workman, 2015). Additionally, variations in the 'primary' structure reflecting different nucleosomal states can lead to large-scale rearrangements in the higher-order structures. These rearrangements are required at particular times throughout the lifetime of the cell, for example in the metaphase stage of mitosis where constrained and non-random structures are required. Finally, chromosomes within the nucleus are organised into particular territories which together constitute the 'nuclear' architecture of the cell. Their organization however is not stable as during the cell cycle chromosomes are required to decondense and move to accommodate replication and transfer of DNA to daughter cells. Generally, chromosomes during interphase and prophase appear in similar locations within the nucleus compared to the metaphase, anaphase,

and telophase stages of mitosis (Cremer and Cremer, 2010). When DNA is compacted, it becomes very difficult for transcription factors and RNA polymerase to bind to the promoter region of a gene to initiate transcription. Genes which are expressed are subsequently found in more accessible regions of DNA than genes which are suppressed (Tsompana and Buck, 2014). The structure of chromatin however is not static and can be rearranged in order to provide access to particular genomic regions. The attraction between DNA and histones can be weakened by modifying the N-terminal tails of histones using acetylation, phosphorylation, and methylation. Acetylation in particular reduces the positive charge of the histones so the negatively charged DNA becomes less tightly wrapped (Görisch et al., 2005). Specific histone proteins can also be replaced by variants (the most common being H2A.X, H2A.Z, and H3.3) which accumulate permissive histone modifications and disrupt nucleosome stability (Henikoff and Smith, 2015). Nucleosomes can additionally be evicted or pushed along the DNA by specific ATP-dependent remodelling proteins thus providing access to the promoters and enhancers of requisite genes (Flaus and Owen-Hughes, 2011). The structure of chromatin therefore presents an additional layer of transcriptional regulation which acts in tandem with transcription factor binding that requires its own level of control from upstream signalling pathways.

1.2 Identification of genome-wide DNA accessibility sites

By mapping regions of the genome which are accessible, information regarding nucleosome positioning and transcription factor occupancy can be generated. Together this information can be integrated to reveal the location of regulatory elements involved in gene expression which can be further studied. The first high-throughput method to map chromatin accessibility was called DNase-seq and is based upon the sensitivity of unbound DNA to cleavage by DNase I nuclease (Crawford et al., 2006).

A library of hypersensitive fragments is generated and sequenced, then mapped back to the genome to identify their original location relative to known genomic features. Although this method has proved highly successful and is still continuously used in research, a number of successive methods were developed to overcome limitations in the original protocol. The method known as FAIRE-seq exploits the fact that nucleosome-depleted regions of the genome are less susceptible to formaldehyde cross-linking, thus after phenol-chloroform extraction these regions are preferentially segregated from nucleosome-bound DNA into the aqueous phase of the solution (Giresi et al., 2007). Unlike DNase-seq, this method does not require permeabilization of the cells or isolation of the nuclei, meaning the experimental protocol is often simpler and can be performed more quickly. Despite this advantage, it has since been shown that FAIRE-seq performs worse than DNase-seq and is often less reproducible (Song et al., 2011). Furthermore, both methods require a large number of cells to achieve a sufficient DNA yield for high-throughput sequencing. The most recently developed method called ATAC-seq was designed to overcome this limitation and is based upon the ability of Tn5 transposase to preferentially insert into open chromatin regions (Buenrostro et al., 2013). The hyperactive Tn5 transposase is first loaded with sequencing adapters and then delivered into cells containing unfixed nuclei to simultaneously fragment and tag the genome with adapters for sequencing. This bypasses the usual inefficient library preparation steps allowing fewer cells to be harvested and more experiments to be performed in biological systems which are limited. Together these methods constitute an array of different approaches to epigenomic profiling and whilst all of them are still used in research, it is becoming clear that the accuracy achieved by ATAC-seq combined with the resolution of single-cell technology is important to achieve a deeper understanding of regulatory elements (Buenrostro et al., 2015).

1.3 Identification of genome-wide DNA binding sites

The study of gene regulation relies on the ability to detect protein-DNA interactions, such as RNA polymerase and transcription factor binding. By characterising which genes are regulated by what proteins, more of the transcriptional network can be pieced together to understand how the myriad of interactions bring about the observed biological complexity (Wilkinson, Nakauchi, and Göttgens, 2017). This endeavour is particularly relevant to understanding cell fate control by transcription factors (Iwafuchi-Doi and Zaret, 2016) and dysregulation of gene expression in cancer (Nebert, 2002). Over the last decade a variety of technologies have been used to identify genome-wide DNA binding sites for transcription factors and other proteins. What follows is an historical overview of their development and current limitations which need to be addressed:

1.3.1 Chromatin immunoprecipitation (ChIP)

All of the technology discussed in this section originate from a technique first described in 1984 by Gilmour and Lis named chromatin immunoprecipitation (Gilmour and Lis, 1984). The authors determined the distribution of RNA polymerase in both bacteria and *Drosophila* by covalently joining proteins to DNA using UV light irradiation, followed by precipitation of a protein antigen out of solution using an antibody that specifically binds to the protein. The eluted DNA fragments in the precipitate were then purified and subjected to hybridization assays in order to identify which DNA sequences were associated *in vivo* with the protein (Gilmour and Lis, 1984; Gilmour and Lis, 1985). After the publication of this technique, many researchers began to adopt it for their own systems and shortly after in 1988 the relatively weak UV fixation was switched to a much stronger formaldehyde fixation in order to prevent loss of genuine but weak DNA-protein complexes during sample preparation

(Solomon, Larsen, and Varshavsky, 1988). Furthermore, with the advent of microarray technology the relatively low-throughput hybridization assays were replaced with whole genome microarrays allowing researchers to measure binding at an unprecedented resolution. This new approach (ChIP-chip) was successfully demonstrated by Ren and colleagues who measured the binding of Gal4 and Ste12 activator proteins in yeast, showing genes whose expression is directly controlled by both proteins in response to changes in carbon source and mating pheromone, respectively (Ren et al., 2000).

1.3.2 Chromatin immunoprecipitation sequencing (ChIP-seq)

Although first described in 2007, the current most popular method to identify genome-wide DNA binding sites is chromatin immunoprecipitation sequencing (ChIP-seq). The concept behind ChIP-seq is to selectively enrich for regions of DNA bound to the protein of interest by first targeting an antibody against an epitope on the protein. The antibody binds to the epitope and the protein-DNA complexes are co-precipitated out of the cellular solution using the antibody. The complexes can then be disassociated leaving just a collection of DNA fragments which are sequenced to identify the genome-wide DNA binding sites (Johnson et al., 2007). In reality, the protocol behind ChIP-seq is much more involved and each step requires consideration. The ChIP-seq protocol starts by treating cells with formaldehyde to cross-link any protein-DNA interactions. This ensures that bound proteins remain fixed to their DNA sequences during chemical handling. The cells are then homogenised, and the chromatin is sheared by sonication. This is used to achieve a greater binding site resolution by reducing the size of the DNA region protruding out from the protein-DNA complex. The protein-DNA complexes are then immunoprecipitated using a highly specific and sensitive antibody. The antibody must target an exposed epitope on the protein or else it cannot bind to the protein-DNA complex. The cross-linking is then reversed, and the

DNA fragments are prepared for high-throughput sequencing (Johnson et al., 2007). In order to identify genome-wide DNA binding sites from ChIP-seq data, the sequencing reads are first mapped to a reference genome using alignment software. Regions of high read coverage in the ChIP sample compared to an input DNA sample (used to correct for non-specific immunoprecipitation) are then located using peak calling software (Pepke, Wold, and Mortazavi, 2009).

Whilst ChIP-seq has been used extensively over the last decade, it suffers from a number of inherent limitations: The quality of the ChIP-seq data is determined by the specificity and sensitivity of the antibody. If a non-specific antibody is used a significant amount of unrelated protein-DNA complexes are immunoprecipitated and sequenced (Kidder, Hu, and Zhao, 2011). The library preparation method also contains a number of inefficient enzymatic steps resulting in sample loss, therefore at least 10^7 cells are needed to generate enough immunoprecipitated DNA for sequencing (Gilfillan et al., 2012). Whilst this number of cells can be achieved by expanding cell lines *in vitro*, it is very difficult to collect enough cells *in vivo* from tissues and rare-cell populations. Lastly, using formaldehyde to cross-link protein-DNA interactions includes its own set of problems: Transient protein-DNA interactions are completely missed (Schmiedeberg et al., 2009), prolonged fixation can recover non-specific interactions at highly occupied regions (Baranello et al., 2016), cross-linking occurs between protein-protein and protein-DNA interactions so direct and in-direct interactions cannot be disassociated (Hoffman et al., 2015), and fixation can lead to irreversible changes in the protein so the antibody can no longer recognise the epitope (Scalia et al., 2017).

1.3.3 New and improved ChIP-seq methods

Over the subsequent years a number of upgraded protocols have been developed to overcome the limitations associated with immunoprecipitation, library construction,

and formaldehyde fixation:

The search for low cell number alternatives began with the development of a nano-ChIP-seq protocol in 2010 which was used to identify trimethylation of histone H3 at lysine 4 (H3K4me3) in 10^4 mouse embryonic stem cells (ESC) (Adli, Zhu, and Bernstein, 2010). In this protocol, already immunoprecipitated DNA is first primed with Sequenase enzyme using a custom primer containing a universal PCR sequence, a BciVI restriction site, and a random 9-mer sequence. The primed DNA is then PCR amplified using another custom primer containing the same universal PCR sequence and BciVI restriction site. The resulting amplicons are lastly digested using the corresponding BciVI restriction enzyme to produce DNA products which can be directly ligated to adapters for high-throughput sequencing. While these changes to the standard library preparation successfully reduced the number of cells required for ChIP-seq, the protocol was only tested on histone modifications which are more abundant and therefore generate a larger amount of starting material than transcription factors.

Three years later came the development of the carrier-assisted chromatin immunoprecipitation protocol (C-ChIP-seq) (Zwart et al., 2013). In this protocol, small numbers of targets cells are first mixed with *Drosophila* cells before a conventional formaldehyde ChIP-seq is performed. The foreign chromatin acts as a so-called carrier for the small amount of target chromatin throughout the isolation procedure and consequently improves enrichment from limited cell numbers. Although the exact mechanism remains undetermined, it is thought that the carrier chromatin bulks up the solution which helps to retain the target chromatin during chemical handling. Using this protocol, the authors were able to identify binding of the transcription factor Estrogen receptor alpha using 10^4 cells from biopsies of human breast tumours.

In 2014, a novel indexing-first chromatin IP protocol (iChIP-seq) was published which

was used to identify histone modifications (H3K4me1 and H3K4me2) from a minimum of 500 hematopoietic precursor cells (Lara-Astiaso et al., 2014). In this protocol, the population of cells are first fixed and then sorted into sub-populations which are sonicated. The chromatin is then immobilised on magnetic beads using an anti-H3 antibody and indexed by ligating sequence-specific adapters. The barcoded chromatin is then released from the beads and pooled before conventional formaldehyde ChIP-seq is performed. While this protocol significantly reduced the number of cells required for ChIP-seq of histone modifications, the minimum required for transcription factors was still 10^4 cells.

Later in the year, a ChIP-seq protocol which used native chromatin and did not require fixation was also published (N-ChIP-seq) (Kasinathan et al., 2014). Instead of formaldehyde cross-linking, the protocol relies on the intermolecular bonds between nucleic acids and amino acids to sustain during the delicate chemical handling. The protocol begins with isolating unfixed nuclei and shearing the chromatin using light MNase digestion. The chromatin is then solubilised by careful needle extraction and protein-DNA complexes are immunoprecipitated using antibody-conjugated magnetic beads. Using this protocol, the authors successfully identified binding of multiple transcription factors in *Drosophila* and *Saccharomyces* using 10^7 cells. Only a year later however, an upgraded N-ChIP-seq protocol for ultra-low-input (ULI-NChIP) was published and was used to identify histone modifications from a minimum of 10^3 embryonic stem cells (Brind'Amour et al., 2015). The authors had optimised multiple steps from the original protocol associated with cell sorting, chromatin fragmentation, and library preparation. Despite these improvements over formaldehyde fixation, there is still a concern that highly dynamic DNA-binding proteins may redistribute during library preparation (Skene and Henikoff, 2017).

In 2015, a protocol for ChIP combined with Tn5 transposase tagmentation (ChIPmentation) was used to identify histone modifications (H3K4me3 and H3K27me3) and CTCF binding from 10^4 cells and 10^5 leukaemia cells, respectively (Schmidl et al., 2015). In this protocol, the usual multi-stage library preparation (end-repair, purification, A-tailing, adapter-ligation, and size selection) was replaced with a single stage using Tn5 tagmentation of the DNA fragments. The Tn5 transposase is loaded *in vitro* with sequencing adapters and in a single step can fragment and tag DNA with the sequencing adapters. This makes the library preparation protocol faster, simpler, and more efficient which reduces the number of cells required to generate a sufficient DNA yield for high-throughput sequencing. While this in itself did not set new records for the minimum number of cells required, this simple change could be combined with other protocols to rely on less starting material.

The utility of Tn5 transposase tagmentation was later expanded by the development of Transposase-Assisted Chromatin Immunoprecipitation (TAM-ChIP) by the Active Motif company (Samuelsson et al., 2015). This technology allowed them to identify histone modifications (H3K4me3 and H3K27me3) in primary human tissues, such as formalin-fixed paraffin-embedded (FFPE) colon tumour cells. The protocol involves fixing cells with formaldehyde, then sonication to produce short DNA fragments, followed by incubation with an antibody directed against the DNA-binding protein. A species-specific antibody chemically linked to Tn5 transposase containing barcoded Illumina sequencing adapters is then added to bind the ChIP antibody. Upon activation of the Tn5 transposase, nearby DNA surrounding the genomic region is tagmented with sequencing adapters and then the antibody-bound DNA-protein complex is immunoprecipitated. The DNA is finally reverse cross-linked and purified for library amplification then sequencing. Not only does this protocol reduce the number of cells required, it also removes many of the downstream steps such as chromatin shearing.

More importantly though, the barcoded sequencing adapters allow multiple targets to be assayed within the sample biological sample, giving rise to a more detailed investigation of multiple epigenetic modifications regardless of biological replicate variability.

Later that year, Brind'Amour and colleagues published a protocol for microfluidics ChIP-seq (Drop-CHIP) which they used to profile histone modifications (H3K4me2 and H3K4me3) in single mouse embryonic stem cells (ESC) and embryonic fibroblast cells (MEF) (Rotem et al., 2015). The protocol combined many of the optimisations from previous publications with a droplet-based microfluidics (DBM) system capable of processing single cell reactions. To avoid pull down of non-specific antibody binding, chromatin from individual cells was indexed (similar to iChIP-seq) before immunoprecipitation. Specifically, the DBM system was used to combine two separate aqueous drops with an enzymatic solution: one drop containing the single cell alongside a weak detergent and MNase enzyme, the other drop containing the unique barcode sequencing adapters, and one small aliquot of enzymatic buffer with DNA ligase. Indexed chromatin fragments from 100 cells were then pooled (similar to ULI-NChIP-seq) and combined with carrier chromatin (similar to C-ChIP-seq) before performing ChIP and using the enriched DNA for high-throughput sequencing. While this was a ground-breaking advancement, the protocol is extremely complicated and requires highly specialised equipment, likewise it has never been used to investigate transcription factor binding because of the smaller amount of input DNA generated.

An important landmark in the search for low input ChIP-seq alternatives, was the recent development of the CUT&RUN protocol (Skene and Henikoff, 2017). Unfixed nuclei are immobilised on magnetic beads and treated successively with a target antibody and protein A-MNase enzyme. The antibody binds to the target protein followed by MNase binding to the antibody. Calcium is then added to activate the MNase and

cleave protein-DNA complexes which are released and diffuse out of the nucleus. The mixture is then centrifuged to recover the supernatant – containing the protein-DNA complexes – and the DNA is used to prepare libraries for high-throughput sequencing. Importantly, this approach does not rely on cross-linking or immunoprecipitation, so artefactual biases associated with formaldehyde fixation and the number of cells required to achieve a sufficient DNA yield are significantly reduced. The recommended sequencing depth is also lower because the MNase only cleaves DNA around binding sites so the amount of non-specific DNA in the reaction is minimised. Excitingly, the original CUT&RUN protocol was then later upgraded to facilitate ultra-low input (abbreviated uliCUT&RUN) and achieve single-cell resolution (Hainer et al., 2018a). The uliCUT&RUN protocol was used to profile CTCF, Nanog, and Sox2 binding from individual mouse embryonic stem cells and blastocyst embryos.

In conclusion, the advancements made over the last decade have culminated in the ability to profile histone modifications and transcription factors from single cells (using Drop-ChIP and uliCUT&RUN). This presents an unprecedented opportunity to investigate protein-DNA interactions at a much higher resolution than previously examined, especially from *in vivo* tissues where the number of cells is limited. However, all of these low cell number methods still rely on the existence of a specific and sensitive antibody (often expensive and difficult to produce) which binds with high affinity to the target protein (Aughey and Southall, 2016). Without an appropriate antibody the advances in low-input protocols cannot be exploited and instead alternative methods which do not rely on antibodies are required. Additionally, both Drop-ChIP and CUT&RUN require cell isolation for cell-specific profiling, compromising their ability to measure *in vivo* binding (Skene and Henikoff, 2017).

1.3.4 DNA adenine methyltransferase identification sequencing (DamID-seq)

An alternative and increasingly popular method to identify genome-wide DNA binding sites is DNA adenine methyltransferase identification (DamID) (Van Steensel and Henikoff, 2000). The protocol involves expressing a fusion protein containing the protein of interest (POI) which is tethered to *Escherichia coli* Dam methyltransferase. When the POI binds to its target sites, Dam adds a methyl group to the N6 position of adenine in nearby GATC sequences. The chromatin is then extracted and fragmented using a DpnI restriction enzyme which cuts at methylated GATC sequences. Sequencing adapters are ligated to the restriction fragments and a DpnII restriction enzyme is used to cut non-methylated GATC sequences. The adapter-ligated DNA is then enriched using PCR amplification and sequenced using high-throughput sequencing (Sun et al., 2003; Wu and Yao, 2013). Importantly, because Dam itself is a DNA-binding protein a control sample containing just Dam is also sequenced and used to measure the amount of non-specific background methylation. In order to identify DNA binding sites from DamID-seq data, the sequencing reads are mapped back to the reference genome using alignment software and restriction fragments which are differentially methylated between the Dam-fusion and Dam only samples are identified. Restriction fragments which have been methylated significantly higher in the Dam-fusion over the Dam only samples are regarded as containing a DNA binding site for the protein of interest. The advantage of DamID is that it does not require cross-linking or antibodies so no artefactual biases from formaldehyde treatment are expected, and no immunoprecipitation step is required so fewer cells ($< 10^4$ cells) can be used as starting material. However, an important consideration in the application of DamID is the expression level of the Dam-fusion protein. High expression can lead to the entire genome becoming methylated making it difficult to identify targeted binding. High

concentrations of the Dam protein can also lead to toxicity of the cells and accumulation of artefactual biases (Pindyurin et al., 2016). Consequently, the expression of the Dam-fusion protein is kept low by exploiting the leakiness of uninduced promoters, such as the hsp70 heat-shock promoter in *Drosophila* cells and the ecdysone promoter in mammalian cells. An expression construct encoded for the Dam-fusion or Dam only protein is placed directly upstream of a leaky promoter limiting expression to levels almost undetectable by western blotting.

Whilst DamID offers a solution to the traditional problems associated with formaldehyde cross-linking and antibody availability, it does not come without certain complications: Given that Dam must be expressed at a sufficiently low level to avoid toxicity, which is usually achieved by inserting an expression vector downstream of leaky promoters, the model system must be tractable to genomic modification in order to generate the transgenic cells required (Aughey, Cheetham, and Southall, 2019). Without these tools, it is difficult but not impossible to achieve the requisite expression level via non-integrating lentiviral vectors, however cell-type specific profiling then becomes limited. Next, in order to achieve a sufficient signal Dam must be expressed for a number of hours ensuring all possible GATC sites have been profiled. The result of this long exposure time is that any binding observed could have occurred at any point, which is the opposite of ChIP-seq data whereby a snapshot of binding is detected at the point in which formaldehyde fixation is administered (Aughey and Southall, 2016). Additionally, because DamID is reliant on the presence of GATC sites to indicate binding the resolution achievable is reliant on the frequency of sites along the genome. Certain regions with a defined sequence bias, such as repeats and promoters, may therefore be under-represented for GATC sequences such that DamID may not be able to identify binding. Finally, it must not be forgotten that physically linking Dam to another DNA-binding protein may have an undetermined effect on

the structure and function of the target protein. It is therefore crucial that control experiments be carried out to ensure the fusion is viable: An electrophoretic mobility shift assay (EMSA) can first be used to confirm the *in vitro* binding ability of the altered target protein. If successful, downstream *in vivo* experiments can be performed to ensure biological function is also retained. For example, an inactive version of Dam can be linked to the target protein and expressed at usual levels in a knock-out system to show that no deleterious effects or altered phenotypes are observed (Tosti et al., 2018).

Although the original DamID protocol was developed nearly two decades ago for specific use in *Drosophila*, it has since been adapted for a number of other organisms including mouse (Tosti et al., 2018), human (Vogel et al., 2006), yeast (Steglich, Sazer, and Ekwall, 2013), plants (Germann et al., 2006), and *C. elegans* (Schuster et al., 2010). One interesting adaptation of DamID is the ability to detect interacting DNA-binding proteins using a method called split DamID (SpDamID) (Hass et al., 2015). The Dam protein is split in half and tethered to separate DNA-binding proteins which are speculated to form a protein-protein interaction. If the two proteins interact, the two inactive halves of Dam are reconstituted to form an active protein and the location of the dimeric protein-DNA interaction along the genome is marked. This method can also be used to differentiate between monomeric and dimeric binding since two of the same halves of Dam cannot reconstitute its methylation activity. Additionally, a targeted DamID approach (TaDa) has been developed in *Drosophila* which can be used to profile binding in a cell-specific manner (Southall et al., 2013). Using the GAL4/UAS targeted gene expression system, different *Drosophila* strains were generated which expressed the Dam-fusion protein only in specific cell-types. This allowed for the *in vivo* profiling of RNA polymerase II binding in different neuronal cell types using fewer than 10^4 cells without needing to isolate the cells. Furthermore, the Bas van

Steensel group recently published a single-cell DamID-seq protocol, where the authors reported the mapping of nuclear lamin B1 interactions (LmnB1) in single human myeloid leukaemia cells (Kind et al., 2015). After 15 hours of Dam-LmnB1 expression, cells were FACS-sorted and captured in 96 well plates containing lysis buffer. The chromatin was then digested with DpnI followed by adapter ligation and PCR amplification as is standard. Importantly, all of the library preparation steps were performed in the same well via sequential addition of reagents so there was no potential for loss of starting material.

Despite these advances, DamID has only recently been combined with high-throughput sequencing. The original DamID protocol by the Bas van Steensel lab and many subsequent studies used microarrays to identify DNA-binding. The resolution of microarrays however is considerably less than high-throughput sequencing and involves its own set of biases which must be investigated (Hurd and Nelson, 2009). In addition, very few published studies have up until now demonstrated transcription factor binding in mammalian cells using DamID-seq (Jesus Domingues et al., 2016). Recently, our lab was the first to develop a protocol to perform mammalian DamID-seq with transcription factors using a minimum of 10^3 cells. While this is an encouraging step forward, there are a number of issues which need to be investigated before DamID-seq can become more widely adopted. There has currently been no investigation into potential systematic biases in the sequencing data produced by DamID-seq experiments. This is concerning because many newly minted technologies exhibit artefacts which are misinterpreted as genuine biology leading to inappropriate conclusions about the results. Due to the lack of sequencing data, the accuracy and sensitivity of DamID-seq has also not been assessed. This is primarily because the overwhelming number of DamID studies were not measured by high-throughput sequencing, but rather microarray technology. The combination of DamID with sequencing is a relatively new

adaptation (Lie-A-Ling et al., 2014) most of the subsequent studies were performed in *Drosophila* rather than in a mammalian system, which is more applicable to our research. A handful of studies have attempted to compare the results from DamID and ChIP, such as GAF (Nègre et al., 2006) and HP1 (Yin et al., 2011) *Drosophila* cells, however these comparisons were based on DamID-chip and ChIP-chip datasets, both of which suffer from a lower resolution than sequencing. Nevertheless, the extent to which both studies attempted to compare results was limited to simply calculating correlation coefficients between microarray intensities (r=0.37 and r=0.77 for GAF and HP1, respectively). Only one other study went further than this, showing that the cognate motif was sufficiently enriched in the regions identified by the DamID method (Bemmel et al., 2010). Additionally, most DamID datasets were produced because of the inability to perform ChIP in their biological system and consequently have relied on the comparisons made by previous studies to ensure confidence in their experiments. Finally, at the time the work in this thesis was carried out there was actually very little DamID-seq data available for transcription factors in mammalian cells, and only one with accompanying ChIP-seq data (Jacinto, Benner, and Hetzer, 2015). Instead larger DNA-binding proteins such as RNA polymerase II and Lamin B1 have been measured in more recent protocols such as TaDa (Southall et al., 2013) and single cell DamID-seq (Kind et al., 2015). Through comparative analyses with ChIP-seq data, it is important to examine whether DamID-seq can detect the same DNA binding sites and measure how this reproducibility decreases with cell number. Additionally, appropriate computational methods to analyse the sequencing data in a streamlined and accurate manner are required.

1.4 Aims

The aim of this work is to measure the accuracy and sensitivity of mammalian DamID-seq experiments by investigating different experimental protocols, comparing to previously established assays, and developing rigorous analysis methods. I will first examine DamID-seq data for potential sources of bias caused by unwanted biological and technical variation. To achieve this, I will determine whether the ability to measure methylation changes according to different sample preparation steps and deterministic features such as GC content. I will then evaluate whether DamID-seq data can be used to accurately measure chromatin binding from minimal numbers of cells. To realise this, I will compare qDamID and DamID-seq methylation to identify an appropriate normalisation and modelling strategy. Following which I will develop a novel analysis method for calling binding sites from DamID-seq data and perform a large-scale comparison using published ChIP-seq data. I will last determine whether DamID-seq data can be used to accurately measure chromatin accessibility from minimal numbers of cells. To accomplish this, I will develop a novel analysis method for calling accessibility sites from DamID-seq data and perform a large-scale comparison with previously established assays using published ATAC-seq, DNase-seq, and FAIRE-seq data. Overall, the results from these analyses will be used to guide the development of a Bioconductor package for the analysis, interpretation, and visualisation of DamID-seq experiments.

Chapter 2

Materials and methods

2.1 Availability of code and data

The computer code and supporting data for this thesis can be downloaded from a private FTP server. The credentials to access the server are:

Hostname `stembio19.med.ed.ac.uk`

Address `/home/s1437643/work/Thesis`

Username `s1437643`

Password `Whiwofbs1`

The code developed to analyse DamID-seq data can also be downloaded from a public GitHub repository (https://github.com/jma1991/Daim) and the sequencing data from published experiments can be downloaded from the relevant public databases (see attribution statement for each chapter and availability of sequencing data sections in this chapter). For clarity and reproducibility, command arguments have also been documented in the relevant sections. The character $ at the beginning of each line represents the command prompt and should not be typed.

2.2 Analysis of DamID-seq experiments

2.2.1 Statement of attribution

The DamID-seq libraries were generated by Dr Luca Tosti - a previous PhD student in Prof. Keisuke Kaji's research group - and analysed by the author, James Ashmore.

2.2.2 Availability of sequencing data

The raw and processed sequencing data is available to download from the private FTP server provided (see Section 2.1).

2.2.3 Processing of sequencing data

Quality control

The DamID-seq libraries were sequenced by either Edinburgh Genomics or the Beijing Genomics Institute on an Illumina HiSeq 4000 machine. Approximately 20 million 1x50 bp reads were generated from each library. Importantly, samples within comparisons were sequenced together so no batch effects were generated. Read quality was evaluated from reports generated by FastQC (version 0.11.7) (Andrews, 2010) and MultiQC (version 1.5) (Ewels et al., 2016). Reads were trimmed using Cutadapt (version 1.16) (Martin, 2011) to remove low quality bases and adapter contaminant sequences:

1. DamID Sequencing Adapters

   AdRt CTAATACGACTCACTATAGGGCAGCGTGGTCGCGGCCGAGGA

   AdRb TCCTCGGCCG

   PCR GGTCGCGGCCGAGGATC

2. Illumina Nextera Adapters

Read1 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

Read2 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

Index1 CAAGCAGAAGACGGCATACGAGAT[N]GTCTCGTGGGCTCGG

Index2 AATGATACGGCGACCACCGAGATCTACAC[N]TCGTCGGCAGCGTC

Read alignment

Reads were aligned to the mm10 assembly of the mouse genome (Karolchik et al., 2004) using BWA-MEM (Li, 2013). Duplicate reads were identified using the MarkDuplicates command from Picard Tools (Broad Institute, 2018) then sorted and indexed using Samtools (Li et al., 2009). Reads were filtered using Bedtools (Quinlan, 2002) and Samtools (Li et al., 2009) with multiple criteria: First, alignments to alternate, mitochondrial, unplaced, and random chromosomes were removed. Next, non-unique, secondary, duplicate, and supplementary alignments were removed. Last, alignments to ENCODE (ENCODE Project Consortium, 2012) and Mitochondrial (Buenrostro et al., 2013) blacklist regions were removed.

Genome coverage

To visualise coverage along the genome, read depth at each genome position was generated in bedGraph format using the *genomecov* command from Bedtools (version 2.27.1) (Quinlan, 2002). The read depth values were then normalised for library size by calculating reads per million (abbreviated RPM) with the formula:

$$RPM = \frac{D}{S} \quad \text{and} \quad S = \frac{T}{10^6} \tag{2.1}$$

where D is defined as the read depth at each genome position, S as the per million scaling factor, and T as the total number of mapped reads. For faster display performance and to save storage space, the plain-text bedGraph files were converted into indexed

binary bigWig files (Kent et al., 2010) using the *bedGraphToBigWig* command from Kent Utilities (version 4.0) (Kuhn, Haussler, and Kent, 2013). Genome browser images were captured using the Integrative Genomics Viewer (version 2.4.9) (Robinson et al., 2011) and pyGenomeTracks (version 2.0) (Ramírez et al., 2016).

2.2.4 Identification of DNA-binding sites

Experimental design

In Dam-fusion protein expressing cells, binding sites are marked by the addition of a methyl group to the N6 position of adenine on nearby GATC sequences. The genomic DNA is then extracted, fragmented using methylated GATC specific restriction enzyme DpnI, subjected to adapter ligation and PCR amplification, sequenced, and aligned to a reference genome. When expressing a Dam-fusion protein, non-specific background methylation caused by non-specific binding of Dam at open chromatin loci is typically observed. To distinguish binding sites from background, the methylation profiles between Dam-fusion and control cells expressing just Dam are compared. Binding sites are therefore identified by locating regions of the genome which have been methylated more frequently in Dam-fusion than Dam protein expressing cells.

Fragment quantification

To measure methylation along the genome, the number of reads aligned to each hypothetical restriction fragment were counted. An annotation file of restriction fragment locations was generated by running a simulated digest on a map of DpnI and DpnII restriction sites in the reference genome. The restriction map was constructed by finding all the occurrences of the pattern *GATC* in the genome sequence using the *matchPattern* command from the Biostrings package (version 2.46.0) (Pagès et al., 2017). The number of reads aligned to each restriction fragment was then counted using the *featureCounts*

command from the Rsubread package (version 1.28.1) (Liao, Smyth, and Shi, 2013). To obtain reliable counts, reads which were marked as duplicate, multi-mapping, or aligned to more than one restriction fragment were discarded.

Sample normalisation

Normalisation is a standard and essential step in the preprocessing of sequence count data. It is required to remove unwanted *technical variation* and transform the raw counts into usable methylation measurements. If not corrected, *technical variation* can often be mistaken for *biological variation* and lead to false discoveries (Hicks et al., 2018). A considerable number of normalisation methods have been developed for sequence count data, but many of these assume that the samples have the same global distribution (Aleksic, Carl, and Frye, 2014). To determine whether these methods were appropriate for DamID-seq count data, a test for the assumptions of quantile normalisation was carried out using the quantro package (version 1.14.0) (Hicks and Irizarry, 2015). The test results indicated that the distribution of the Dam-fusion libraries were significantly different to the Dam libraries, so instead the count data was normalised with an appropriate weighted quantile method called smooth quantile normalisation using the qsmooth package (version 0.0.1) (Hicks et al., 2018). In addition, sequence count data typically exhibits a linear relationship with the length of the genomic feature being counted. In order to compare between features of different lengths, the counts are typically divided by the feature length (Mortazavi et al., 2008). To determine whether DamID-seq count data follows this trend, the relationship between restriction fragment length and count abundance was modelled using the CQN package (version 1.24.0) (Hansen, Irizarry, and Wu, 2012). The modelling showed a non-linear relationship with restriction fragment length, which was corrected by applying CQN computed offsets to the smooth quantile normalised count data.

Differential methylation

To identify differentially methylated restriction fragments, the length-corrected smooth quantile normalised count data was analysed using the limma package (version 3.34.9) (Ritchie et al., 2015). Multidimensional scaling (abbreviated MDS) plots indicated that the Dam-fusion samples exhibited significantly more within group variation than the Dam samples. To account for this disparity, linear modelling was performed using the *arrayWeights* (Ritchie et al., 2006) function to compute Dam-fusion and Dam group weights.

Fragment clustering

Methylation by Dam has been shown to extend from any given DNA-binding site to a distance of up to 5 kb (Van Steensel and Henikoff, 2000). Differentially methylated restriction fragments were therefore merged into clusters using the *mergeWindows* command from the csaw package (Lun and Smyth, 2016). For clustering, the maximum distance between adjacent fragments was defined as 261 bp (the average size of fragments in the mouse genome) and the maximum size of merged clusters was defined as 5 kb (the maximum spread of Dam methylation). Clusters with a FDR < 0.1 and $\log_2$FC > 0 were defined as statistically significant DNA-binding sites.

2.2.5 Identification of DNA-accessibility sites

Experimental design

To distinguish accessibility sites from non-specific methylation, the methylation profile between Dam and the average methylation rate is compared. Accessibility sites are therefore identified by locating regions of the genome which have been methylated significantly more in Dam protein expressing cells than the average methylation rate.

Background modelling

In order to identify non-random enrichment of restriction fragments, methylation measurements were compared to a neighbourhood background methylation rate. For each fragment, the rate was defined as:

$$Rate = \max(\lambda 25k, \lambda 50k, \lambda 100kb) \tag{2.2}$$

where $\lambda 25k$, $\lambda 50k$ and $\lambda 100k$ are estimated from 25 kb, 50 kb, and 100 kb neighbouring windows. Accessibility sites were then identified from differentially methylated fragments using the procedure described previously (see Subsection 2.2.4).

2.2.6 Implementation of R package

The analysis, interpretation, and visualisation of DamID-seq experiments was performed using the Daim package (`https://github.com/jma1991/Daim`). It includes functions to process sequencing data, quantify fragment abundance, and detect DNA binding-and-accessibility sites. Additionally, the package provides functions to perform downstream functional and sequence analyses. Daim is implemented in R and takes advantage of the Bioconductor ecosystem (`http://bioconductor.org`). It is freely available under a Massachusetts Institute of Technology (abbreviated MIT) license.

2.2.7 Comparison of DamID-seq analysis pipelines

There are currently four different published analysis pipelines for DamID-seq data, all of which implement novel approaches but none of which have been benchmarked:

The first pipeline was released in 2015 by Li and colleagues and is described as a non-parametric algorithm for peak calling (Li, Hempel, and Jiang, 2015). It begins by

randomly sampling 90% of alignments from Dam and then counting how many Dam-fusion and resampled Dam reads align within non-overlapping 100 bp windows along the genome. To adjust for sequencing depth, read counts are normalised to reads per million (RPM) and the signal in each window is represented as the log2 fold change of the RPM values (Dam-fusion/Dam). The windows are then filtered by removing those with a negative log2 fold change (Dam > Dam-fusion) which by definition do not contain binding. The median log2 fold change of the remaining windows is then calculated to estimate the average level of methylation across the genome. To define significant windows, the previous steps are repeated (N = 200) and then the distribution of the median log2 fold changes (MFC) is calculated. Any windows which have a log2 fold change greater than the 95% percentile of the MFC distribution are declared significant. Adjacent windows are then merged together to define peak regions. As a piece of software, the pipeline by Li and colleagues suffers from a number of problems: The pipeline is provided "as is" meaning all of the file paths are actually hard coded to the authors computer and environment. You have to manually go through the code and replace any instances of a hard path with the relevant path on your computer. This is obviously fraught with issues, not only because it limits reproducibility but also because it is sometimes not immediately obvious what type of file the hard-coded paths refer to on the author's computer. There are also many "magic numbers" in the code (i.e. a hard-coded value which has no obvious description or justification) which limits the overall readability and understanding of the software. For example, chromosome sizes are hard-coded to the Drosophila genome and because there is no automated way to update these values (besides a manual search and replace) it becomes very difficult to adapt the algorithm to any other reference genome. Additionally, the peak regions identified by the algorithm do not contain any peak-level statistics such as enrichment or significance values so no downstream filtering (e.g. remove weakly bound regions) can be performed. Finally, no manual is provided to

explain how to run the pipeline and there is no list of required packages, so it is particularly difficult to debug when an error has occurred. Given these multiple serious issues, this pipeline should be disregarded in any future work looking to evaluate DamID-seq analysis software.

The second pipeline was released later on in 2015 by Marshall and colleagues and is described as an automated pipeline for processing DamID-seq datasets (Marshall and Brand, 2015). It begins by counting how many Dam reads align within GATC restriction fragments along the genome. The fragments are then filtered by removing any with zero read counts and the remaining fragments are divided into deciles by total read count. Those within the first three deciles and the highest 10% are also removed because they produce inconsistent normalisation factors and contain genuine binding, respectively. The distribution of the log2 fold change values (Dam-fusion/Dam) for all remaining fragments is then calculated and the Dam-fusion counts are normalised such that the point of maximum kernel density of the log2 fold change values equals zero. Coverage tracks are finally generated containing either the normalised Dam-fusion and Dam read counts or the log2 fold change values. The pipeline does not actually include a peak calling stage, instead an additional piece of software based upon a separate publication can be downloaded and applied to the coverage tracks to identify binding sites (Wolfram et al., 2012). It is difficult to describe the exact methodology of the peak caller because it was adapted from an algorithm used for DamID microarray experiments, as such it has not been published or adequately described (see algorithm section of the following website: `https://github.com/owenjm/find_peaks`). According to the algorithm described by Wolfram and colleagues, the peak caller first appears to merge consecutive restriction fragments and defines those greater than 900 bp in length which have a greater than 2-fold change as potential peak regions. A false discovery rate (FDR) is then assigned to these regions using a Monte Carlo resampling

method whereby the frequency of fold changes for a range of peak widths located in the chromosome arms is calculated. This data is then used to model the exponential decay of the FDR with respect to increasing fold change and peak width, therefore enabling extrapolation of FDR values for higher and broader peaks. While the lack of details regarding the peak calling software is concerning, the pipeline itself is well documented and can be readily installed. The assumptions of the normalisation procedure however are subject to critique (as discussed extensively at the end of Subsection 4.4.2). Briefly, the assumption that the majority of fragments should exhibit a log2 fold change of zero and that the Dam-fusion read counts should be normalised to achieve this assumption goes against the theory behind DamID which shows that most non-target sites should be primarily methylated by Dam and therefore should have a log2 fold change less than zero. In order to establish whether the assumptions of this normalisation procedure are practically appropriate, would require a large-scale analysis of multiple DamID-seq datasets measuring a range of DNA-binding proteins which exhibit different binding dynamics. In fact, the pipeline itself even has a separate setting whereby if their assumption does not hold for the data it will instead perform a standard read depth normalisation. Given this work has established that retaining the significant differences in the methylation distributions between the Dam-fusion and Dam proteins is important for peak calling accuracy, this pipeline no longer becomes particularly novel.

The third pipeline was released in 2016 by Gutierrez-Triana and colleagues and is based upon an improved protocol called iDamID-seq which inverts the DpnI and DpnII digestion order and adds steps that involve a phosphatase and exonuclease (Gutierrez-Triana et al., 2016). The pipeline begins by generating all possible GATC restriction fragments in the specified genome, then removes any which are outside the range of 200 to 2000 bp in length. Next, the number of Dam-fusion and Dam reads

aligned within this subset of fragments is calculated and only fragments with a minimum number of reads relative to the fragment length are retained. This threshold was computed as three times the total number of reads in all fragments divided by their total length. Fragments that were not further apart than the smallest fragment length were then joined together, and the number of reads aligned within these join regions were calculated to produce a count matrix. Significant differences between the Dam-fusion and Dam replicates are then computed using the DESeq2 package which was originally developed for RNA-seq analysis (Love, Huber, and Anders, 2014). Unlike the other pipelines, this one requires the reference genome to be available as a BSgenome package, so it is not readily applicable to users with a non-reference genome in FASTA format. Additionally, because fragments with high Dam-fusion and Dam read counts are merged prior to differential methylation analysis the algorithm cannot differentiate between sharp binding events (e.g. transcription factors) whereby individual fragments in the merged region are differentially methylated in both directions. Finally, because DESeq2 by default uses a two-tailed hypothesis test to detect differential gene expression in both directions, it is more difficult to achieve significance in one direction (i.e. testing only whether Dam-fusion > Dam). Specifically, the two-tailed test splits the user-specified significance level and applies it to both directions, thus each direction is only half as strong as a one-tailed test which applies the significance in just one direction.

The fourth pipeline was released later on in 2016 by Maksimov and colleagues and is specific to a particular DamID-seq protocol which only sequences DNA flanking GATC sequences (Maksimov, Laktionov, and Belyakin, 2016). The pipeline first aligns reads to the specified genome and then counts how many reads align immediately upstream or downstream of a GATC sequence. Next, fragments showing greater than

2-fold standard deviation (referred to as the dynSD between replicates) from the Dam-fusion group average are removed. To identify significantly methylated fragments the level of biological and technical variability in the data is assessed using a one-sided Fisher's exact test between Dam-fusion and Dam replicates (biological variability) and within Dam-fusion replicates (technical variability). This information is then used to assign a false discovery rate to the P values returned by the Fisher's exact test for each GATC fragment. Surprisingly, this pipeline does not actually create peak regions from the significant fragments, instead only the fragment-level statistics are returned. This means that a single DNA-binding event will be represented by multiple fragments in the output file, making downstream analyses cumbersome. Whilst the user can take it upon themselves to merge the fragments afterwards, all of the fragment-level statistics would also have to be summarised using an appropriate calculation. This is particularly difficult for P values where simply averaging the P value from independent tests can result in loss of type I error control (Lun and Smyth, 2014). More worryingly, because the pipeline only counts reads which are adjacent to GATC sequences (which is protocol-specific) it is not applicable to other protocols which sequence the entirety of the GATC fragment allowing the read counts to be normalised with respective to fragment length and GC content.

2.2.8 Development of analysis workflow

Processing of the raw sequencing data was automated using a custom pipeline built with the Conda package manager (version 4.5.1) (Anaconda, 2018) and Snakemake workflow engine (version 4.7.0) (Köster and Rahmann, 2012). The analyses described in this section can be reproduced by downloading and running the relevant DamID-seq workflow.

2.3 Analysis of quantitative DamID (qDamID) experiments

2.3.1 Attribution statement

The qDamID template genomic DNA was prepared by Prof. Keisuke Kaji and the qDamID experiments were performed by the author, James Ashmore.

2.3.2 Restriction fragment screen

In order to investigate how well DamID-seq data represented the actual *in vivo* DNA methylation levels, I aimed to compare the DamID-seq data with qDamID data, a qPCR-based method to detect methylation in cells expressing Dam (Van Steensel and Henikoff, 2000). For this purpose, I first identified 192 restriction fragments which cover the entire methylation levels and fold changes between Dam and Dam-Oct4 protein expressing ES cells. The advantages of sampling are that the laboratory cost is lower and data collection is faster than measuring the entire population, which would also be impractical in this experiment. The average abundance and fold change for each restriction fragment was calculated from DamID-seq data described in Section 2.2: First, alignments from replicates were merged to produce a single sorted BAM file using the *merge* command from Samtools (Li et al., 2009). The number of reads aligned to each restriction fragment was then counted using the *featureCounts* command from Rsubread (Liao, Smyth, and Shi, 2013). Reads aligned to more than one restriction fragment, multi-mapping reads, and reads marked as PCR duplicates were not counted. Next, to minimize differences between samples for restriction fragments with small counts and normalise with respect to library size, a regularized log (abbreviated rlog) transformation was applied using the *rlog* command from the DESeq2 package (Love, Huber, and Anders, 2014). After this, methylation differences between samples were visualised with a Bland-Altman plot (also called an MA plot) generated using the *smoothScatter* function from R (R Core Team, 2017). On the graph, each

restriction fragment is represented by a point, the x-axis measures the average abundance over the mean of normalised counts (A-values), and the y-axis measures the $\log_2$ fold change between samples (M-values). Finally, a 25 x 25 reference grid consisting of 625 equally spaced points was drawn over the MA plot and the 10 restriction fragments closest to each point as measured by the Euclidean distance were selected.

2.3.3 Primer design pipeline

To design accurate and efficient primer pairs for the 192 restriction fragments in a high-throughput manner, an automated computational pipeline was developed. For a given restriction fragment, primer pairs were designed flanking each restriction site using Primer3 (Untergasser et al., 2012). The search parameters were configured (Thornton and Basu, 2011) to increase the probability of finding a large number of efficient primer pairs (see Appendix B). For each primer pair, amplimers were identified by performing an *in-silico* polymerase chain reaction (abbreviated PCR) against the mm10 assembly of the mouse genome (Karolchik et al., 2004) using isPCR (Kuhn, Haussler, and Kent, 2013). The alignment parameters were configured to increase the probability of finding all potential amplimers (maxSize=4000, minSize=15, and minGood=15), and only primer pairs with zero off-target amplimers were retained. For each reference point, only the restriction fragment with the lowest primer pair penalty was selected.

2.3.4 Quantitative DamID protocol

The PCR templates for quantitative DamID (abbreviated qDamID) were generated by extracting genomic DNA from Dam and Dam-Oct4 protein expressing ES cells using the DNeasy Blood & Tissue Kit manufactured by QIAGEN. Of the DNA that was

extracted, 4 µg of genomic DNA (gDNA) was diluted in 32 µl of purified water (abbreviated H_2O). The solution was then divided into two tubes: one containing DpnII buffer (2 µl) and H_2O (2 µl) for undigested samples; and the other containing DpnII buffer (2 µl) and DpnII enzyme (2 µl) for digested samples. Both tubes were then incubated overnight at 37°C and the DNA was diluted to 10 ng/µl by adding 180 µl of H_2O.

The qDamID experiments were performed in triplicate using 4.5 µl of PCR Master Mix, 0.9 µl of digested or undigested DNA (10 ng/µl), and 1.8 µl each of the forward and reverse primers. The qPCR reactions were performed in 384 well plates, in a 9 µl reaction, using the LightCycler 480 SYBR Green I Master Mix and LightCycler 480 system. Inefficient primer pairs or off-target amplimers were identified by extremely high cycle threshold values or a bimodal peak in the melting temperature curve analysis, respectively, and excluded from the analysis. For the full list of primers see Appendix A.

2.3.5 Differential methylation analysis

To calculate restriction fragment abundance, the Ct values were normalised using control restriction fragments (zero reads in the DamID-seq data) and the difference in normalised Ct values (abbreviated ΔCt) between DpnII digested and undigested libraries was calculated using the *deltaCt* command from the NormqPCR package (Perkins et al., 2012) with the formula:

$$\Delta Ct = 2^{(Ct_{\text{Undigested}} - Ct_{\text{Digested}})} \tag{2.3}$$

To calculate restriction fragment fold change, the ΔCt values were normalised using

controls - restriction fragments with zero fold change - and the difference in ΔCt values (abbreviated ΔΔCt) between Dam and Dam-Oct4 samples was calculated using the *deltadeltaCt* command from the NormqPCR package (Perkins et al., 2012) with the formula:

$$\Delta\Delta Ct = 2^{(\Delta Ct_{\text{Dam}} - \Delta Ct_{\text{Dam-Oct4}})} \tag{2.4}$$

2.4 Analysis of ChIP-seq experiments

2.4.1 Statement of attribution

The previously published ChIP-seq experiments were performed by the relevant research groups and re-analysed by the author, James Ashmore.

2.4.2 Availability of sequencing data

Raw sequencing data from published studies was downloaded from either the ArrayExpress, ENA, ENCODE or SRA database. This included Oct4 ESC ChIP-seq libraries from PRJNA185339 (Aksoy et al., 2013), PRJNA127937 (Ang et al., 2011), PRJNA242533 (Buecker et al., 2014), PRJNA106455 (Chen et al., 2008), PRJNA356297 (Chronis et al., 2017), PRJNA185048 (Das et al., 2014), PRJNA284634 (Flynn et al., 2016), PRJNA242892 (Galonska et al., 2015), ENCSR392DGA (Yue et al., 2014), PRJNA151337 (Hu et al., 2013), PRJNA269747 (Jacinto, Benner, and Hetzer, 2015), PRJNA153273 (Jang et al., 2012), PRJEB1833 (Karwacki-Neisius et al., 2013), PRJNA347885 (King and Klose, 2017), PRJNA252515 (Krishnakumar et al., 2016), PRJNA299026 (Liu and Kraus, 2017), PRJNA358612 (Liu et al., 2017), PRJNA106023 (Marson et al., 2008), PRJEB13059 (Miller et al., 2016), PRJDB4490 (Okashita et al., 2016), PRJNA336049 (Shen et al., 2017), PRJNA312531 (Shin et al., 2016), PRJNA291779 (Tu et al., 2016),

PRJNA213131 (Wang et al., 2014), PRJNA189323 (Whyte et al., 2013), PRJNA272971 (Xu et al., 2015), PRJEB6095 (Yang et al., 2014); Sox2 ESC ChIP-seq libraries from PRJNA356297 (Chronis et al., 2017), PRJNA242892 (Galonska et al., 2015), and PRJNA106023 (Marson et al., 2008); and Sox2 NSC ChIP-seq libraries from PRJNA152295 (Lodato et al., 2013), PRJEB5253 (Mateo et al., 2015), and PRJNA286988 (Mistri et al., 2015).

2.4.3 Collection of sequencing data

A comprehensive list of Oct4 ESC ChIP-seq samples was obtained by querying public sequence databases. The ArrayExpress (abbreviated AE) (Kolesnikov et al., 2015), European Nucleotide Archive (abbreviated ENA) (Silvester et al., 2015), and Encyclopedia of DNA Elements (abbreviated ENCODE) (Sloan et al., 2016) databases were queried using a Representational State Transfer (abbreviated REST) application programming interface (abbreviated API). The Gene Expression Omnibus (abbreviated GEO) (Edgar, Domrachev, and Lash, 2002) and Sequence Read Archive (abbreviated SRA) (Leinonen et al., 2011) databases were queried using the GEOmetadb (version 1.42.0) (Zhu et al., 2008) and SRAdb (version 1.42.2) (Zhu et al., 2013) packages, respectively. To avoid missing samples, multiple terms for a single phrase were used to query the metadata. The Ontology Lookup Service (abbreviated OLS) (Jupp et al., 2015) was used to search for terms related to the query phrases *POU5F1* and *EMBRYONIC STEM CELL*: the Ontology of Genes and Genomes (abbreviated OGG) (Liu, Zhao, and He, 2016) database was queried with the phrase *POU5F1*, and the list of returned terms included *OCT-3*, *OCT-4*, *OCT3*, *OCT4*, *OTF-3*, *OTF3*, and *OTF4*; and the Cell Ontology (abbreviated CL) (Diehl, 2017) database was queried with the phrase *EMBRYONIC STEM CELL*, and the list of returned terms included *EMBRYONIC CELL*, *ES CELL*, *ESC*, and *STEM CELL*. For each search result, the project submission page

and cited research article was manually checked to verify the target proteins, cell populations, and experimental procedures were relevant. After manual curation, a total of 31 Oct4 ESC ChIP-Seq samples were used for analysis.

2.4.4 Processing of sequencing data

Quality control

The raw sequenced reads were downloaded from the SRA database using a parallel version of the fastq-dump command from the SRA Toolkit (version 2.8.2) (Kodama et al., 2012) called parallel-fastq-dump (version 0.6.2) (Valieris, 2018). The quality of the reads was evaluated from reports generated by FastQC (Andrews, 2010) and summarised using MultiQC (Ewels et al., 2016) for easier comprehension. The reads were then trimmed using Cutadapt (Martin, 2011) to remove low quality bases and adapter contaminant sequences:

1. Illumina TruSeq Adapters

   Universal AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

   Index GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[N]ATCTCGTATGCCGTCTTCTGCTTG

After adapter and quality trimming, new FastQC reports were generated to judge the quality of the trimmed reads before proceeding to read alignment.

Read alignment

The trimmed reads were aligned to the mm10 assembly of the mouse genome (Karolchik et al., 2004) using BWA-MEM (Li, 2013) with default settings. Duplicate reads were identified with the MarkDuplicates command from Picard Tools (Broad Institute, 2018) and alignments were sorted and indexed using Samtools (Li et al., 2009). To produce a set of high quality alignments, the BAM files were filtered using Bedtools (Quinlan,

2002) and Samtools (Li et al., 2009) based on multiple criteria: Alignments to alternate, mitochondrial, unplaced, and random chromosomes were removed; Non-unique, secondary, duplicate, and supplementary alignments were removed; and alignments to ENCODE (ENCODE Project Consortium, 2012) blacklist regions were removed.

Genome coverage

To visualise read coverage along the genome, the read depth at each base pair was calculated and reported in bedGraph format using the *callpeak* command from MACS2 (Zhang et al., 2008). The read depth values in the ChIP and input bedGraph files were then compared to produce a log2 of the ratio or the subtracted difference using the *bdgcmp* command from MACS2. For faster display performance and to save storage space, the plain text bedGraph files were converted into an indexed binary bigWig file (Kent et al., 2010) using the *bedGraphToBigWig* command (version 4.0) from the Kent Tools software collection (Kuhn, Haussler, and Kent, 2013). Genome browser images of read coverage and peak locations were captured using the Integrative Genomics Viewer (version 2.3.98) (Robinson et al., 2011) and pyGenomeTracks (version 3.0.0) (Ramírez et al., 2016).

2.4.5 Calculation of quality metrics

The quality of each ChIP-seq library was measured using a combination of metrics developed by the CISTROME platform (Liu et al., 2011) and ENCODE consortia (Landt et al., 2012):

Sequence quality

The read number, average read length, and average Phred quality score was calculated for each library using FastQC (Andrews, 2010). Libraries with an average Phred

quality score >= 25 were considered to have a good sequence quality.

Mapping quality

The number of mapped, unmapped, unique, duplicate, useable, blacklist, filtered, and peak alignments were counted using Samtools (Li et al., 2009). Blacklist Number of reads mapped onto the ENCODE blacklist regions (ENCODE Project Consortium 2012) for the mm10 assembly

Libraries with an average Phred quality score >= 25 were considered to have a good sequence quality.

Library complexity

Library complexity was measured using the Non-Redundant Fraction (NRF) and PCR Bottlenecking Coefficients 1 and 2 (PBC1 and PBC2). ChIP libraries with a NRF > 0.9, PBC1 > 0.9, and PBC2 > 10 were considered to have an acceptable library complexity.

ChIP enrichment

Normalised strand cross-correlation coefficient (NSC), relative strand cross-correlation coefficient (RSC), and average fragment length were calculated using ccQualityControl (version 1.11) (Marinov et al., 2014).

2.4.6 Identification of DNA-binding sites

Peak calling

In order to identify DNA-binding sites, enriched ChIP regions relative to the input background were located using MACS2 (Zhang et al., 2008). Statistically significant peaks were defined based on a *false discovery rate* (abbreviated FDR) significance level:

narrow peaks of punctuate enrichment (Oct4, Sox2, H3, and H3.3) were called using a 0.05 FDR threshold; broad peaks of diffuse enrichment (H3K27me3, H3K36me3, H3K79me2, H3K9me3) were called using a 0.1 FDR threshold; and gapped peaks of compact enrichment (H3K27ac, H3K4me1, H3K4me2, H3K4me3, and H3K9ac) were called using a 0.1 FDR threshold. For more reliable region identification and better summit resolution, the signal profile was smoothed by extending read length to the average fragment size, which was estimated by cross-correlation analysis using SPP (Kharchenko, Tolstorukov, and Park, 2008) (see Listings 2.1 and 2.2).

```
$ macs2 callpeak --treatment <TREATMENT> --control <CONTROL> --format BAM --gsize mm
    --keep-dup all --outdir <OUTDIR> --name <NAME> --bdg --SPMR --nomodel --shift 0
    --extsize <EXTSIZE> -q 0.05
```

LISTING 2.1. Command arguments for narrow peak calling using MACS2

```
$ macs2 callpeak --treatment <TREATMENT> --format BED --gsize mm --keep-dup all --outdir
    <OUTDIR> --name <NAME> --bdg --SPMR --nomodel --shift 0 --extsize <EXTSIZE> --broad
    --broad-cutoff 0.1
```

LISTING 2.2. Command arguments for broad and gapped peak calling using MACS2

Peak reproducibility

The majority of published ChIP-seq experiments analysed in this work were not replicated, so reproducible peaks within studies could not be identified using post-hoc measurements such as the *irreproducible discovery rate* (abbreviated IDR). In order to

treat libraries from different experiments equally, libraries from replicated ChIP-seq experiments were merged to create a single library and peaks were called on the merged library. Reproducible peaks across studies were generated by first building a master list of non-overlapping regions from all inputs using BEDOPS (Neph et al., 2012). Peaks that initially overlap were ranked by score, and the highest scoring peak was then added to the master list. The original peaks from each study were then replaced by the "master" peak and only those found in multiple studies were used for downstream analysis. The source code used to create the master list was adapted from the BEDOPS tutorial: `https://bedops.readthedocs.io/en/latest/content/usage-examples/master-list.html`

Peak comparisons

The presence of peaks across conditions (binary analysis) was calculated using the *intersect* command from Bedtools (Quinlan, 2002) and visualised using proportional euler diagrams generated using the eulerr package (Larsson, 2018). Given these diagrams suppose a one-to-one relationship between set elements - which is usually false because a peak in one study can overlap two peaks in a different study - unique peaks were defined as those which do not overlap any peak in another set of peaks, and common peaks were defined as those which overlap any number of peaks in another set of peaks which afterwards are merged to produce a set of non-overlapping regions.

To visualise peak occupancy between multiple experiments, a binary heatmap depicting common and unique peaks was generated. To start with, a binary matrix of intersecting peaks was created using the *multiinter* command from Bedtools (Quinlan, 2002). Then, the distances between the rows and columns of the binary matrix were calculated using the *parDist* command with the "binary" distance method from parallelDist (Eckert, 2017). Next, the rows and columns of the distance matrix were

clustered using the *hclust* command with the "complete" agglomeration method from fastcluster (Müllner and Others, 2013). Finally, the clustered binary matrix was plotted using pheatmap (Kolde, 2018).

2.4.7 Functional analysis of DNA-binding sites

Peak annotation

To categorise the global distribution of DNA-binding sites, called peaks were associated with functionally relevant genomic regions using the *annotatePeaks* command from Homer (Heinz et al., 2010). Genes were characterised using the UCSC Genes annotation table for the mm10 assembly of the mouse genome (Karolchik et al., 2004). Peaks were assigned to the nearest gene TSS (either upstream or downstream) and genomic feature (either TSS, TTS, CDS, Exons, 5' UTR, 3' UTR, CpG Islands, Repeats, Introns, or Intergenic) occupied by its centre. It is important to note that because non-unique alignments were filtered from all of the sequencing data, coverage at regions such as interspersed repeats, low complexity, and tandem repeat regions will be artifically reduced. This is particularly important to remember when interpreting the chromatin state model results from ChromHMM in Chapter 2, and as such this issue has also been documented in the relevant sections to serve as a reminder for the reader.

Target prediction

To determine how gene regulation is affected by DNA-binding, ChIP-seq and RNA-seq data were integrated using BETA (Wang et al., 2013). Target genes and activating/repressive functions were predicted using differential gene expression data from a knock-out Oct4 ESC ChIP-seq experiment (King and Klose, 2017).

Gene ontology analysis

In order to interpret the functional profile of a set of peaks, *gene ontology* (abbreviated GO) analysis was performed with the GREAT algorithm (McLean et al., 2010) using the rGREAT package (Gu, 2017). Genomic regions were associated to genes using the basal plus extension rule, and a background set of peaks from an Oct4 ESC ChiP-seq knockout experiment (King and Klose, 2017) was provided as control regions.

2.4.8 Sequence analysis of DNA-binding sites

De novo motif discovery

Motif discovery and enrichment was performed on 500 base pair regions centred on ChIP-seq peak summits using MEME-ChIP (version 4.12.0) (Ma, Noble, and Bailey, 2014). To reduce the influence of repeat sequences on de novo motif finding, peak regions were masked with Repeat Masker (Tarailo-Graovac and Chen, 2009), Window Masker (Morgulis et al., 2006), and Tandem Repeats Finder (Benson, 1999) annotations for the mm10 assembly of mouse genome (Karolchik et al., 2004). To improve the likelihood of identifying true motif occurrences, a background model and position specific priors were generated from a knock-out ChIP-seq library (King and Klose, 2017) and ATAC-seq library (Chronis et al., 2017), respectively. Discovery was performed on both DNA strands and motifs with an E value less than a 0.05 significance level were defined as statistically significant.

Known motif search

Peak regions were scanned for occurrences of a given motif with *position weight matrices* (abbreviated PWM) from the JASPAR 2018 core non-redundant motif database (Khan et al., 2018) using the *FIMO* command (Grant, Bailey, and Noble, 2011) from the MEME suite (Ma, Noble, and Bailey, 2014). Searches were performed on both DNA

strands and matches with a P value less than a 0.0001 significance level (default for FIMO) were defined as statistically significant. The PWM accession numbers used to scan for the Oct4 and Sox2 motifs are MA0142.1 and MA0143.3, respectively.

Sequence conservation

To examine the sequence conservation over peak regions, the phastCons scores (Siepel et al., 2005) for the mm10 assembly of the mouse genome (Karolchik et al., 2004) were mapped onto overlapping peaks using the *computeMatrix* and *plotProfile* commands from Deeptools (Ramírez et al., 2016). Conservation profiles were drawn 250 bp upstream and downsteam of the reference point (defined as the peak centre) using 10 bp bins. For bins without overlapping conservation scores, missing data was replaced by zero values. The phastCons score track was downloaded through the GenomicScores package (Castelo, 2018), exported as a bedGraph file with the rtracklayer package (Lawrence, Gentleman, and Carey, 2009), and converted to a bigWig file using the *bedGraphToBigWig* command from the Kent Tools software collection (Kuhn, Haussler, and Kent, 2013)

2.4.9 Identification of chromatin states

State modelling

Chromatin models for ESC and MEF cell types were built using ChromHMM (version 1.14) (Ernst and Kellis, 2017) with default settings. First, alignments from ATAC-seq, ChIP-seq, and DamID-seq experiments were binarized using the *BinarizeBed* command. To better represent their enrichment patterns: ATAC-seq alignments were shifted by -36 bp and extended to a size of 73 bp; ChIP-seq alignments were extended to the average fragment size; and DamID-seq alignments were extended to the nearest restriction site. Second, a single shared model with cell type specific annotations was

derived using the *LearnModel* command. An 18 state model was decided upon (after visually inspecting models with up to 30 states) as it provided an easily interpretable yet multi-faceted segmentation of the chromatin landscape.

State profiling

Chromatin states were profiled by calculating an enrichment score for each state over multiple genomic annotations using the *OverlapEnrichment* and *NeighbourhoodEnrichment* commands. For overlap enrichment analysis, genomic annotations were generated from either the AnnotationHub package (Morgan, 2017) or the TxDb.Mmusculus.UCSC.mm10.knownGene annotation database (Team and Maintainer, 2016) using the GenomicFeatures package (Lawrence et al., 2013). The genomic annotations used in the analysis were: CDS, CpG, Exons, Gap, Genes, Intergenic, Introns, Microsatellite, Promoters, RepeatMasker, SimpleRepeats, TES, TSS, Transcripts, 3′ UTR, 5′ UTR, and WindowMasker. For neighbourhood enrichment analysis, the position of each restriction site in the mm10 assembly of the mouse genome (Karolchik et al., 2004) was generated using the *matchPattern* command from the Biostrings package (Pagès et al., 2017).

2.4.10 Development of analysis workflow

Processing of the raw sequencing data was automated using a custom pipeline built with the Conda package manager (version 4.5.1) (Anaconda, 2018) and Snakemake workflow engine (version 4.7.0) (Köster and Rahmann, 2012). The analyses described in this section can be reproduced by downloading and running the relevant ChIP-seq workflow.

2.5 Analysis of ATAC-seq, DNase-seq, and FAIRE-seq experiments

2.5.1 Statement of attribution

The previously published ATAC-seq, DNase-seq, and FAIRE-seq experiments were performed by the relevant research groups and the raw sequencing data was re-analysed by the author, James Ashmore.

2.5.2 Availability of sequencing data

The raw sequencing data is available from the SRA database. The BioProject accession numbers include ESC ATAC-seq from PRJNA356293 (Chronis et al., 2017), PRJNA279456 (Maza et al., 2015), and PRJNA369204 (Simon et al., 2017); MEF ATAC-seq from PRJNA356293 (Chronis et al., 2017), PRJNA359484 (Li et al., 2017), and PRJNA279456 (Maza et al., 2015); ESC DNase-seq from PRJNA281090 (Domcke et al., 2015), PRJNA233390 (Sherwood et al., 2014), and SRP015984 (Yue et al., 2014); MEF DNase-seq from PRJNA269282 (Deng et al., 2015), PRJEB21708 (Herdman et al., 2017), and SRP015984 (Yue et al., 2014); ESC FAIRE-seq from PRJNA242533 (Buecker et al., 2014), PRJNA272126 (Dieuleveult et al., 2016), and PRJNA252824 (Murtha et al., 2015); and MEF FAIRE-seq from PRJNA252824 (Murtha et al., 2015), PRJNA276442 (Schick et al., 2015), and PRJNA188177 (Wapinski et al., 2013).

2.5.3 Processing of sequencing data

Quality control

The raw sequencing data was downloaded from the SRA database using the parallel-fastq-dump command (version 0.6.2) (Valieris, 2018) command from the SRA Toolkit (version 2.8.2) (Kodama et al., 2012). The read quality was evaluated from reports

generated by FastQC (version 0.11.7) (Andrews, 2010) and MultiQC (version 1.4.0) (Ewels et al., 2016). The reads were trimmed using Cutadapt (Martin, 2011) to remove low quality bases and adapter contaminant sequences:

1. Illumina TruSeq Adapters

   Universal AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT

   Index GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[N]ATCTCGTATGCCGTCTTCTGCTTG

2. Illumina Nextera Adapters

   Read1 TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG

   Read2 GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG

   Index1 CAAGCAGAAGACGGCATACGAGAT[N]GTCTCGTGGGCTCGG

   Index2 AATGATACGGCGACCACCGAGATCTACAC[N]TCGTCGGCAGCGTC

Before proceeding, new FastQC and MultiQC reports were generated to make sure all low quality bases and adapter sequences were removed.

Read alignment

The trimmed reads were aligned to the mm10 assembly of the mouse genome (Karolchik et al., 2004) using BWA-MEM (Li, 2013). Duplicate reads were identified using the *MarkDuplicates* command from Picard Tools (Broad Institute, 2018) and alignments were sorted and indexed using Samtools (Li et al., 2009). The alignment files were filtered using Bedtools (Quinlan, 2002) and Samtools (Li et al., 2009) based on multiple criteria: Alignments to alternate, mitochondrial, unplaced, and random chromosomes were removed; non-unique, secondary, duplicate, and supplementary alignments were removed; and alignments to ENCODE (ENCODE Project Consortium, 2012) and mitochondrial (Buenrostro et al., 2013) blacklist regions were removed. To

better represent the centre of the transposon binding event in ATAC-seq experiments, alignments to the plus and minus strands were shifted 4 bp and -5 bp, respectively.

Genome coverage

To visualise read coverage along the genome, the read depth at each base pair was calculated and reported in bedGraph format using the *callpeak* command from MACS2 (Zhang et al., 2008). For faster display performance and to save storage space, the plain text bedGraph files were converted into an indexed binary bigWig file (Kent et al., 2010) using the *bedGraphToBigWig* command (version 4.0) from Kent Tools (Kuhn, Haussler, and Kent, 2013). Genome browser images of read coverage were captured using the Integrative Genomics Viewer (version 2.3.98) (Robinson et al., 2011) and pyGenomeTracks (version 3.0.0) (Ramírez et al., 2016).

2.5.4 Identification of DNA-accessibility sites

Peak calling

In order to identify DNA-accessibility sites, enriched chromatin regions relative to the background were located using MACS2 (version 2.1.1) (Zhang et al., 2008). Broad peaks of diffuse enrichment (ATAC-seq, DNase-seq, FAIRE-seq, and ChIP-seq) were called with a 0.1 FDR threshold. To better represent the centre of each accessibility site, alignments were shifted by -36 bp and extended to a size of 73 bp corresponding to the length of DNA wrapped around a single nucleosome (see Listing 2.3).

```
macs2 callpeak --treatment <TREATMENT>
               --format BED
               --gsize mm
               --keep-dup all
```

```
--outdir <OUTDIR>
--name <NAME>
--bdg
--SPMR
--nomodel
--shift -36
--extsize 73
--broad
--broad-cutoff 0.1
```

LISTING 2.3. Parameters for broad peak calling using MACS2

Peak reproducibility

The majority of published ATAC-seq, DNase-seq, and FAIRE-seq experiments analysed in this work were not replicated. In order to treat libraries from different experiments equally, libraries from replicated experiments were merged to create a single library and peaks were called on the merged library.

For experiments with biological replicates, alignments from all replicates were merged to create a single bulk replicate, in agreement with un-replicated experiments. The broadPeak file format was used to represent broad regions of enrichment typically seen in accessibility experiments. Command line parameters used are listed below:

Peak comparisons

Reproducible regions were generated by comparing peaks from three different studies using BEDOPS (version 2.4.3) (Neph et al. 2012) with a 1 bp overlap threshold. Firstly, a master list of regions was generated by ranking overlapping peaks from each study and choosing the peak with the highest score. The original peaks from each study

were then replaced by the overlapping peaks in the master list and only those present in all studies were used for downstream analyses. Code used to create the master list was adapted from the website listed below:

https://bedops.readthedocs.io/en/latest/content/usage-examples/master-list.html

Differential accessibility

To identify differential DNA-accessibility sites, binding analysis was performed with coverage data using DiffBind (version 2.6.5) (Ross-Innes et al., 2012). Sites with a FDR < 0.1 and absolute $\log_2$FC > 1 were defined as statistically significant. For comparisons between technologies, bulk replicates from each study and reproducible peaks between studies were used.

2.5.5 Visualisation of DNA-accessibility sites

Peak accessibility

To visualise DNA-accessibility at, heatmaps of read coverage over broadPeak regions were produced using the *computeMatrix* and *plotHeatmap* commands from Deeptools (Ramírez et al., 2016). Reads within 4 kb of the enhancer centre were counted into 10 bp bins, then the rows of the heatmap were sorted by average read coverage.

Enhancer accessibility

To visualise DNA-accessibility at enhancers, heatmaps of read coverage over H3K4me1 and H3K27ac chromatin states (see Subsection 2.4.9) were generated using the *computeMatrix* and *plotHeatmap* commands from Deeptools (Ramírez et al., 2016). Reads within 4 kb of the enhancer centre were counted into 10 bp bins, then the rows of the heatmap were sorted by average read coverage. In addition, super-enhancers from from a previously published experiment (Whyte et al., 2013) were visualised. The

super-enhancer coordinates were re-mapped from the mm9 to mm10 assembly of the mouse genome (Karolchik et al., 2004) using the *liftOver* command (version 4.0) from Kent Utilities (Kuhn, Haussler, and Kent, 2013).

Promoter accessibility

To visualise DNA-accessibility at promoters, heatmaps of read coverage around transcription start sites (abbreviated TSS) were generated using the *computeMatrix* and *plotHeatmap* commands from Deeptools (Ramírez et al., 2016). An annotation file of TSS locations was generated from the TxDb.Mmusculus.UCSC.mm10.knownGene annotation database (version 3.4.0) (Team and Maintainer, 2016) using the GenomicFeatures package (version 1.30.3) (Lawrence et al., 2013). Reads within 2 kb of a TSS were counted into 10 bp bins, then the rows of the heatmap were sorted by normalised gene expression (see Subsection 2.6.4).

2.5.6 Development of analysis workflow

Processing of the raw sequencing data was automated using a custom pipeline built with the Conda package manager (version 4.5.1) (Anaconda, 2018) and Snakemake workflow engine (version 4.7.0) (Köster and Rahmann, 2012). The analyses described in this section can be reproduced by downloading and running the relevant ATAC-seq, DNase-seq, and FAIRE-seq workflows.

2.6 Analysis of RNA-seq experiments

2.6.1 Statement of attribution

The previously published RNA-seq experiments were performed by the relevant research groups and the raw sequencing data was re-analysed by the author, James Ashmore.

2.6.2 Availability of sequencing data

The raw sequencing data is available from the SRA database. The BioProject accession numbers are PRJNA286869 (Milagre et al., 2017) and PRJNA347884 (King and Klose, 2017).

2.6.3 Processing of sequencing data

Quality control

The raw sequencing data was downloaded from the SRA database using the *parallel-fastq-dump* command (version 0.6.2) (Valieris, 2018) from the SRA Toolkit (version 2.8.2) (Leinonen et al., 2011). The read quality was evaluated from reports generated by FastQC (version 0.11.7) (Andrews, 2010) and MultiQC (version 1.5) (Ewels et al., 2016). The reads were trimmed using Cutadapt (version 1.16) (Martin, 2011) to remove low quality bases and adapter contaminant sequences:

1. Illumina TruSeq Adapters

Universal	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATCT
Index	GATCGGAAGAGCACACGTCTGAACTCCAGTCAC[N]ATCTCGTATGCCGTCTTCTGCTTG

Before proceeding, new FastQC and MultiQC reports were generated to make sure all low quality bases and adapter sequences were removed.

Transcript quantification

To quantify transcript abundance, trimmed reads were pseudo-aligned to the UCSC mm10 assembly of the mouse transcriptome (Karolchik et al., 2004) using the *quant* command from Kallisto (version 0.43.1) (Bray et al., 2016). Transcript sequences (n = 63,759) were generated from the TxDb.Mmusculus.UCSC.mm10.knownGene annotation database (version 3.4.0) (Team and Maintainer, 2016) using the *extractTranscriptSeqs* command from the GenomicFeatures package (version 1.30.3) (Lawrence et al., 2013).

Gene summarisation

To generate gene-level count matrices, transcript abundance estimates were summarised per gene using the tximport package (version 1.6.1) (Soneson, Love, and Robinson, 2015). Transcripts were mapped to genes (n = 24,116) by the *GENEID* and *TXNAME* columns from the TxDb.Mmusculus.UCSC.mm10.knownGene annotation database (version 3.4.0) (Team and Maintainer, 2016).

2.6.4 Identification of differentially expressed genes

Quality control

Gene-level count matrices were imported into R (version 3.4.3) (R Core Team, 2017) and assembled into DESeqDataSet objects for processing using the DESeq2 package (version 1.18.1) (Love, Huber, and Anders, 2014). Genes with very low counts (less than 10) across all samples were filtered because there was no purpose in analysing genes that were not expressed. In addition, fewer genes reduced the processing time and severity of multiple testing correction. To compare expression levels among samples, gene counts were normalised using the median ratio method from the *estimateSizeFactors* function. Between and within sample group differences were evaluated

from principal component analysis (abbreviated PCA) and Brand-Altman (abbreviated MA) plots generated using the *plotPCA* and *plotMA* functions.

Differential expression

Differentially expressed genes (abbreviated DEG) were identified using the standard procedure implemented in the *DESeq2* function: estimate size factors, estimate dispersions, fit generalised linear models (abbreviated GLM) based on the negative binomial distribution, and compute Wald statistics. Genes with a FDR < 0.05 and absolute $\log_2$ FC > 1 were defined as statistically significant. To visualise the most meaningful DEG, volcano plots and heatmaps were created using the ggplot2 (version 2.2.1) (Wickham, 2009) and pheatmap (version 1.0.8) (Kolde, 2018) packages.

Functional profiling

To characterise the molecular functions and pathways in which DEG are involved, over-representation analyses for Gene Ontology (abbreviated GO) terms and Kyoto Encyclopedia of Genes and Genomes (abbreviated KEGG) pathways were performed using the GOseq package (version 1.3.0) (Young et al., 2010). Terms and pathways with an over-represented P value < 0.005 were defined as statistically significant. To visualise the most meaningful profiles, bar plots of the -$\log_{10}$ over-represented P values for the 10 most significant terms and pathways were generated using the ggplot2 package (version 2.2.1) (Wickham, 2009)

Multiomics integration

Downstream analyses integrating RNA-seq data with other functional genomics technologies (ATAC-seq, ChIP-seq, DamID-seq, DNase-seq, and FAIRE-seq) were carried out using regularised log transformed gene counts generated using the *rlog* function.

The transformation corrects for differences in library size and reduces the contribution of lowly expressed genes by shrinking their variance.

2.6.5 Development of analysis workflow

Processing of the raw sequencing data was automated using a custom pipeline built with the Conda package manager (version 4.5.1) (Anaconda, 2018) and Snakemake workflow engine (version 4.7.0) (Köster and Rahmann, 2012). The analyses described in this section can be reproduced by downloading and running the relevant RNA-seq workflow.

Chapter 3

Bias detection and protocol optimization in DamID-seq data

3.1 Introduction

The development of DamID-seq has provided researchers with an alternative method to identify DNA binding sites in situations where conventional ChIP-seq has been limited. It has been used for a variety of DNA-binding proteins, from transcriptional regulators to chromatin and nuclear organisers (Aughey and Southall, 2016). While the majority of these experiments were performed in *Drosophila* (Greil, Moorman, and Van Steensel, 2006), considerable progress has been made in recent years adapting the technology to mammals, principally to investigate rare cell populations (Tosti et al., 2018). Whilst celebrated, these technological advancements have been applied rather uncritically. Since the development of DamID nearly 20 years ago, there has been relatively little consideration regarding the potential sources of bias in the DamID-seq experimental procedure or the sequencing data generated. Currently, known biases in the experimental procedure include modulating the level of Dam expression to

achieve a good signal to noise ratio (Greil, Moorman, and Van Steensel, 2006), methylation of the Dam plasmid in transient transfection experiments (Pindyurin et al., 2016) and the relatively low resolution of binding sites detected (Greil, Moorman, and Van Steensel, 2006). However, to the best of our knowledge there has been no systematic review or publication of biases present in the sequencing data, unlike other high-throughput technologies such as ChIP-seq (Furey, 2012; Meyer and Liu, 2014) and RNA-seq (Conesa et al., 2016; Zheng, Chung, and Zhao, 2011). This presents concerns for past results and future experiments, as technical artifacts may be misinterpreted as biologically-relevant DNA binding. Historically, the application of newly-developed technology has often exceeded our understanding of the data generated. Biases such as chromatin fragmentation, nucleic acid isolation, PCR amplification, and read mapping in next-generation sequencing experiments are continually being documented (Meyer and Liu, 2014). The accuracy and sensitivity of any given technology is greatly influenced by multiple technical and biological factors. Without investigating their impact on the data generated, technology with hidden potential may never be fully realised, or worse technology which produces false-positives may become widely adopted.

3.2 Aims

The aim of this chapter is to identify potential sources of bias in the DamID-seq experimental procedure and demonstrate how these biases affect the sequencing data generated. This is necessary so that technical variation between the Dam and Dam-fusion libraries is not misinterpreted as biologically-relevant DNA binding. Sources of bias will be surveyed using DamID-seq data for transcription factors Oct4 and Sox2 in mouse embryonic stem cells (abbreviated ESC), embryonic fibroblast cells (abbreviated MEF), and neural stem cells (abbreviated NSC). To ensure cell number biases

are not overlooked, biases will also be surveyed using Oct4 ESC DamID-seq data prepared with 10^6, 10^4, 10^3, and 10^2 cells. Information about any potential sources of bias can be used to guide future experimental design and inform the development of rigorous analysis methods.

3.3 Attribution

The DamID-seq libraries were generated by Dr Luca Tosti, and the sequencing data was analysed by the author, James Ashmore. All of the ChIP-seq libraries were taken from public experiments and re-analysed.

3.4 Results

3.4.1 DamID-seq data yields broad regions of enrichment

In order to develop rigorous computational methods for analysing DamID-seq data, it was first necessary to gain an understanding of the enrichment pattern generated by DamID-seq experiments. This was evaluated by looking at the pileup of aligned sequencing reads along the genome, particularly around regions likely to be enriched (e.g. Transcription factor binding sites). Genome browser tracks showed that Oct4 and Sox2 DamID-seq data exhibited much broader regions of enrichment around DNA binding sites than ChIP-seq data (see Figure 3.1). Previous studies have demonstrated that Dam can methylate DNA multiple kilobases away from the DNA binding site (Van Steensel and Henikoff, 2000; Van Steensel, Delrow, and Henikoff, 2001). This occurs because the tether in the fusion-protein allows Dam to diffuse locally around the DNA binding site and methylate any DNA in close proximity. To estimate the average reach of Dam, the distribution of DamID-seq reads around all DNA binding sites identified from ChIP-seq data was plotted (see Figure 3.2). The spread of

the distributions showed that DamID-seq data display localised but much broader regions of enrichment (~3 kb from the DNA binding site) than ChIP-seq data (~250bp from the DNA binding site). One of the main features of ChIP-seq data used to increase its resolution is the strand-specific structure of the reads aligned around the DNA binding site (Pepke, Wold, and Mortazavi, 2009). In ChIP-seq experiments, immunoprecipitated DNA fragments are equally likely to be sequenced from both ends, so the read density around a DNA binding site should show a bimodal enrichment pattern (i.e. reads aligned to the forward strand should be enriched upstream of the DNA binding site and vice versa) (Zhang et al., 2008). Peak calling algorithms exploit this shift to more accurately identify the location of the DNA binding site. This strand-specific structure was immediately visible in the genome browser tracks of the ChIP-seq data, but not the DamID-seq data (see Figure 3.3). To substantiate this observation, the distribution of strand-specific DamID-seq reads at DNA binding sites identified from ChIP-seq data was plotted (see Figure 3.4). The completely overlapping forward and reverse strand distributions in the DamID-seq data verified there was no strand-specific structure which could be exploited to increase its resolution. Instead it appears the resolution of DamID-seq experiments is limited by the reach of Dam and the size of the restriction fragment in which the DNA binding site is located (see Figure 3.5). In the mouse genome, the average restriction fragment length is approximately 260 bp, meaning the average resolution for each DNA binding site will be between 260 bp and 6 kb (combined upstream and downstream reach of Dam). These results imply that DamID-seq will detect DNA binding sites at a much lower resolution than ChIP-seq, which is limited to the size of chromatin fragments produced by sonication (typically between 200 and 500 bp). However, altered methods such as ChIP-exo and X-ChIP-seq claim to acquire single base pair resolution using exonucleases to degrade the DNA protruding either side of the DNA-protein complex

(Skene and Henikoff, 2015). Lower resolution binding sites also hinder many downstream analyses such as *de-novo* motif enrichment and discovery by having to scan much larger sequence regions (Ma, Noble, and Bailey, 2014).

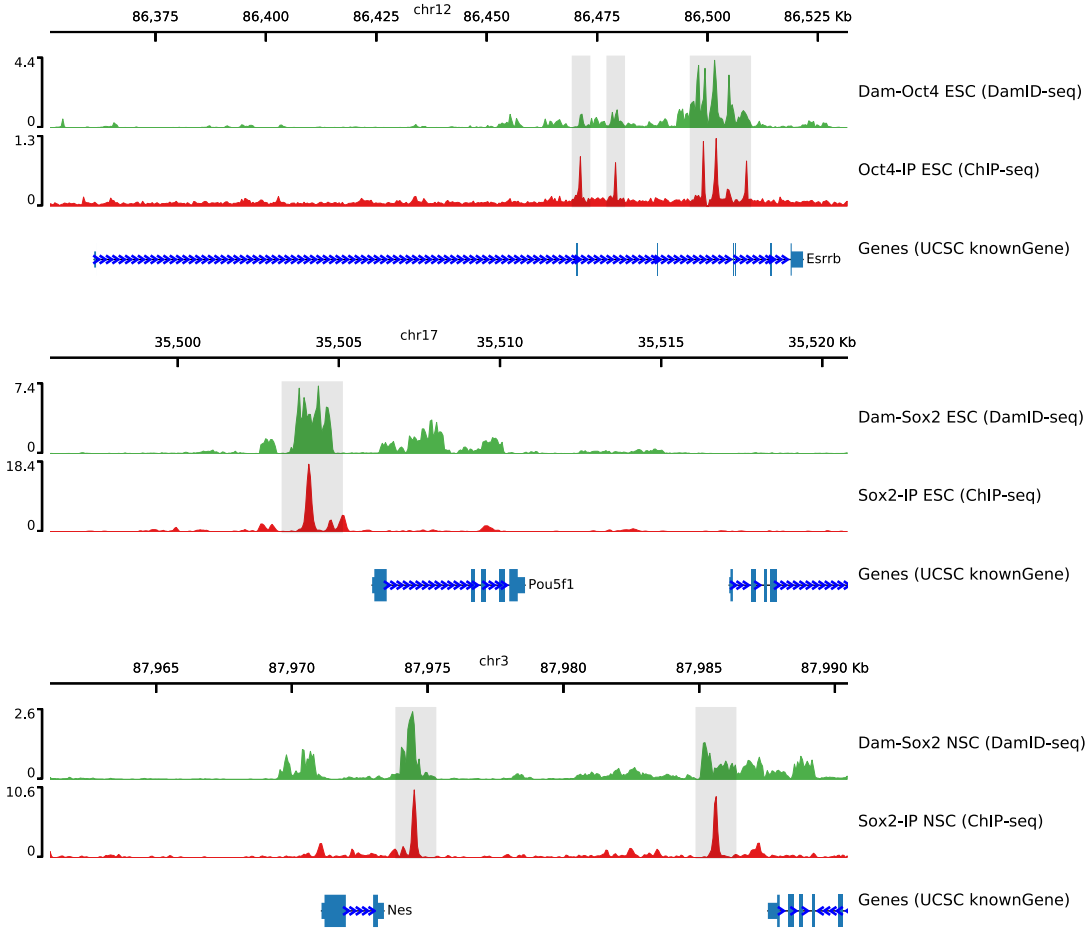

FIGURE 3.1. Tracks of ChIP-seq and DamID-seq read coverage at DNA binding sites.

These tracks show ChIP-seq and DamID-seq read coverage at select Oct4 and Sox2 DNA binding sites in mouse embryonic stem cells (abbreviated ESC) and neural stem cells (abbreviated NSC). The ChIP-seq tracks were generated by re-analysing data from public Oct4 ESC (Chronis et al., 2017), Sox2 ESC (Marson et al., 2008), and Sox2 NSC (Lodato et al., 2013) experiments.

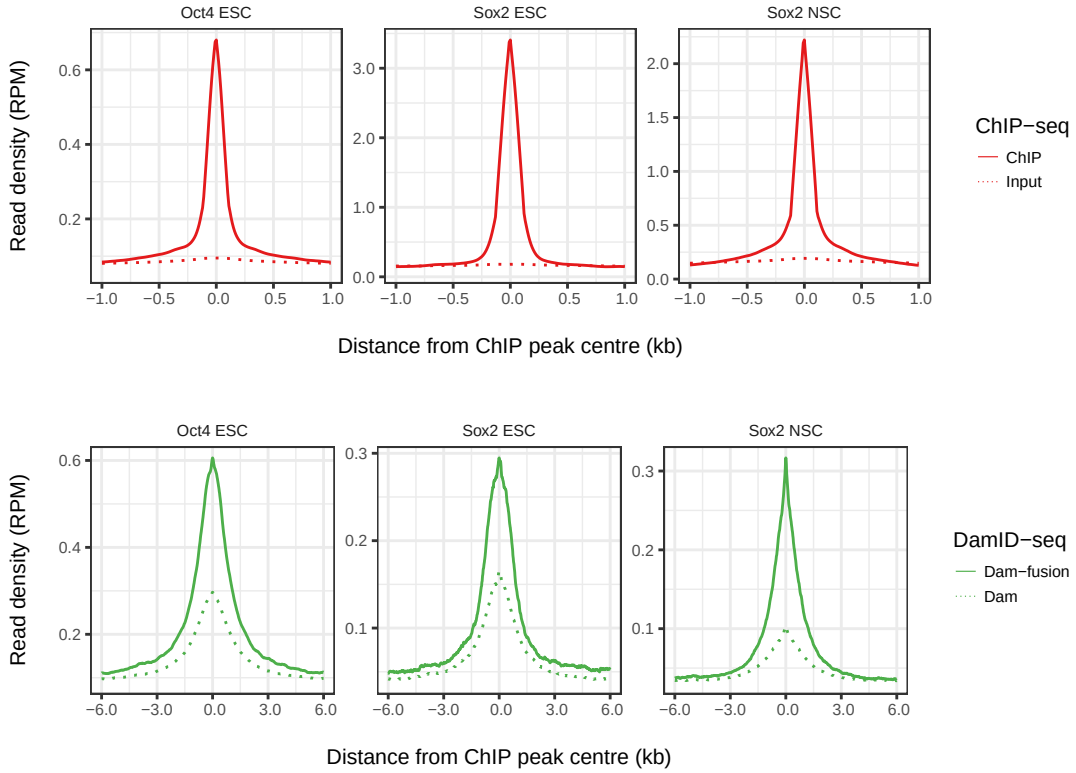

FIGURE 3.2. Graphs of ChIP-seq and DamID-seq read coverage at DNA binding sites.

These graphs show ChIP-seq and DamID-seq read coverage at all Oct4 and Sox2 DNA binding sites in mouse embryonic stem cells (abbreviated ESC) and neural stem cells (abbreviated NSC). The top panels are from three ChIP-seq experiments (Chronis et al., 2017; Marson et al., 2008; Lodato et al., 2013) and the bottom panels are from three DamID-seq experiments (Tosti et al., 2018). The DNA binding sites were identified by calling peaks from the Oct4 ESC (n = 37,925), Sox2 ESC (n = 15,690), and Sox2 NSC (n = 22,481) ChIP-seq data.

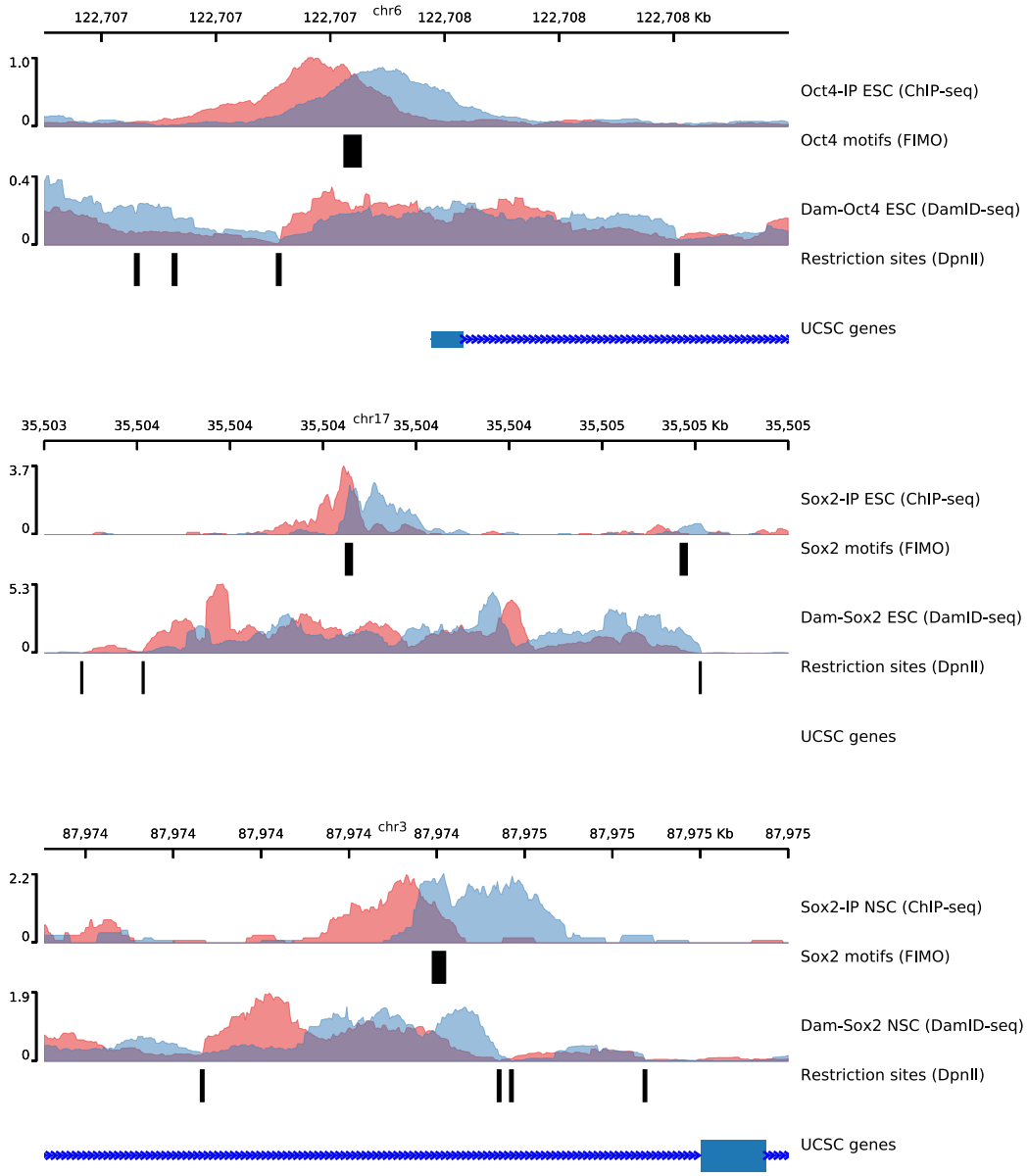

FIGURE 3.3. Tracks of strand-specific ChIP-seq and DamID-seq read coverage at DNA binding sites.

These tracks show strand-specific ChIP-seq and DamID-seq read coverage at select Oct4 and Sox2 DNA binding sites in mouse embryonic stem cells (abbreviated ESC) and neural stem cells (abbreviated NSC). Reads aligned to the forward and reverse strands are coloured red and blue, respectively. The panels also contain the location of the relevant transcription factor binding motif (identified by FIMO) and the Dam binding motif (DpnII restriction site).

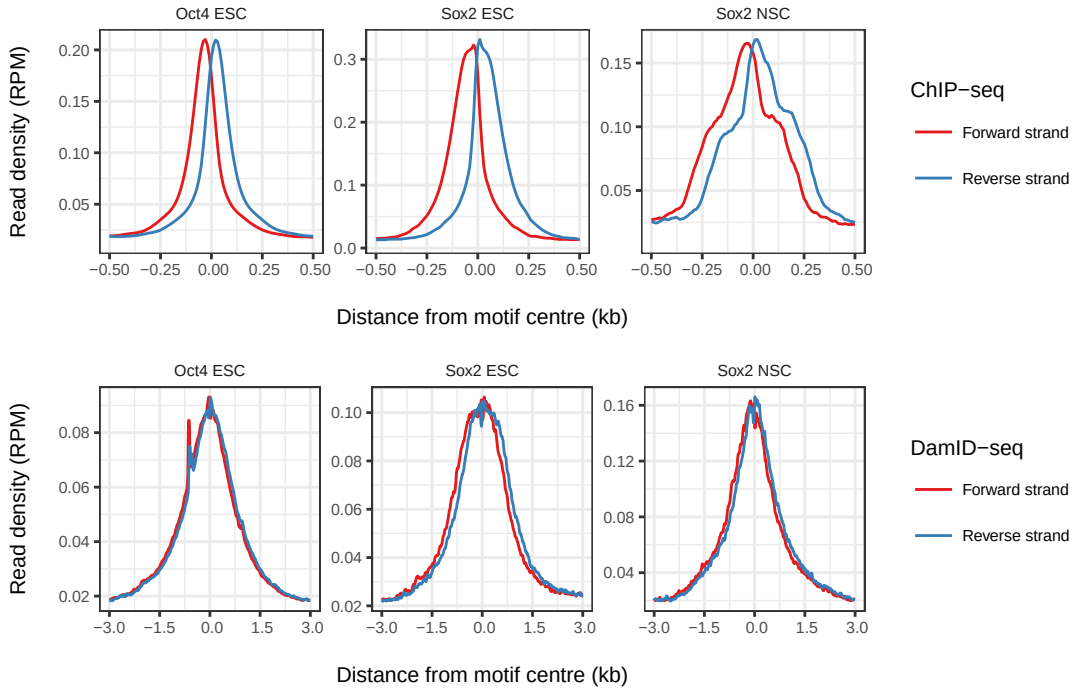

FIGURE 3.4. Graphs of strand-specific ChIP-seq and DamID-seq read coverage at DNA binding sites.

These graphs show strand-specific ChIP-seq and DaID-seq read coverage at all Oct4 and Sox2 DNA binding sites in mouse embryonic stem cells (abbreviated ESC) and neural stem cells (abbreviated NSC). The top panels are from three ChIP-seq experiments (Chronis et al., 2017; Marson et al., 2008; Lodato et al., 2013) and the bottom panels are from three DamID-seq experiments (Tosti et al., 2018). There is a characteristic shift between reads aligned to the forward and reverse strands at the DNA binding sites in the ChIP-seq experiments.

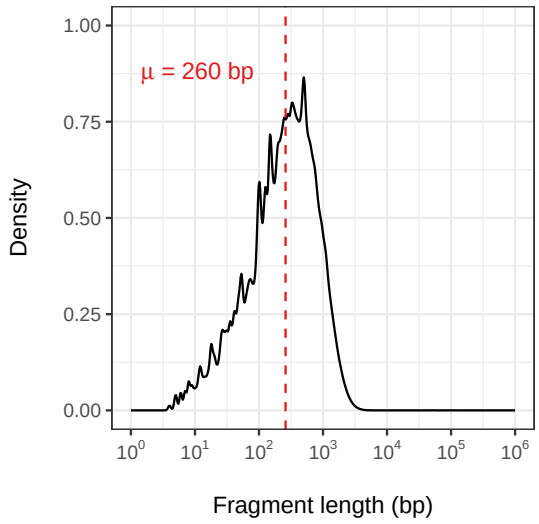

FIGURE 3.5. Distribution of restriction fragment sizes in the mouse genome.

This graph shows the distribution of restriction fragment sizes generated from an in-silico digest of the mouse genome by restriction enzyme DpnII (i.e. cutting all GATC sequences between the adenine and thymine bases). The median restriction fragment length is 260 base pairs.

3.4.2 Assessment of duplication rates in DamID-seq data

An important decision in the analysis of high-throughput sequencing data is whether to remove duplicate sequencing reads. Duplicates are classed as reads which have the same genomic sequence and consequently align to the same position in the reference genome. There are multiple ways in which duplicates can arise, categorised either by technical or natural duplication (Bansal, 2017). Technical duplicates are generated by PCR amplification, which is necessary in most library preparations to enrich for adapter-ligated DNA fragments for sequencing. This sometimes produces an altered or unrepresentative library composition because polymerase efficiency is influenced by the length and nucleotide content of the DNA fragment being amplified (Aird et al., 2011). Natural duplicates on the other hand are generated when DNA fragments with

a high copy number in the starting material are independently sequenced. In RNA-seq experiments, highly expressed genes produce thousands of copies of the same mRNA transcript which are then extracted, fragmented, and copied into cDNA for sequencing. Shorter mRNA transcripts expressed at the same level as longer ones also tend to generate more duplicates because the space of possible start and end positions for mRNA fragmentation is saturated (Parekh et al., 2016).

To remove duplicates from sequencing data, computer software locates reads with the same genomic sequence (Shen et al., 2016) or alignment position (Broad Institute, 2018) and filters all but the highest scoring read. The problem with this strategy is that it cannot determine whether the reads were generated by technical or natural duplication (i.e. spurious technical copies or legitimate biological copies). Whilst there are alternative library preparations which allow each individual DNA fragment to be uniquely marked (e.g. universal molecular identifiers and cellular barcodes), these are normally reserved for situations where a large number of PCR cycles is needed or the library complexity is small enough to accommodate the maximum number of unique barcodes which can be generated (Kivioja et al., 2011). The removal of duplicates affects the quantification of the biology under investigation, and therefore influences the results of the experiment. The general consensus for RNA-seq data is that duplicates should not be removed because the probability of natural duplication is inherently high given the level of transcription (Conesa et al., 2016). In comparison with ChIP-seq data, duplicates are routinely removed because the maximum copy number of DNA from a single cell is two (a single chromosome pair), and DNA fragmentation is unbounded so the probability of sequencing the exact same fragment is low (Carroll et al., 2014). For DamID-seq data, it is unclear what the rate of technical to natural duplication is because the maximum copy number of a restriction fragment from a single cell is two, but the possible number of DNA fragmentation sites is bounded by

the length of the restriction fragment.

In order to assess the behaviour of duplicates in DamID-seq experiments, a logistic regression model (i.e. duplicates predicted from methylation) was fitted to Oct4 and Sox2 DamID-seq data (see Figure 3.6). The models showed no relationship between the percentage of duplicate reads and the level of methylation. Instead the percentage of duplicates varied between 0% and 100% for a fixed level of methylation (e.g. 10^2 reads/kb). Only at exceedingly high methylation levels did the percentage of duplicate reads increase linearly (e.g. 10^4 reads/kb). This same pattern was also observed in the low cell number Oct4 DamID-seq data, but as the number of cells decreased a much larger proportion of restriction fragments exhibited duplicates (see 3.7). In this situation, the increased duplication rate can be explained by the higher number of PCR cycles employed in the library preparation which suggests they are technical rather than natural duplicates. For comparison, RNA-seq data which displays a high rate of natural duplication exhibits a linear relationship between expression and duplication. The DamID-seq data did not exhibit this relationship, which indicated very few natural duplicates in the DamID-seq data, and that duplicate removal should be performed. To evaluate the affect of duplicate removal, the distribution of DamID-seq reads around DNA binding sites identified from ChIP-seq data with and without duplicates removed was plotted (see 3.8). In the Sox2 ESC and Sox2 NSC DamID-seq data the removal of duplicates either increased or did not affect the read coverage at DNA binding sites. The increase in coverage can be explained by the removal of reads not associated with the DNA binding sites, therefore decreasing the overall library size and consequently when normalised to reads per million causing an increase in coverage at the DNA binding sites. Surprisingly, coverage at DNA binding sites was also not affected in the low cell number DamID-seq data (see 3.9). Together these results

indicated that the number of natural duplicates in DamID-seq data is low and consequently PCR duplicates can be safely removed without affecting coverage at DNA binding sites.

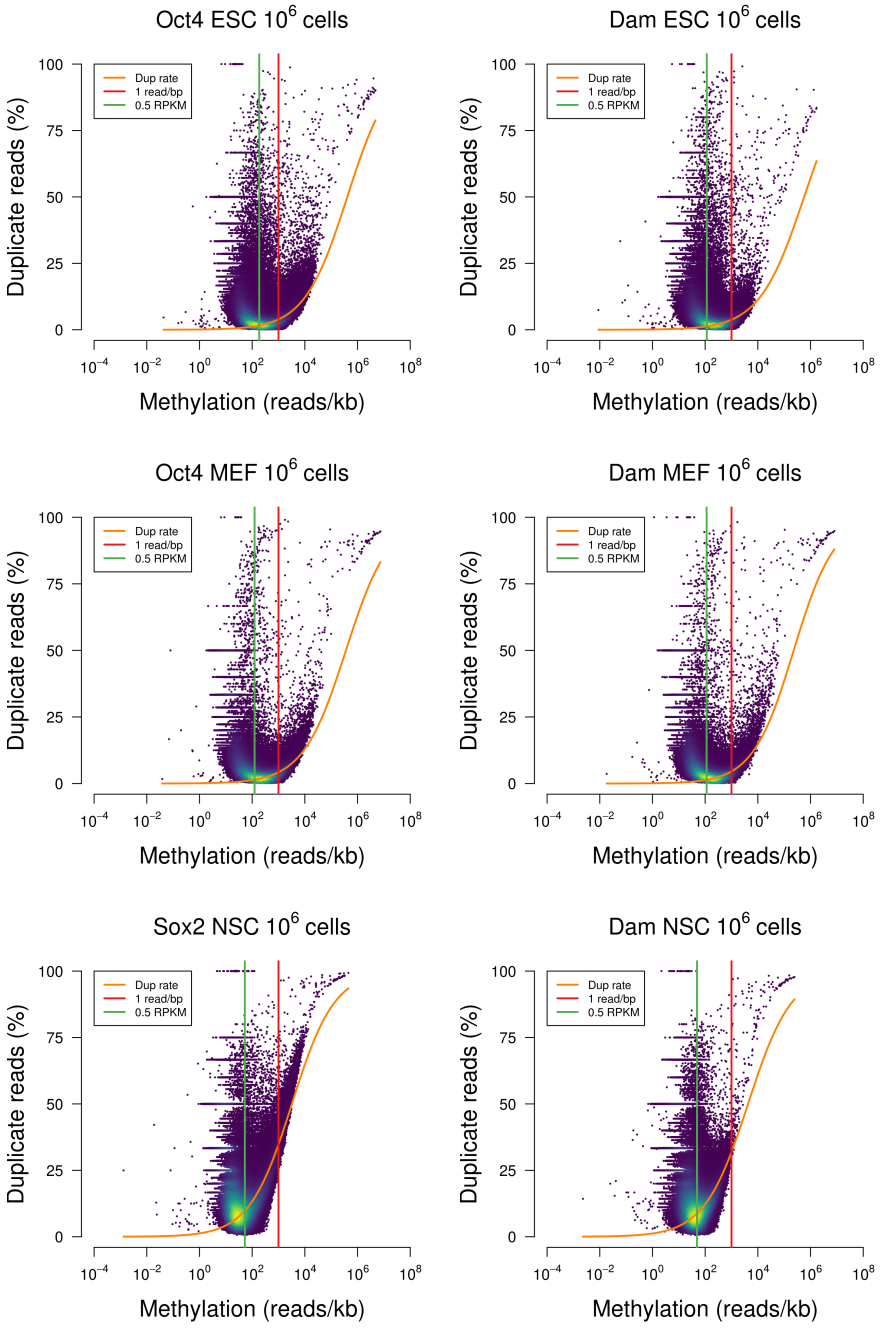

FIGURE 3.6. Graphs of duplication rates from Oct4 and Sox2 DamID-seq experiments.

These graphs show the duplication rates from Oct4 and Sox2 DamID-seq experiments in mouse embryonic stem cells (abbreviated ESC), fibroblast cells (abbreviated MEF) and neural stem cells (abbreviated NSC). The X-axis measures the number of reads per kilobase pair aligned to each restriction fragment, and the Y-axis measures the percentage of reads aligned to each restriction fragment which are marked as PCR duplicates. There is no strong relationship between methylation and duplication, meaning the number of natural duplicates is low.

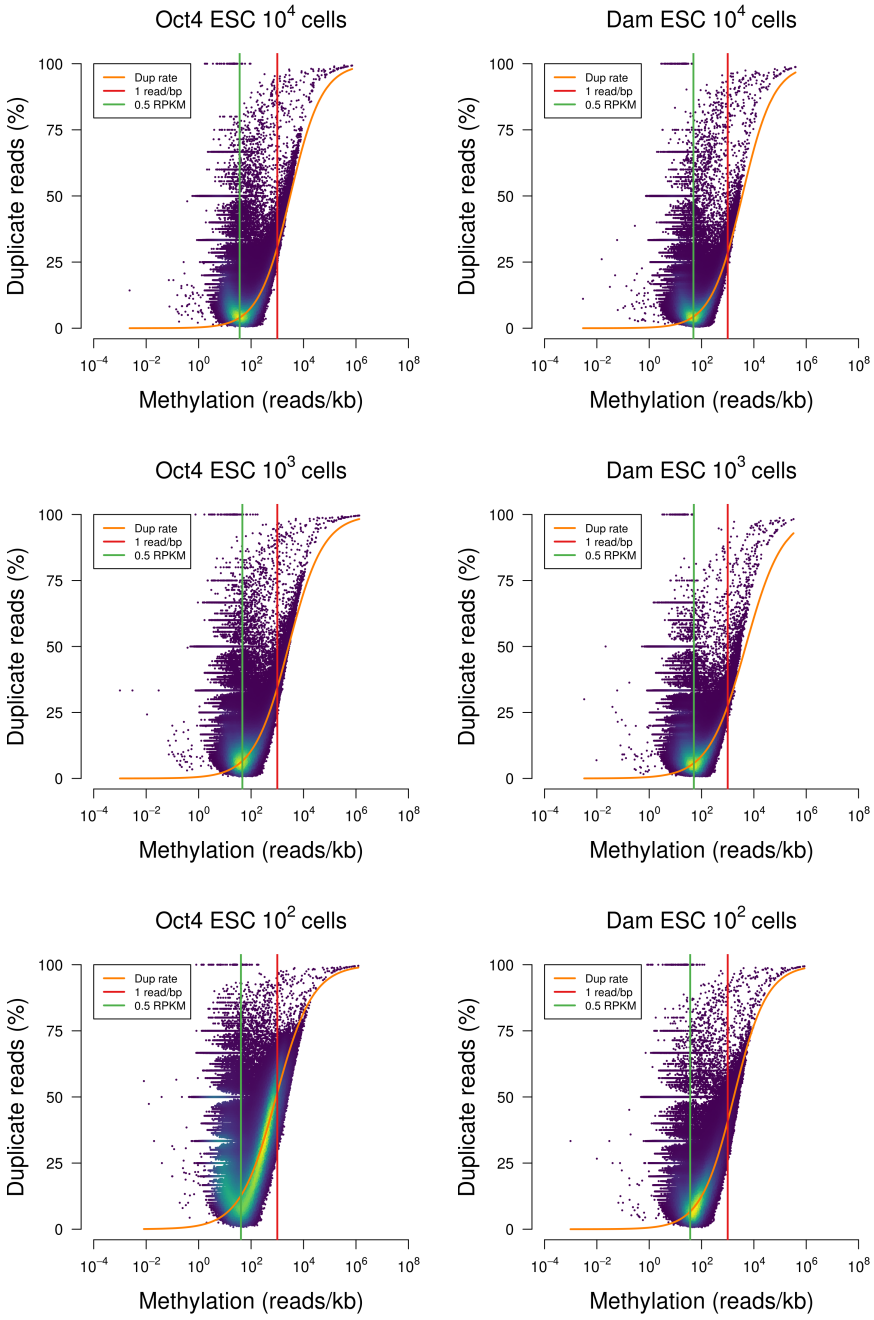

FIGURE 3.7. Graphs of duplication rates from low cell number Oct4 DamID-seq experiments.

These graphs show the duplication rates from Oct4 DamID-seq experiments in 10^6, 10^4, 10^3, and 10^2 mouse embryonic stem cells. Each point represents one restriction fragment in the mouse genome. The X-axis shows the number of reads per kilobase pair aligned to each restriction fragment, and the Y-axis shows the percentage of reads aligned to each restriction fragment which are marked as PCR duplicates. There is no strong relationship between methylation and duplication, meaning the number of natural duplicates is low.

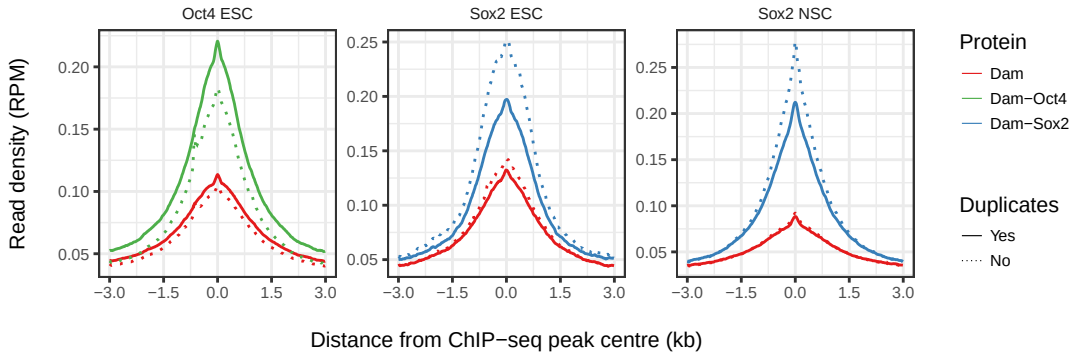

FIGURE 3.8. Graphs of DamID-seq read coverage with and without PCR duplicates removed at DNA binding sites.

These graphs show DamID-seq read coverage with and without PCR duplicates removed at Oct4 and Sox2 DNA binding sites in mouse embryonic stem cells (abbreviated ESC) and neural stem cells (abbreviated NSC). The DNA binding sites were identified by calling peaks from the Oct4 ESC (n = 37,925), Sox2 ESC (n = 15,690), and Sox2 NSC (n = 22,481) ChIP-seq data.

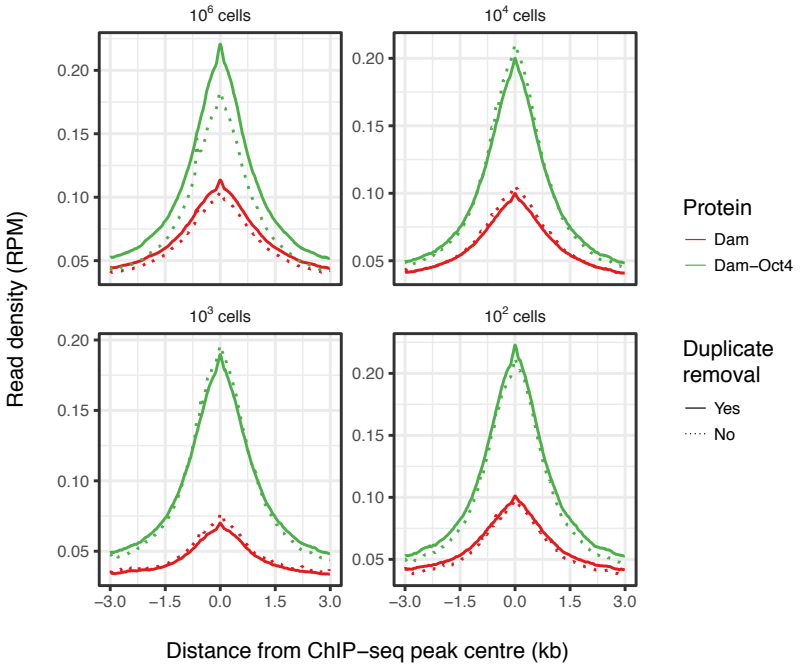

FIGURE 3.9. Graphs of low cell number Oct4 DamID-seq read coverage with and without PCR duplicates removed.

These graphs show DamID-seq read coverage with and without PCR duplicates removed at Oct4 DNA binding sites in 10^6, 10^4, 10^3, and 10^2 mouse embryonic stem cells (abbreviated ESC). The DNA binding sites were identified by calling peaks from the Oct4 ESC (n = 37,925), Sox2 ESC (n = 15,690), and Sox2 NSC (n = 22,481) ChIP-seq data.

3.4.3 Dam binding is not biased by nucleotide composition

The DNA adenine methyltransferase (abbreviated Dam) used in DamID-seq experiments is taken from *Escherichia coli* bacteria. Methylation of adenine is widespread across many bacteria, but is largely absent in eukaryotes (Aughey and Southall, 2016). When expressed in an entirely new system, it is possible that the binding specificity of Dam is altered or preferentially binds restriction sites with a specific local nucleotide composition. This invariably would alter which DNA binding sites were identified, because restriction sites with an unfavourable local nucleotide composition would be

less methylated or not methylated at all compared to other more favourable restriction sites. This type of bias can be seen in sequencing data generated by the Illumina Nextera DNA preparation kit, whereby the Tn5 transposase preferentially inserts itself into chromatin at a particular sequence (Ason and Reznikoff, 2004). To the best of our knowledge, only two previous studies have sought evidence for Dam exhibiting compositional DNA preference in binding. Horton and colleagues generated a crystal structure of Dam in complex with non-cognate DNA, lacking any GATC sequences. Their structures exhibited an apparent 5 bp DNA binding sequence, which was also found flanking GATC sites in some Dam-regulated promoters in the *Drosophila* genome (Horton et al., 2015). While this is unexpected, it is not clear whether Dam would bind to such an element *in vivo* away from the concentrated solution used in this study to achieve a crystal structure of high resolution. Additionally, Bergerat and colleagues observed that at very low temperature (0°C) and in the presence of S-Adenosyl methionine (Ado-Met) Dam can bind to non-specific DNA with low affinity (Bergerat and Guschlbauer, 1990). Again, this altered specificity for binding is only observed *in vitro* and it is unknown whether Dam would exhibit such behaviour at a physiological temperature and condition. Importantly, neither of these experiments have been carried out on a genome-wide scale where the ability to measure methylation at every individual GATC site is possible, and can be accurately quantified.

To determine if Dam preferentially binds restriction sites with a specific local nucleotide composition, the DNA compositional bias around methylated and unmethylated restriction sites in multiple DamID-seq data sets was plotted (see Figure 3.10). In all libraries a uniform sequence composition around both methylated and un-methylation restriction sites was observed. The difference between the guanine-cytosine (GC) and

adenine-thymine (AT) compositions was caused by the GC (42%) and AT (58%) content of the mouse genome (Ruvinsky and Marshall Graves, 2005). To visualise more subtle patterns in the data, the plots were scaled to remove the influence of the GATC site on the nucleotide frequency (see Figure 3.11). When scaled, the plots showed a small decrease in A/T nucleotides flanking the GATC motif in methylated versus unmethylated restriction fragments (methylated = 0.35 and unmethylated = 0.38). However, it is not clear whether such a small decrease (3%) is biologically significant reflecting an innate bias in Dam binding, or whether some technical factor such as PCR amplification bias, differential fragmentation efficiency of DNA templates, or simply changing the cut-off threshold for defining methylated versus unmethylated restriction fragments would change this observation. Either way, such a small decrease is unlikely to alter the genome-wide methylation levels to any extent that downstream peak calling would be affected. To ensure there was no sequence-specificity at different levels of methylation, the DNA compositional bias around restriction sites separated into quartiles by methylation level was also plotted (see Figure 3.12). Again, all libraries displayed a uniform sequence composition regardless of the level of methylation. These results indicated that Dam does not preferentially bind restriction sites in a sequence-specific manner and that no correction of the data with respect to local nucleotide composition is required.

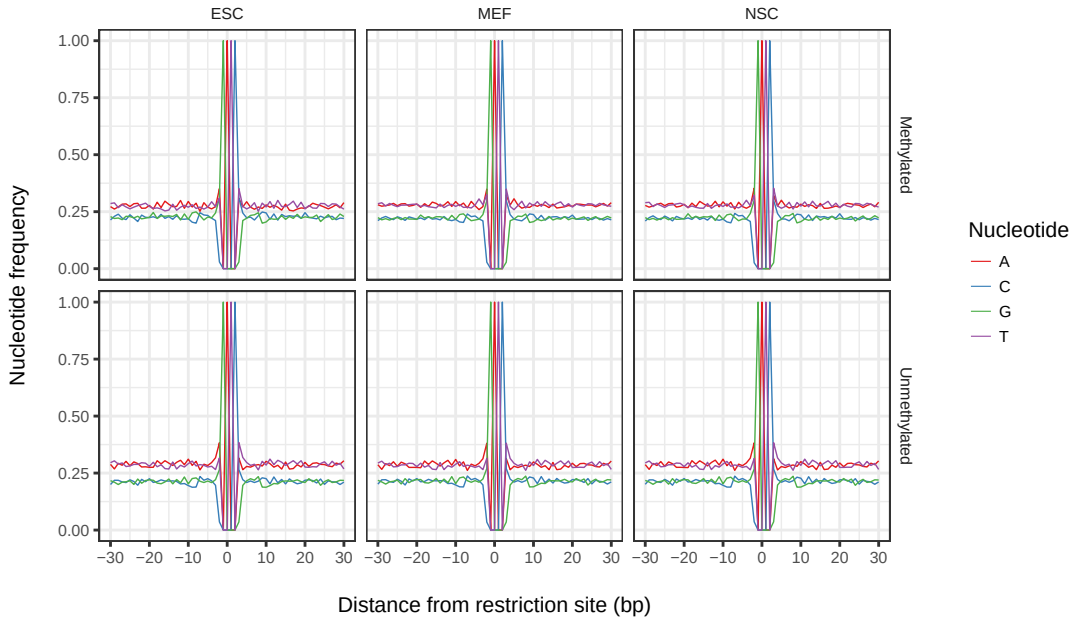

FIGURE 3.10. Graphs of nucleotide frequency around methylated and unmethylated restriction sites.

Graphs of nucleotide frequency around methylated and un-methylated restriction sites from DamID-seq data for embryonic stem cells (ESC), embryonic fibroblast cells (MEF), and neural stem cells (NSC). Restriction sites neighbouring a restriction fragment with greater than 10 reads were classed as methylated, and vice versa. Distance zero was defined as the position of adenine in the restriction site GATC sequence.

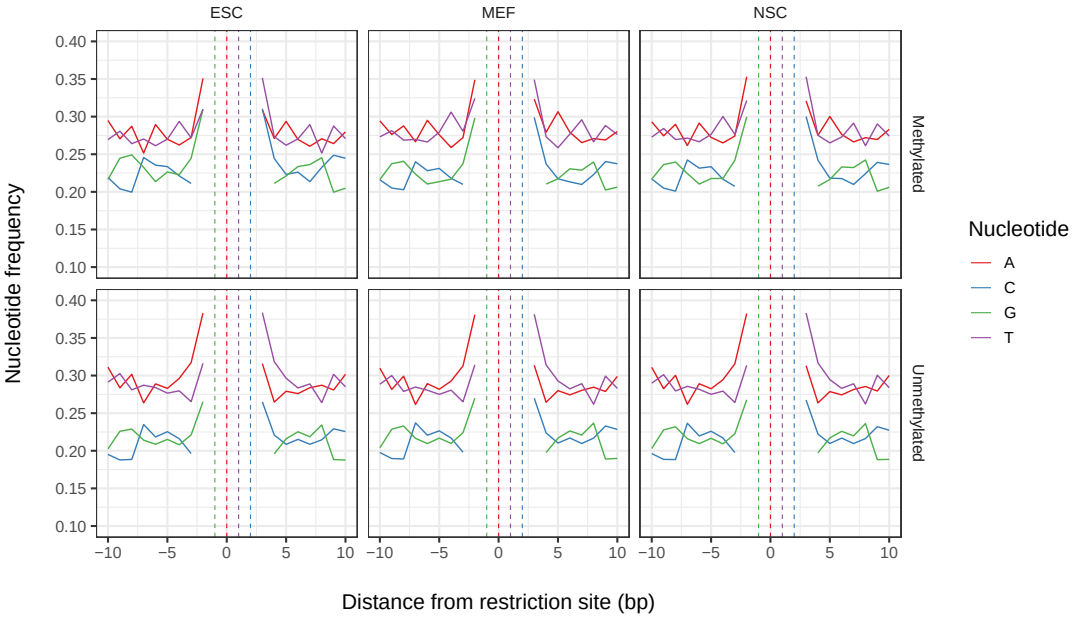

FIGURE 3.11. Graphs of nucleotide frequency around methylated and unmethylated restriction sites.

Graphs of nucleotide frequency around methylated and un-methylated restriction sites from DamID-seq data for embryonic stem cells (ESC), embryonic fibroblast cells (MEF), and neural stem cells (NSC). Restriction sites neighbouring a restriction fragment with greater than 10 reads were classed as methylated, and vice versa. Distance zero was defined as the position of adenine in the restriction site GATC sequence. The graphs caled to remove the influence of the GATC site on the nucleotide frequency. The vertical dashed lines represent the position of each nucleotide in the GATC site.

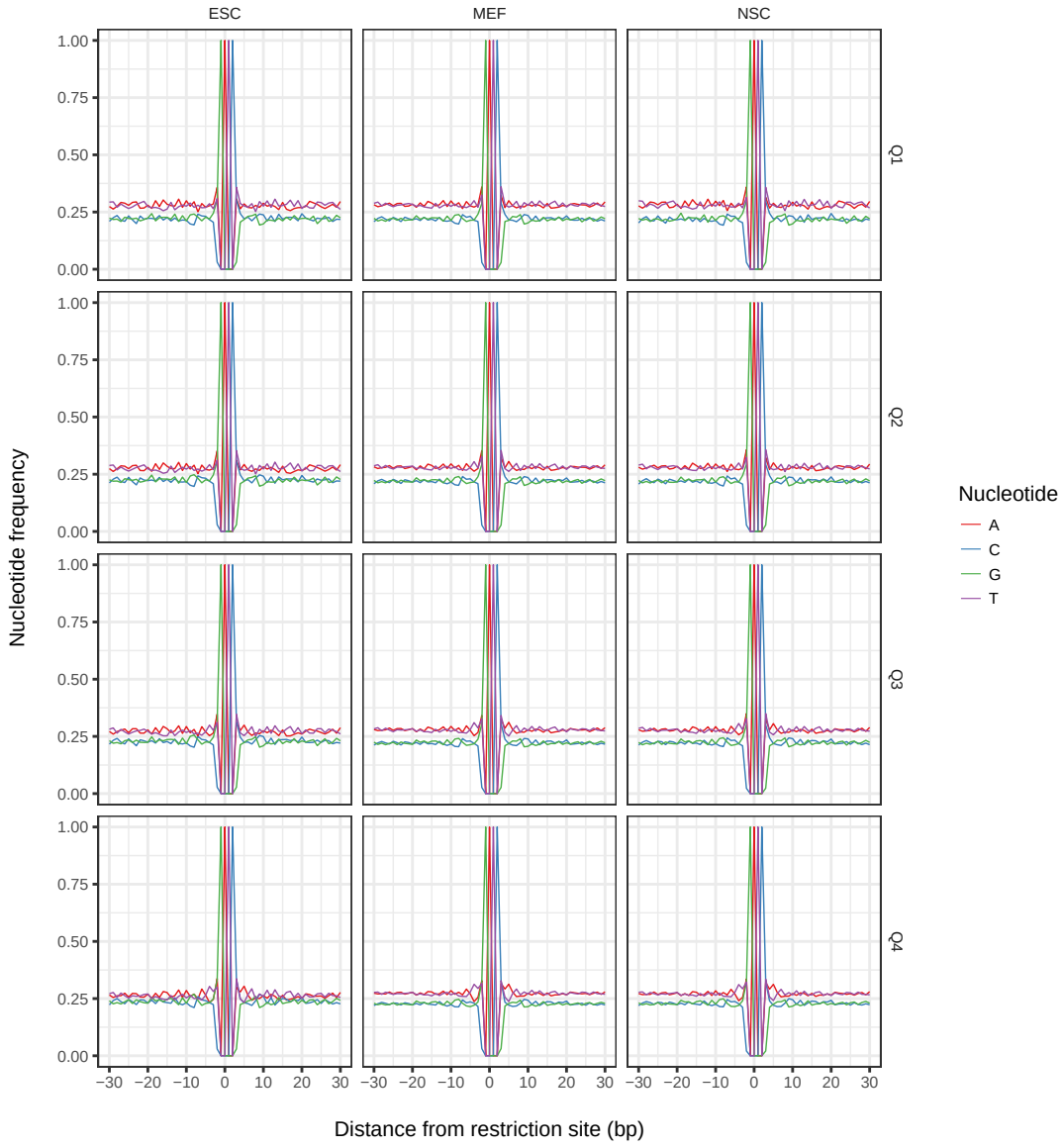

FIGURE 3.12. Nucleotide frequency around restriction sites divided into quartiles by methylation.

Graphs of nucleotide frequency around methylated restriction sites from DamID-seq data for embryonic stem cells (ESC), embryonic fibroblast cells (MEF), and neural stem cells (NSC). Restriction sites neighbouring a restriction fragment with greater than 10 reads were classed as methylated. Restriction sites were then divided into quartiles my methylation. Distance zero was defined as the position of adenine in the restriction site GATC sequence.

3.4.4 DpnII digestion is required for enrichment of factor-bound chromatin

In the original DamID-seq experimental protocol, adapted-ligated DNA was digested with restriction enzyme DpnII to avoid amplifying fragments containing unmethylated restriction sites (Greil, Moorman, and Van Steensel, 2006). However, recent publications aiming to improve upon this protocol have removed the digestion step entirely, citing reasons of redundancy and simplicity (Hass et al., 2015; Kind et al., 2015; Pindyurin et al., 2016). They argue that in a typical library there should be very few adapter-ligated fragments which contain unmethylated restriction sites, and that their limited presence should not impact the sequencing data generated.

To determine the effect of DpnII digestion, read coverage of restriction fragments in Oct4 DamID-seq data prepared with and without DpnII digestion (abbreviated +DpnII/-DpnII) was measured. Genome browser tracks showed that read coverage at multiple of Oct4 binding sites was drastically reduced without DpnII digestion (see Figure 3.13). More precisely, read coverage at all Oct4 binding sites was approximately four-fold lower in the -DpnII libraries, and the ratio of Dam-Oct4 over Dam was almost two-fold lower (see Figure 3.14). The reduction in coverage at DNA binding sites makes it more challenging to identify statistically significant binding, because the level of methylation is closer to the level of random noise in the sequencing data, such as non-specific background methylation. In a typical ChIP-seq experiment, the input DNA library is expected to display a uniform distribution of reads along the genome. There should be no preferential enrichment at particular genomic features. By comparison, the antibody-treated library is expected to display a biased distribution of reads along the genome. There should be preferential enrichment at the binding sites of the target protein (Diaz, Nellore, and Song, 2012). A similar pattern should be observed in the DamID-seq data, with the Dam-fusion library localised to binding sites, and the Dam library distributed more widely across the genome. If the reads in

both the treatment and control libraries are spread equally across the genome, it becomes harder to differentiate signal from background. Without DpnII digestion, both the Dam and Dam-Oct4 reads were more widely distributed along the genome, and the difference between the Dam and Dam-Oct4 distributions was narrower (see Figure 3.15). These observations suggested that DpnII digestion is required for adequate read coverage at DNA binding sites, and that without DpnII digestion it is harder to differentiate between the foreground and background methylation.

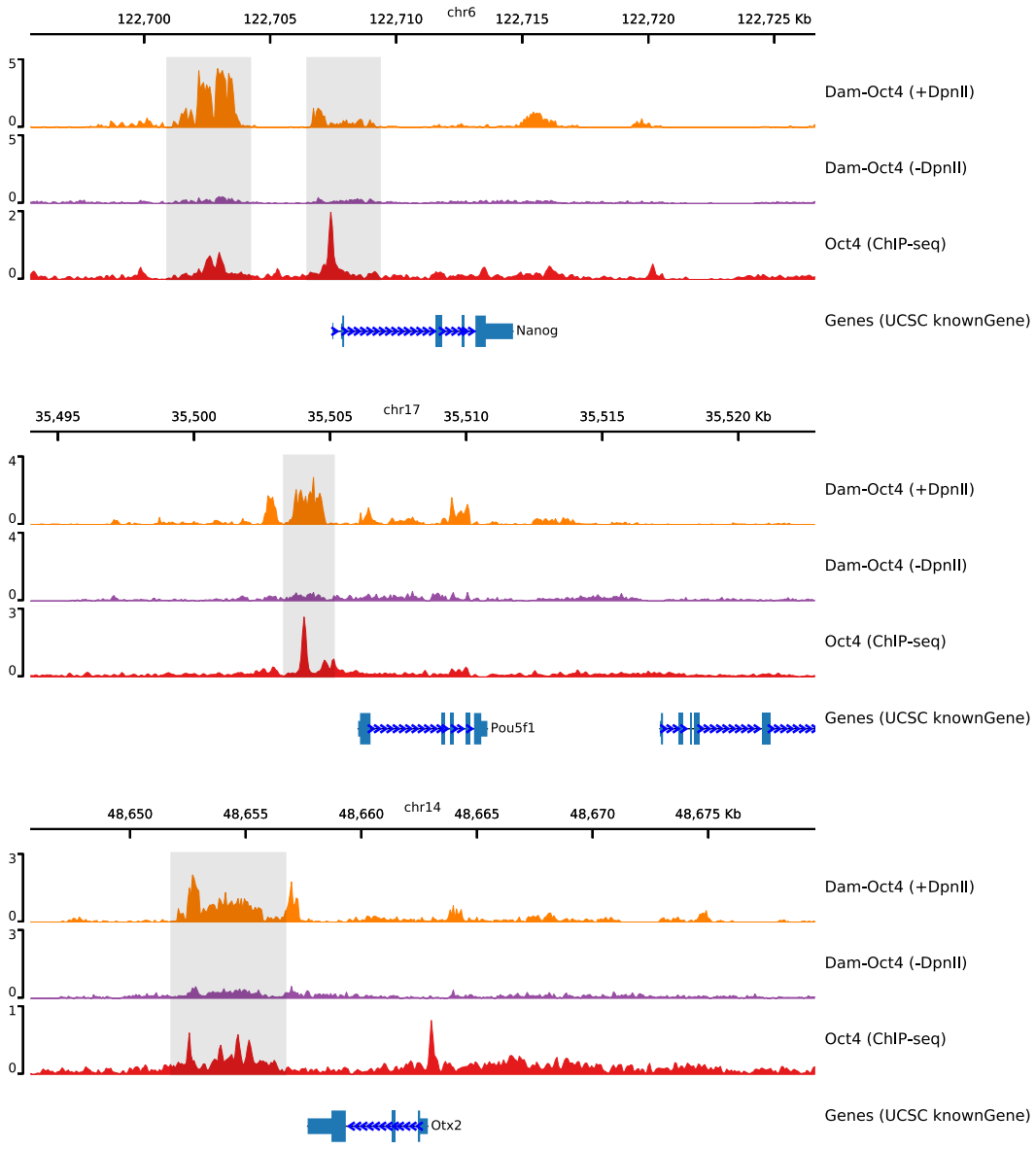

FIGURE 3.13. Tracks of +DpnII/-DpnII Oct4 DamID-seq read coverage at Oct4 binding sites.

Tracks of Oct4 DamID-seq data generated with and without DpnII digestion (abbreviated +DpnII/-DpnII). In the absence of DpnII digestion, enrichment at Oct4 DNA-binding sites is drastically reduced.

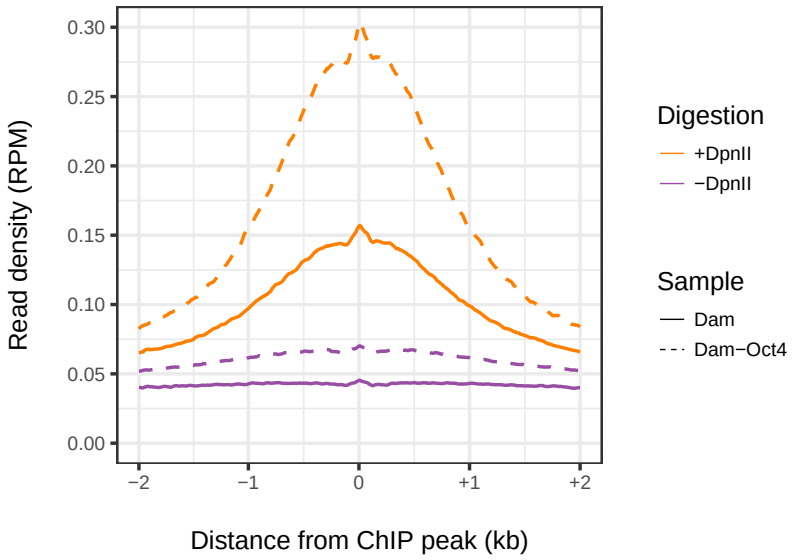

FIGURE 3.14. Graph of +DpnII/-DpnII Oct4 DamID-seq read coverage at Oct4 binding sites.

Graph of Oct4 DamID-seq read coverage at all Oct4 binding sites in mouse embryonic stem cells (abbreviated ESC) with and without DpnII digestion. The Oct4 binding sites were identified by calling peaks from the Oct4 ESC (n = 37,925) ChIP-seq data. Distance zero was defined as the position of the summit in the ChIP-seq peaks.

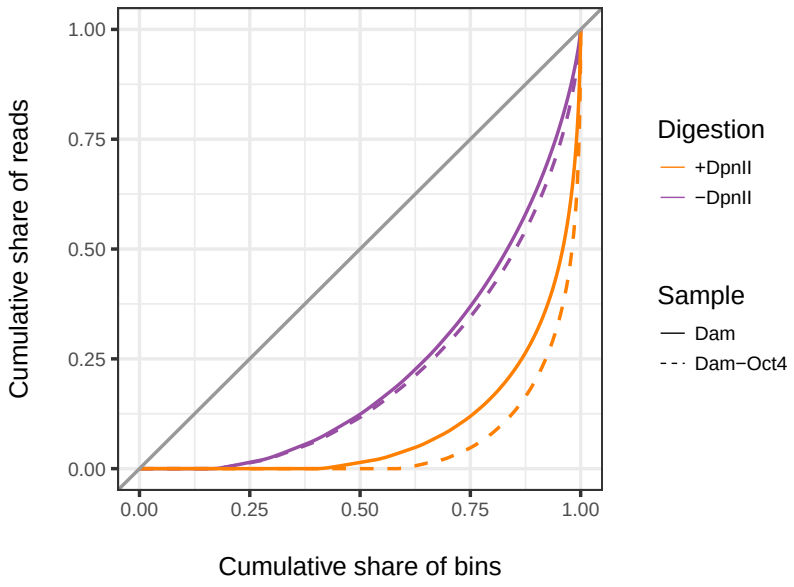

FIGURE 3.15. Graph of the distribution of reads from +DpnII/-DpnII Oct4 DamID-seq data

Graph showing the proportion of DamID-seq reads aligned in 1 kb bins along the entire genome. The curve shows the bottom x fraction of bins have y fraction of the total sequencing. Libraries prepared with DpnII digestion have more localised and stronger enrichment than those without DpnII digestion.

3.4.5 Polymerase efficiency impacts restriction fragment amplification

In the original DamID-seq experimental procedure, adapter-ligated DNA was amplified using the polymerase chain reaction (abbreviated PCR) with an Advantage II polymerase from Takara Clontech (Van Steensel and Henikoff, 2000). Since then, manufacturers have engineered more processive and sensitive polymerases yet the majority of published DamID-seq experiments have not upgraded. The PCR amplification process is worth exploring because it has been shown that different polymerases generate libraries with varying complexity and coverage, which impacts the accuracy and sensitivity of the DNA-seq experiment (Brandariz-Fontes et al., 2015). To determine the affect of PCR amplification on DamID-seq experiments, read coverage of Oct4 ESC DamID-seq libraries prepared from the same DNA sample but amplified with

two different polymerases (the original Advantage II polymerase from Takara Clontech and the more recent KAPA HiFi from Kapa Biosystems) were compared. Genome browser tracks showed that sequencing libraries generated with the Kapa polymerase contained enriched regions which were entirely absent in those generated with the Clontech polymerase (see Figure 3.16). More specifically, the number of sequenced bases along the genome (i.e. genome coverage) was consistently higher in the Kapa than Clontech libraries over a range of sequencing depths (see Figure 3.17). Additionally, libraries prepared with the Kapa polymerase were revealed to have higher read and restriction fragment complexity than those prepared with the Clontech polymerase (see Figures 3.18 and 3.19). These results indicated that libraries prepared with the Kapa polymerase successfully amplified a higher number of restriction fragments, which potentially would also contain DNA binding sites for the protein being assayed. Generally, different polymerases greatly affected the PCR amplification process, and that to achieve higher quality DamID-seq libraries, future experiments should be performed with a more accurate polymerase (e.g. KAPA HiFi from Kapa Biosystems).

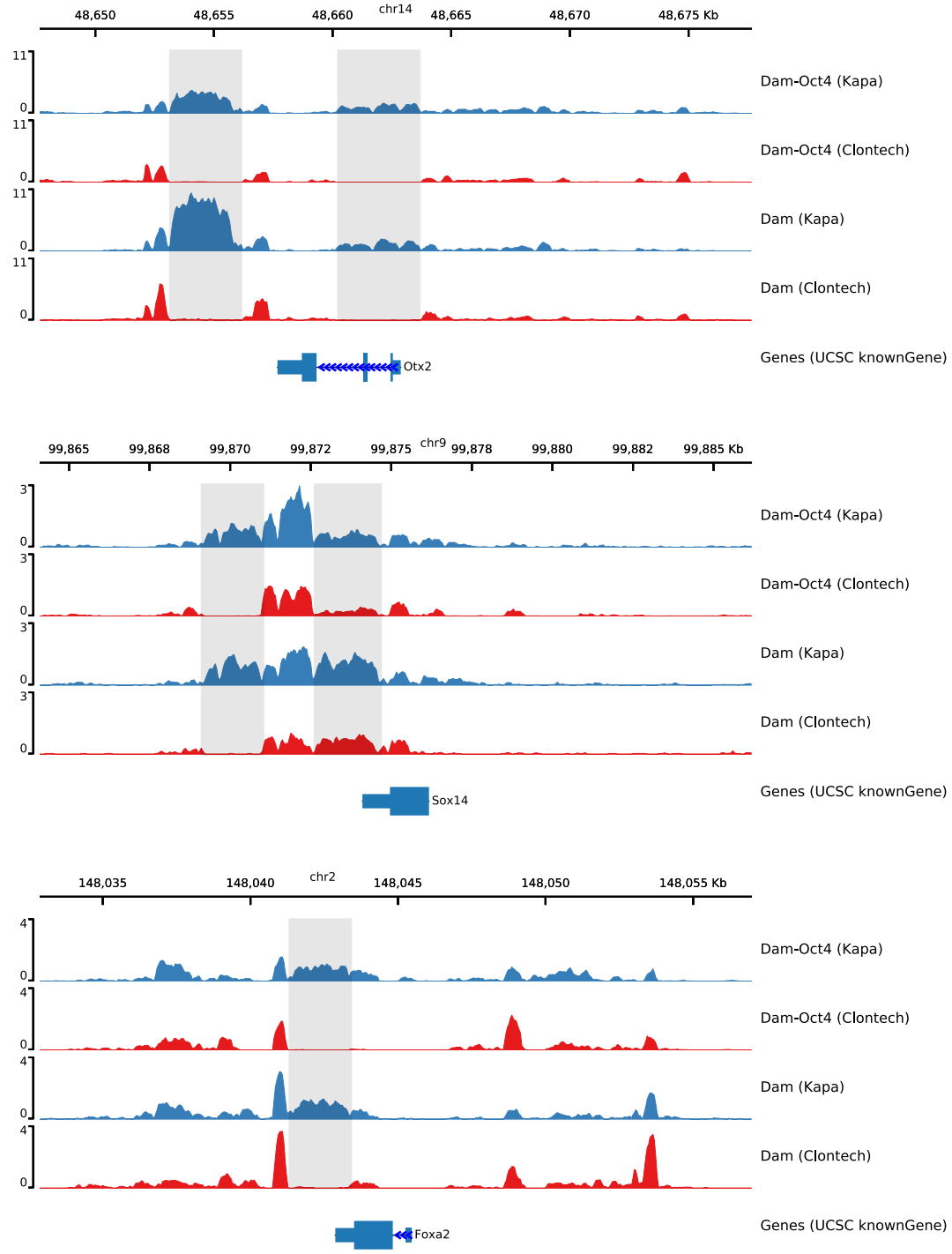

FIGURE 3.16. Tracks of Oct4 DamID-seq data amplified using Clontech and Kapa polymerases.

Tracks of DamID-seq read coverage prepared from the same DNA library but amplified with either the Advantage II polymerase from Takara Clontech or the KAPA HiFi polymerase from Kapa Biosystems. In all tracks, the Kapa libraries display enriched regions which are missing from the Clontech libraries.

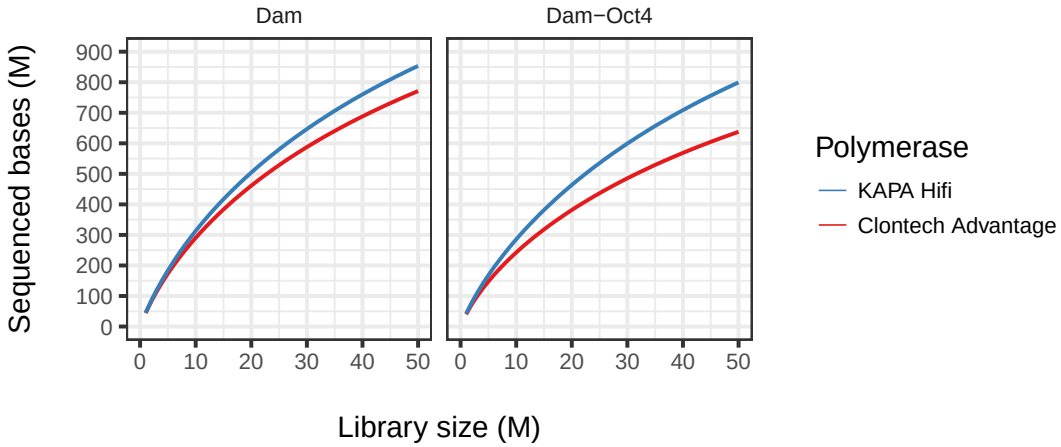

FIGURE 3.17. Coverage of DamID-seq data generated by Clontech and Kapa polymerases.

Graphs displaying the number of distinct bases sequencing in the mouse genome over a range of library sizes for Oct4 ESC DamID-seq libraries prepared from the same DNA library but amplified with either the Advantage II polymerase from Takara Clontech or the KAPA HiFi polymerase from Kapa Biosystems. In both the Dam and Dam-Oct4 sequencing data, the Kapa libraries display a higher coverage than Clontech libraries.

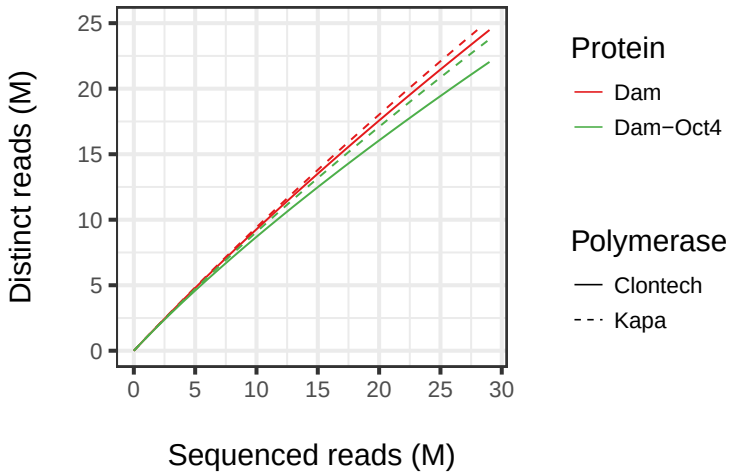

FIGURE 3.18. Read complexity of DamID-seq data generated by Clontech and Kapa polymerases.

Graph of read complexity for Oct4 ESC DamID-seq libraries prepared from the same DNA sample but amplified with either the Advantage II polymerase from Takara Clontech or the KAPA HiFi polymerase from Kapa Biosystems. In both the Dam and Dam-Oct4 sequencing data, the Kapa libraries are more complex compared to the Clontech libraries.

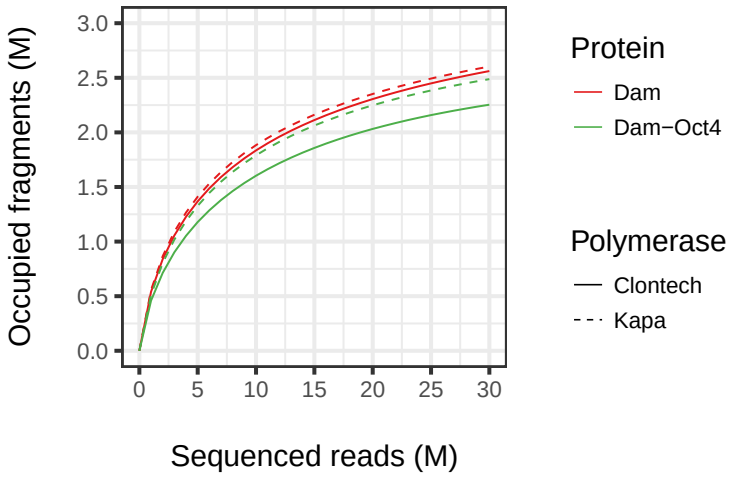

FIGURE 3.19. Restriction fragment complexity of DamID-seq data generated by Clontech and Kapa polymerases.

Graph of restriction fragment complexity for Oct4 ESC DamID-seq libraries prepared from the same DNA sample but amplified with either the Advantage II polymerase from Takara Clontech or the KAPA HiFi polymerase from Kapa Biosystems. In both the Dam and Dam-Oct4 sequencing data, the Kapa libraries are more complex compared to the Clontech libraries.

3.4.6 Restriction fragment size affects the level of methylation

Read coverage in next-generation sequencing experiments is meant to provide an accurate and quantitative measurement of the biology under investigation. For example, coverage can be used to quantify enrichment of factor-bound chromatin (ChIP-seq), expression of mRNA transcripts (RNA-seq), or interaction of associated chromatin (Hi-C). For DamID-seq experiments, read coverage is used to quantify the methylation of factor-bound chromatin. Previous work has shown that read coverage can be biased by technical factors such as transcript length in RNA-seq experiments (Oshlack and Wakefield, 2009) or restriction fragment length in Hi-C experiments (Yaffe and Tanay, 2011). In RNA-seq data, longer transcripts tend to have greater coverage than shorter ones, even if they have the same level of expression. In Hi-C data however, longer restriction fragments tend to have less coverage than shorter ones,

even if they have the same physical contact. This variation in coverage between features of different lengths presents two problems: it confounds comparisons between features because coverage is influenced by factors unrelated to abundance, and biological replicates can be biased differently which reduces reproducibility. To identify DNA binding sites from DamID-seq data, each restriction fragment is tested for differential methylation between the Dam and Dam-fusion libraries. In theory, differential methylation should not be affected by fragment length because proteins bind along the genome without regard for the location of restriction sites, and therefore the level of methylation should not be influenced.

To determine if there is a relationship between restriction fragment length and differential methylation, Oct4 and Sox2 DamID-seq read counts were binned according to restriction fragment length and the percentage of differentially methylated fragments was plotted (see Figure 3.20). The data showed that as fragment length increased the ability to detect differential methylation decreased. This same pattern was also observed in the low cell number Oct4 DamID-seq data (see Figure 3.21). To investigate the exact relationship between read coverage and restriction fragment length, a quantile regression model (i.e. coverage predicted from length) was fitted to data from multiple DamID-seq experiments (see Figure 3.22). All of the models displayed strong sample-specific non-linear relationships between coverage and length, which were different between fusion proteins and cell types. Similar sample-specific non-linear relationships were also observed for the low cell number DamID-seq data (see Figure 3.23), and despite measuring the same underlying biology (i.e. Oct4 binding in ESCs) there was no common relationship between libraries from different cell numbers. There was however a notable convergence in coverage within all of the libraries for restriction fragments between 10^3 and 10^4 base pairs in length. However, coverage

of restriction fragments smaller or greater than this range were either continuously increasing or decreasing. This suggest that regardless of their size, restriction fragments in this 10^3 to 10^4 range were methylated equally and preferentially over those outside this range. This is concerning because comparisons between DNA binding sites will be influenced by the size of the restriction fragment where they are located. In addition, the data also showed that the non-linear relationships were distinct between the Dam and Dam-fusion proteins. This is also concerning because in order to correctly test for differential methylation we assume that the same restriction fragment can be methylated with equal opportunity across the Dam and Dam-fusion libraries. If this assumption does not hold, restriction fragments could be incorrectly identified as being differentially methylated because of Dam and Dam-fusion length biases rather than differences in DNA binding between the Dam and Dam-fusion proteins. Together these results highlight a significant need for analysis methods which test differential restriction fragment methylation whilst accounting for restriction fragment length biases.

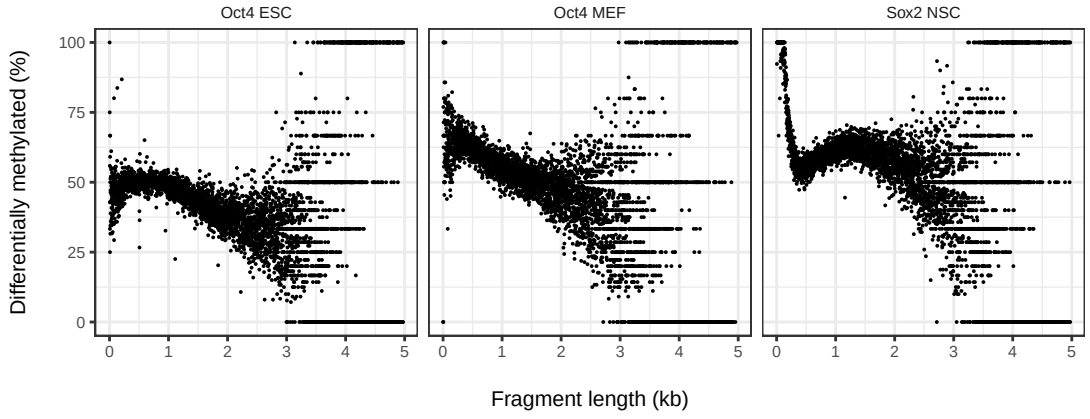

FIGURE 3.20. Graphs of differential methylation by fragment length in Oct4 and Sox2 DamID-seq data.

Graphs showing the percentage of differentially methylated fragments binned according to restriction fragment length in Oct4 and Sox2 DamID-seq data. Differential methylation was tested using limma (FDR < 0.1).

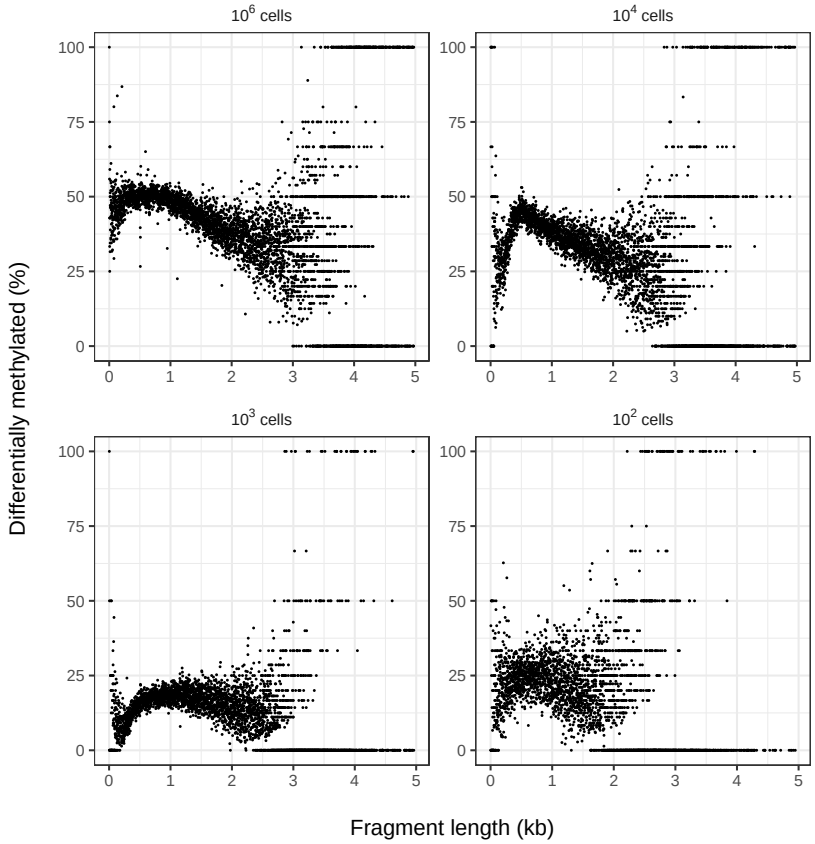

FIGURE 3.21. Graphs of differential methylation by fragment length in low cell number Oct4 DamID-seq data.

Graphs showing the percentage of differentially methylated fragments binned according to restriction fragment length in low cell number Oct4 DamID-seq data. Differential methylation was tested using limma (FDR < 0.1).

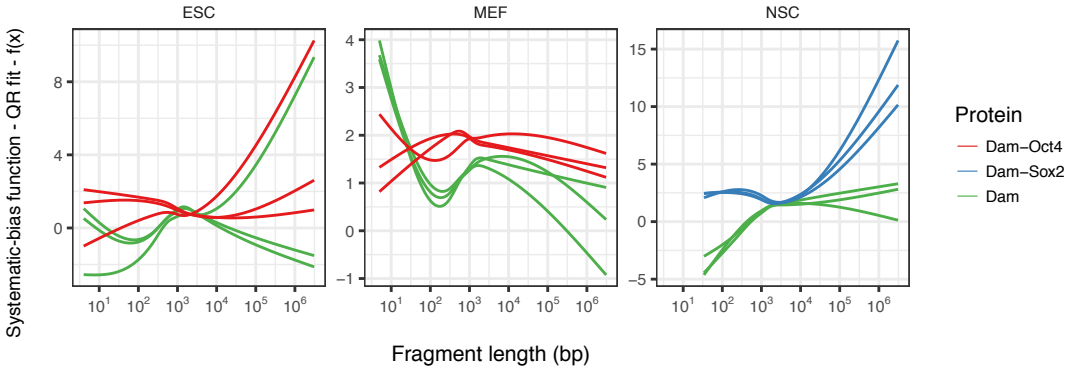

FIGURE 3.22. Graphs of the fragment length effect on methylation in Oct4 and Sox2 DamID-seq data.

Graphs of the estimated fragment length effect on methylation in Oct4 and Sox2 DamID-seq data from mouse embryonic stem cells (abbreviated ESC), embryonic fibroblast cells (abbreviated MEF), and neural stem cells (abbreviated NSC). The X-axis measures the restriction fragment length in kilobases and the Y-axis measures the fit from the quantile regression model. Specifically, for each sample a systematic bias function using natural cubic splines can be estimated. For each value on the x-axis, the y-axis plots the fitted value generated from each of the sample-specific bias functions. The sample-specific affect of the covariate on read counts can be visualised by plotting the estimates from the bias functions. Each line represents the trend from a biological sample.

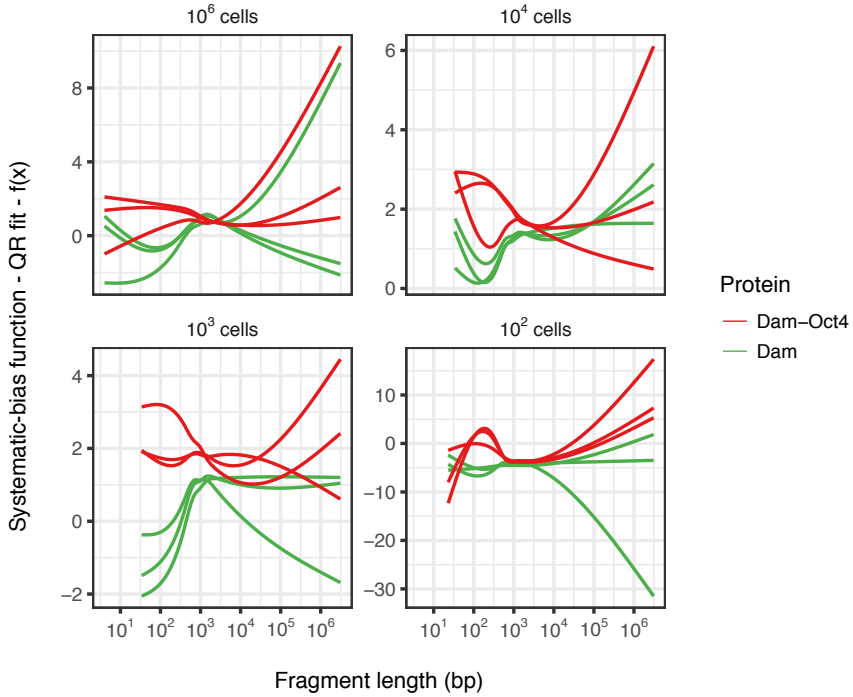

FIGURE 3.23. Graphs of the fragment length effect on methylation in low cell number Oct4 DamID-seq data.

Graphs of the estimated fragment length effect on methylation in low cell number DamID-seq data from mouse embryonic stem cells (abbreviated ESC). The X-axis measures the restriction fragment length in kilobases and the Y-axis measures the fit from the quantile regression model. Specifically, for each sample a systematic bias function using natural cubic splines can be estimated. For each value on the x-axis, the y-axis plots the fitted value generated from each of the sample-specific bias functions. The sample-specific affect of the covariate on read counts can be visualised by plotting the estimates from the bias functions.

3.4.7 Restriction fragment GC content affects methylation levels

Similar to feature length, read coverage can also be affected by sequence composition such as guanine-cytosine (GC) content (Hansen, Irizarry, and Wu, 2012). In RNA-seq data, mRNA transcripts expressed at the same level can be measured differently because of their GC content and this can substantially bias differential expression analysis (Risso et al., 2011). In addition, Hi-C data can also be strongly affected by GC content near the ligated restriction fragment ends (Yaffe and Tanay, 2011). For DamID-seq

data, differential methylation should not be affected by restriction fragment GC content because proteins bind along the genome without regard for the GC content of the restriction fragment, and therefore the level of methylation should not be influenced. To determine if there is a relationship between restriction fragment GC content and differential methylation, DamID-seq read counts were binned according to restriction fragment GC content and the percentage of differentially methylated fragments was plotted (see Figure 3.24). The data showed that as GC content increased the ability to detect differential methylation increased. This same pattern was also observed in the low cell number DamID-seq data (see Figure 3.25). To investigate the exact relationship between read coverage and restriction fragment GC content, a regression model (i.e. coverage predicted from GC content) was fitted to data from multiple DamID-seq experiments (see Figure 3.26). All of the models displayed strong sample-specific non-linear relationships between coverage and GC content, which were also different between fusion proteins and cell types. Similar sample-specific non-linear relationships were also observed for the low cell number DamID-seq data (see Figure 3.23), and despite measuring the same underlying biology (i.e. Oct4 binding in ESCs) there was no common relationship between libraries from different cell numbers. Similar to the restriction fragment length bias identified previously, the affect of GC content must also be accounted for in the analysis of DamID-seq data.

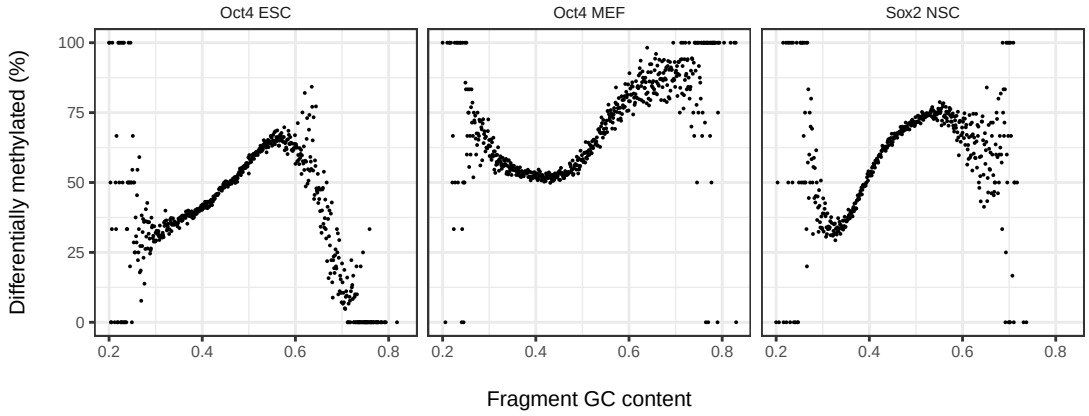

FIGURE 3.24. Graphs of differential methylation by fragment GC content in Oct4 and Sox2 DamID-seq data.

Graphs showing the percentage of differentially methylated fragments binned according to restriction fragment GC content in Oct4 and Sox2 DamID-seq data. Differential methylation was tested using limma (FDR < 0.1).

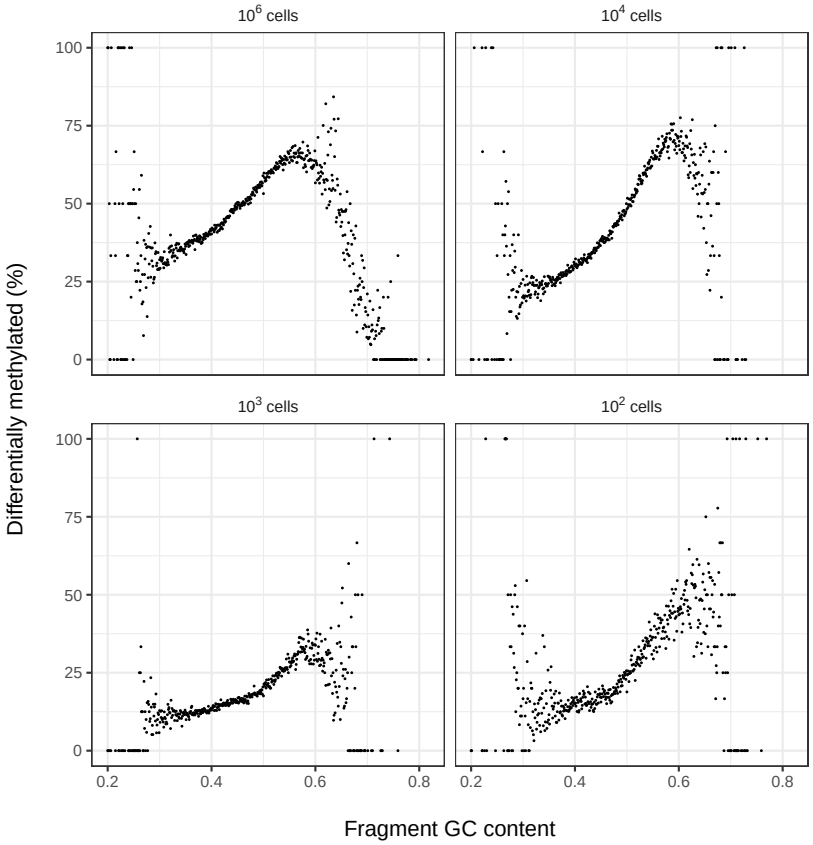

FIGURE 3.25. Graphs of differential methylation by fragment GC content in low cell number Oct4 DamID-seq data.

Graphs showing the percentage of differentially methylated fragments binned according to restriction fragment GC content in low cell number Oct4 DamID-seq data. Differential methylation was tested using limma (FDR < 0.1).

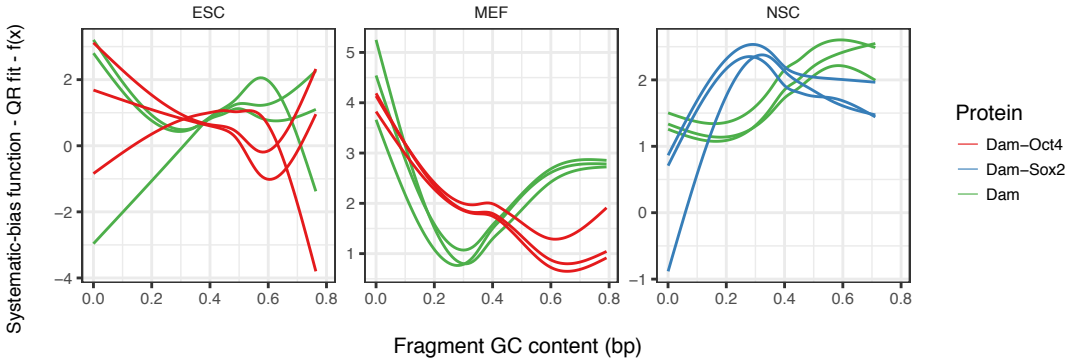

FIGURE 3.26. Graphs of the fragment GC content effect on methylation in Oct4 and Sox2 DamID-seq data.

Graphs of the estimated fragment GC content effect on methylation in Oct4 and Sox2 DamID-seq data from mouse embryonic stem cells (abbreviated ESC), embryonic fibroblast cells (abbreviated MEF), and neural stem cells (abbreviated NSC). The X-axis measures the restriction fragment GC content and the Y-axis measures the fit from the quantile regression model. Specifically, for each sample a systematic bias function using natural cubic splines can be estimated. For each value on the x-axis, the y-axis plots the fitted value generated from each of the sample-specific bias functions. The sample-specific affect of the covariate on read counts can be visualised by plotting the estimates from the bias functions.

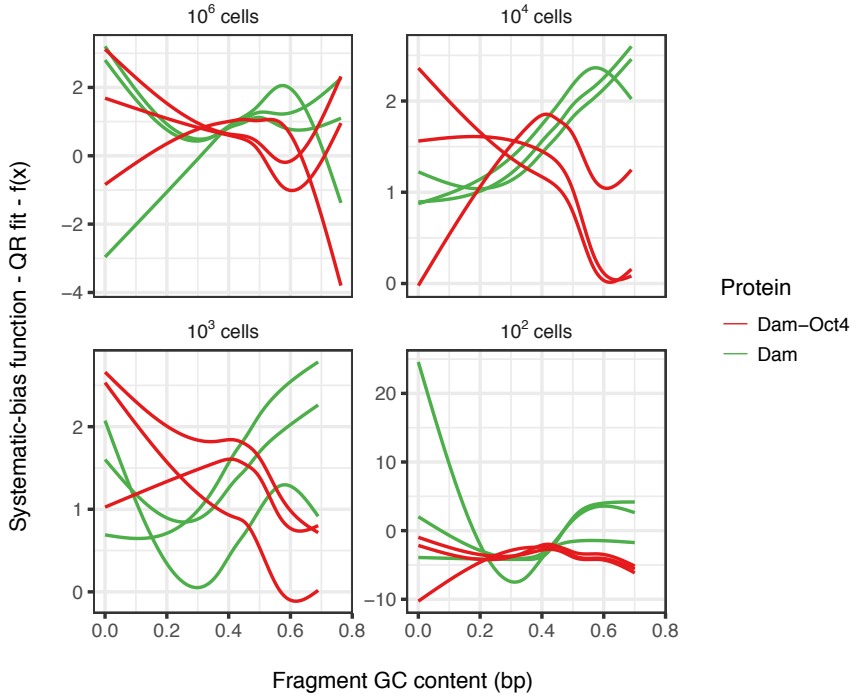

FIGURE 3.27. Graphs of the fragment GC content effect on methylation in low cell number Oct4 DamID-seq data.

Graphs of the estimated fragment GC content effect on methylation in low cell number Oct4 DamID-seq data from mouse embryonic stem cells (abbreviated ESC). The X-axis measures the restriction fragment GC content and the Y-axis measures the fit from the quantile regression model. Specifically, for each sample a systematic bias function using natural cubic splines can be estimated. For each value on the x-axis, the y-axis plots the fitted value generated from each of the sample-specific bias functions. The sample-specific affect of the covariate on read counts can be visualised by plotting the estimates from the bias functions.

3.4.8 Regional GC content does not affect methylation levels

Finally, it was important to check whether a relationship between restriction fragment length and GC content could also be observed. The restriction fragment GC content may reflect the regional GC content in the genome, which then influences the frequency of GATC sites and therefore the length and methylation level of restriction

fragments. To determine if read coverage is affected by the relationship between restriction fragment length and GC content, DamID-seq read counts were first length-normalised then binned according to restriction fragment GC content and the percentage of differentially methylated fragments was plotted (see Figures 3.28 and 3.29). If there is a relationship between restriction fragment length and GC content, a reduction in GC bias would be observed after normalising read counts for length. The plots however showed that a similar GC content bias was still observable after normalising for fragment length in both the multiple cell type and low cell number DamID-seq data (compare Figure 3.28 to 3.24 and Figure 3.29 to 3.25). To investigate the exact relationship between length-normalised read coverage and GC content, a regression model (i.e. length-normalised coverage predicted from GC content) was fitted to the DamID-seq data (see Figures 3.30 and 3.31). All of the models again displayed strong sample-specific non-linear relationships between length-normalised coverage and GC content (compare Figure 3.30 to 3.26 and Figure 3.31 to 3.27). It is important to note that although a strong bias was observed in all the datasets, the regression models appeared to be slightly different to the ones previously calculated. This was expected given that when the original GC content bias models were calculated, the effect of restriction fragment length was held constant in the model in order to visualise just the effect of GC content on read coverage. Ultimately, none of the length-normalised bias plots showed a reduction in GC content bias suggesting that there is no strong relationship between regional GC content and methylation level.

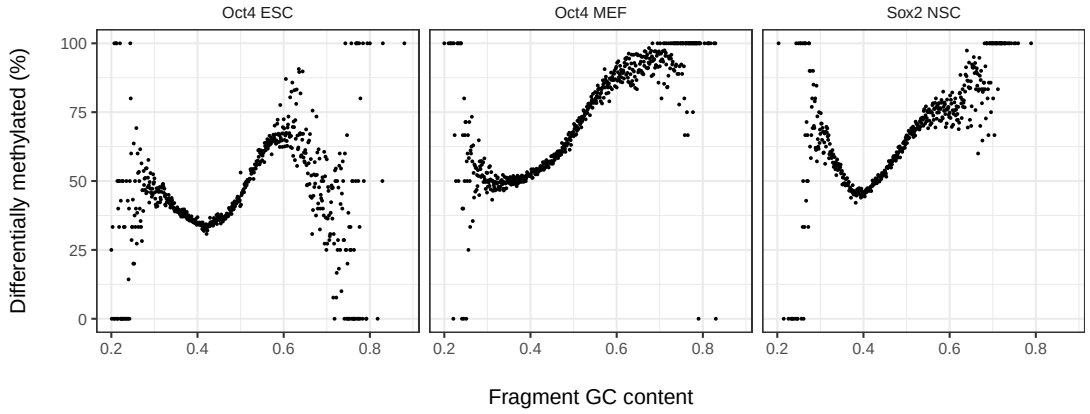

FIGURE 3.28. Graphs of differential methylation by fragment GC content in length-normalised Oct4 and Sox2 DamID-seq data.

Graphs showing the percentage of differentially methylated fragments binned according to restriction fragment GC content in length-normalised Oct4 and Sox2 DamID-seq data. Differential methylation was tested using limma (FDR < 0.1). Read counts were normalised for restriction fragment length using conditional quantile normalization from the cqn package (Hansen, Irizarry, and Wu, 2012).

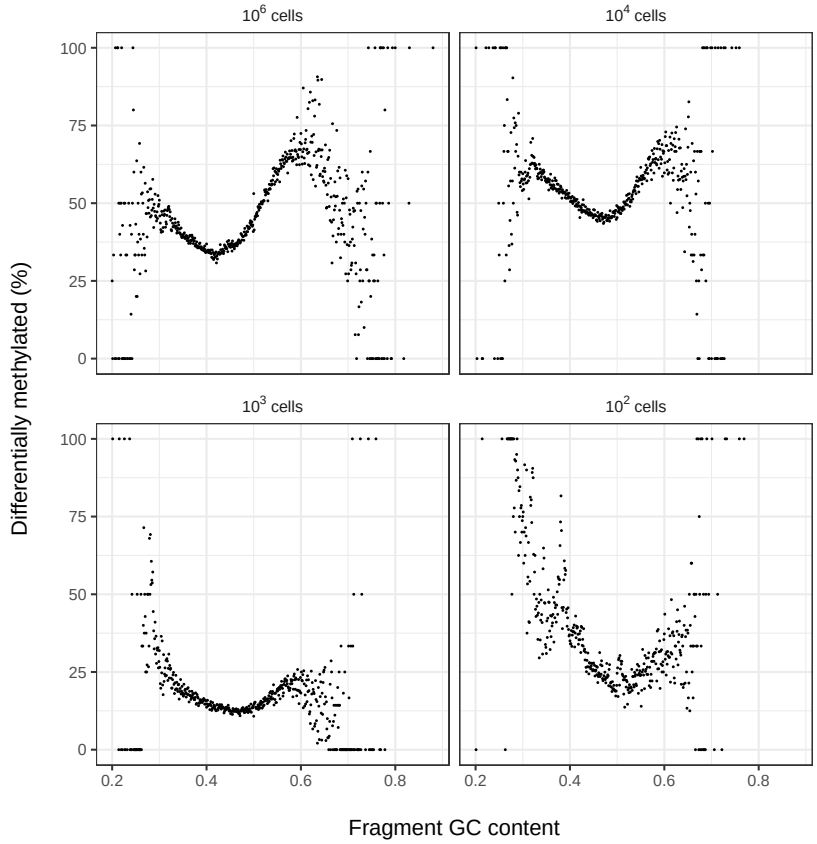

FIGURE 3.29. Graphs of differential methylation by fragment GC content in low cell number length-normalised Oct4 DamID-seq data.

Graphs showing the percentage of differentially methylated fragments binned according to restriction fragment GC content in low cell number length-normalised Oct4 DamID-seq data. Differential methylation was tested using limma (FDR < 0.1). Read counts were normalised for restriction fragment length using conditional quantile normalization from the cqn package (Hansen, Irizarry, and Wu, 2012).

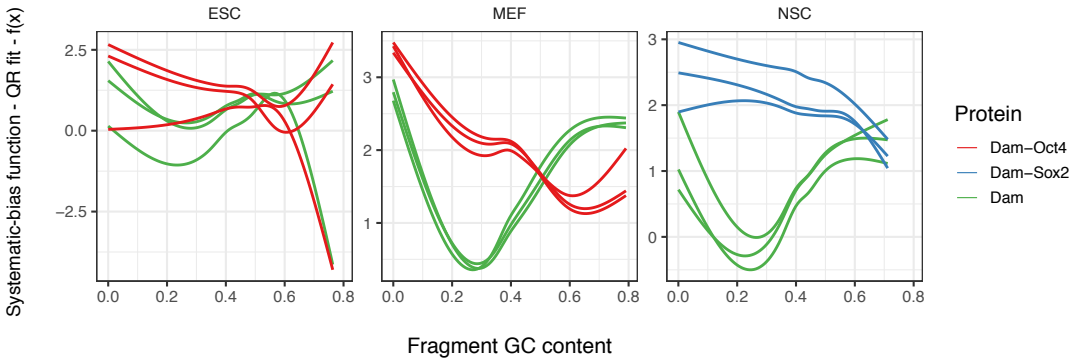

FIGURE 3.30. Graphs of the fragment GC content effect on methylation in length-normalised Oct4 and Sox2 DamID-seq data.

Graphs of the estimated fragment GC content effect on length-normalised methylation in Oct4 and Sox2 DamID-seq data from mouse embryonic stem cells (abbreviated ESC), embryonic fibroblast cells (abbreviated MEF), and neural stem cells (abbreviated NSC). Read counts were normalised for restriction fragment length using conditional quantile normalization from the cqn package (Hansen, Irizarry, and Wu, 2012). The X-axis measures the restriction fragment GC content and the Y-axis measures the fit from the quantile regression model. Specifically, for each sample a systematic bias function using natural cubic splines can be estimated. For each value on the x-axis, the y-axis plots the fitted value generated from each of the sample-specific bias functions. The sample-specific affect of the covariate on read counts can be visualised by plotting the estimates from the bias functions.

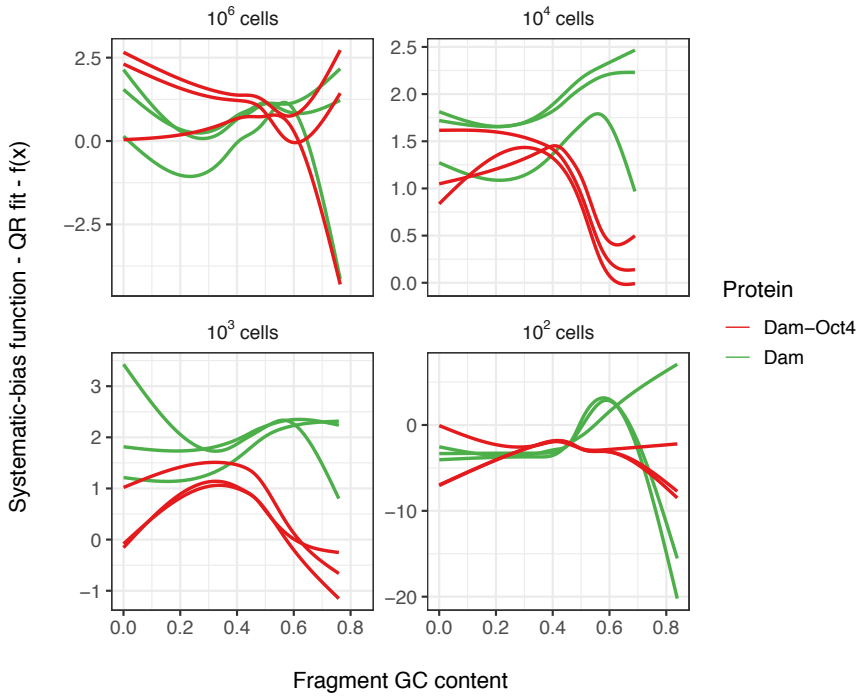

FIGURE 3.31. Graphs of the fragment GC content effect on methylation in low cell number length-normalised Oct4 DamID-seq data.

Graphs of the estimated fragment GC content effect on length-normalised methylation in low cell number Oct4 DamID-seq data from mouse embryonic stem cells (abbreviated ESC). Read counts were normalised for restriction fragment length using conditional quantile normalization from the cqn package (Hansen, Irizarry, and Wu, 2012). The X-axis measures the restriction fragment GC content and the Y-axis measures the fit from the quantile regression model. Specifically, for each sample a systematic bias function using natural cubic splines can be estimated. For each value on the x-axis, the y-axis plots the fitted value generated from each of the sample-specific bias functions. The sample-specific affect of the covariate on read counts can be visualised by plotting the estimates from the bias functions.

3.4.9 Dam preferentially binds euchromatin and regulatory regions

One disadvantage of ChIP-seq in comparison to DamID-seq, is that it requires a highly specific antibody to bind with high affinity to the target protein. Such antibodies can be difficult or expensive to produce, thus limiting most applications of ChIP-seq to

proteins with already validated antibodies. Even then, antibodies for the same protein from multiple manufacturers display different binding characteristics and the experimental procedure must be specifically adapted. In theory, DamID can be applied to any protein for which it is possible to engineer a Dam-fusion expression construct, providing the tether does not interfere with the protein's function. However, just like ChIP-seq antibodies, Dam may exhibit its own binding characteristics which restrict its usefulness to already validated proteins. For example, DNA binding sites in particular chromatin states may be easily accessible by the target protein, but not by the Dam protein. If Dam is unable to methylate restriction sites within these chromatin states, the technology would be unsuitable for a great number of proteins.

To determine the binding characteristics of Dam, a chromatin state model for ESCs and MEFs was built using histone modification, chromatin accessibility, and Dam sequencing data (see Figures 3.32 and 3.33). The model showed Dam was enriched in multiple states (E3, E4, E14, E15, E16, E17, and E18) associated with chromatin accessibility (ATAC) and multiple histone modifications (H3K4me1, H3K4me2, H3K27ac, H3K9ac, H3K4me3, and H3K36me3). These typically mark elements that are responsible for regulating gene expression: H3K4me1 at active and primed enhancers (Local et al., 2018), H3K4me2 at transcription factor binding regions (Wang, Li, and Hu, 2014), H3K27ac at active enhancers (Creyghton et al., 2010), H3K9ac at active promoters (Barski et al., 2007), H3K4me3 at transcribed genes (Liu et al., 2016), and H3K36me3 at nucleosomes (Sims and Reinberg, 2009). Next, the enrichment of each chromatin state overlapping different genomic features was calculated (see Figure 3.34). Out of the seven chromatin states which displayed Dam binding, three of these (E3, E15, and E16) were highly enriched over intergenic regions and the other three (E14, E17, and E18) were moderately enriched over gene and promoter regions. Interestingly,

only one chromatin state E3 was uniquely bound by Dam, and was enriched over intergenic, repeat and microsatellite regions. This is surprising, given that non-unique alignments were filtered out of the sequencing data for all downstream analyses. In theory this would have greatly reduced coverage at regions with a low complexity and large repeat sequences such as those mentioned previously. It is therefore possible that the enrichment observed for state E3 over these repeat regions may actually be underestimated. To establish whether Dam is prevented from methylating restriction sites in certain chromatin states, the enrichment of each chromatin state around all restriction sites in the genome was calculated (see Figure 3.35). All chromatin states except for E3, E4, and E15 were highly enriched, but these were generally unique to Dam so enrichment at the site itself (where DpnI and DpnII cleave) was expected to be reduced. Promoter regions tend to be highly occupied given multiple DNA-binding proteins work together to regulation gene expression. To check if Dam can be obstructed from methylating restriction sites in promoter regions, the enrichment of each chromatin state around all transcription start sites in the genome was calculated (see Figure 3.36). Only chromatin states E17 and E18 (which exhibit a high likelihood of observing Dam) were enriched suggesting that Dam is able to methylate restriction sites within promoter regions. Interestingly, chromatin state E4 (which also exhibits a high likelihood of observing Dam) was enriched around all transcription end sites in the genome (see Figure 3.37). Together these results indicate that Dam predominantly binds accessible chromatin, including enhancers and promoters within intergenic and transcribed regions.

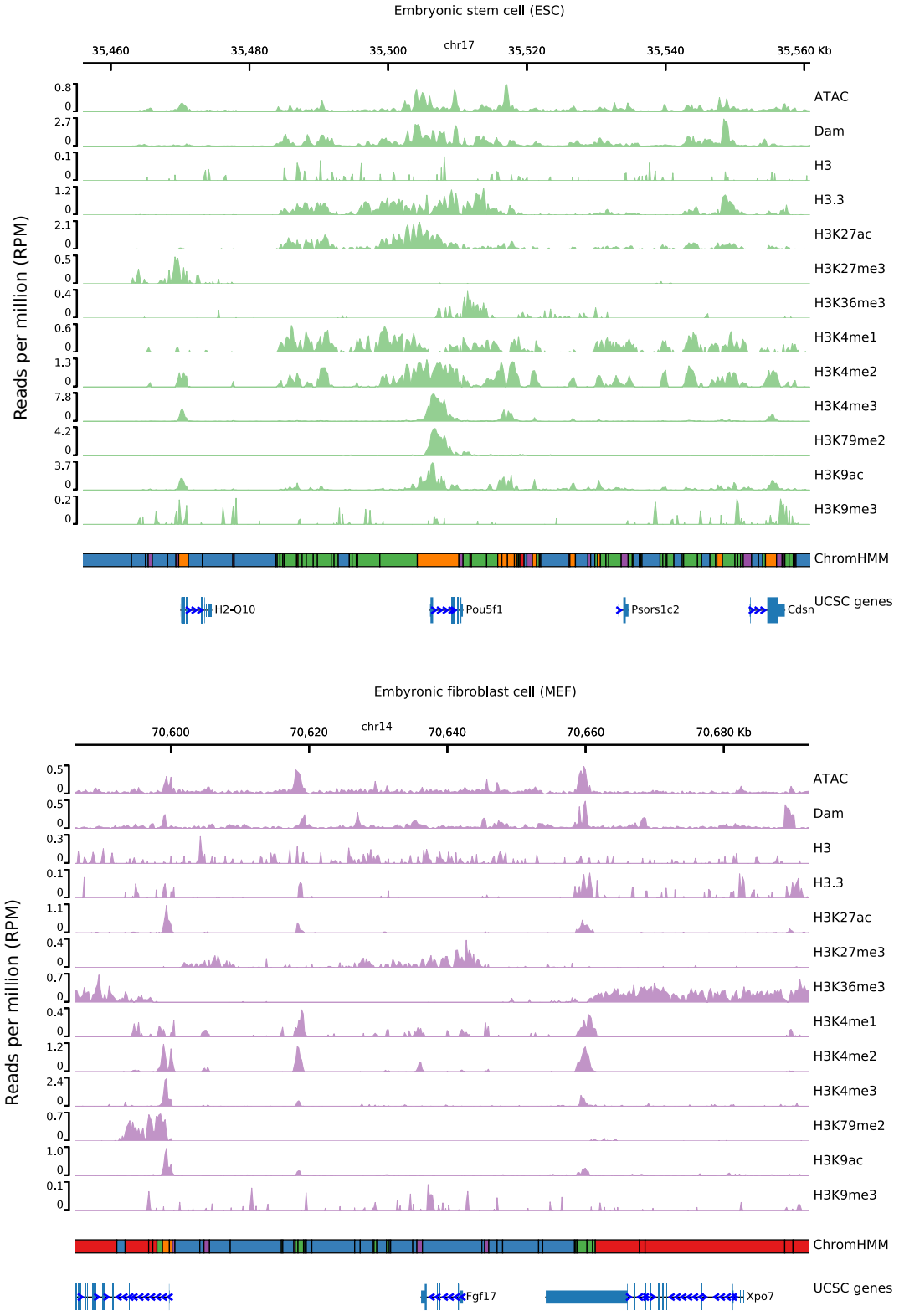

FIGURE 3.32. Tracks of chromatin state annotations produced by ChromHMM.

Tracks of ESC and MEF chromatin state annotations produced by ChromHMM using chromatin accessibility, histone modification, and Dam sequencing data. Descriptions of each chromatin state are listed in Figure 3.33.

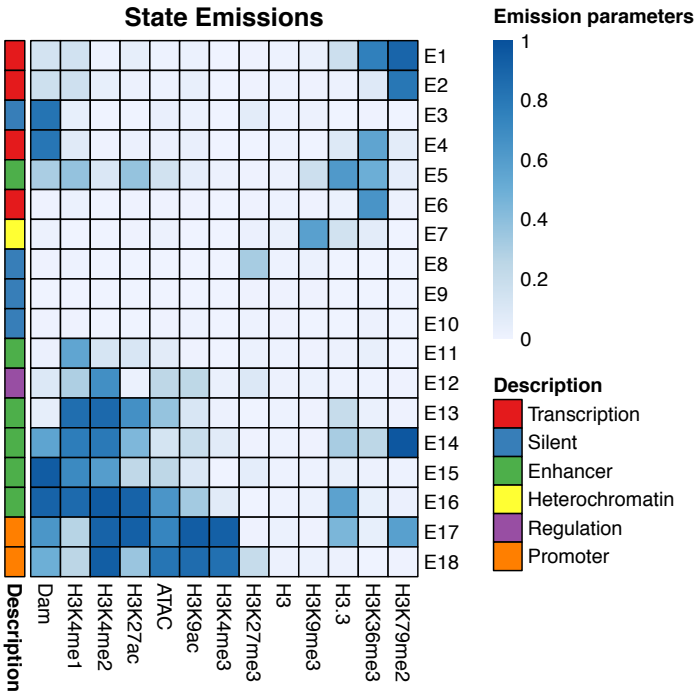

FIGURE 3.33. Heatmap of chromatin state emissions produced by ChromHMM.

Heatmap of ESC and MEF chromatin state emissions produced by ChromHMM using chromatin accessibility, histone modification, and Dam sequencing data. Each row corresponds to a different chromatin state, and each column corresponds to a different epigenomic mark. A darker colour corresponds to a greater probability of observing the epigenomic mark in the chromatin state. All of the sequencing data was re-analysed from public ATAC-seq and ChIP-seq experiments (Chronis et al., 2017).

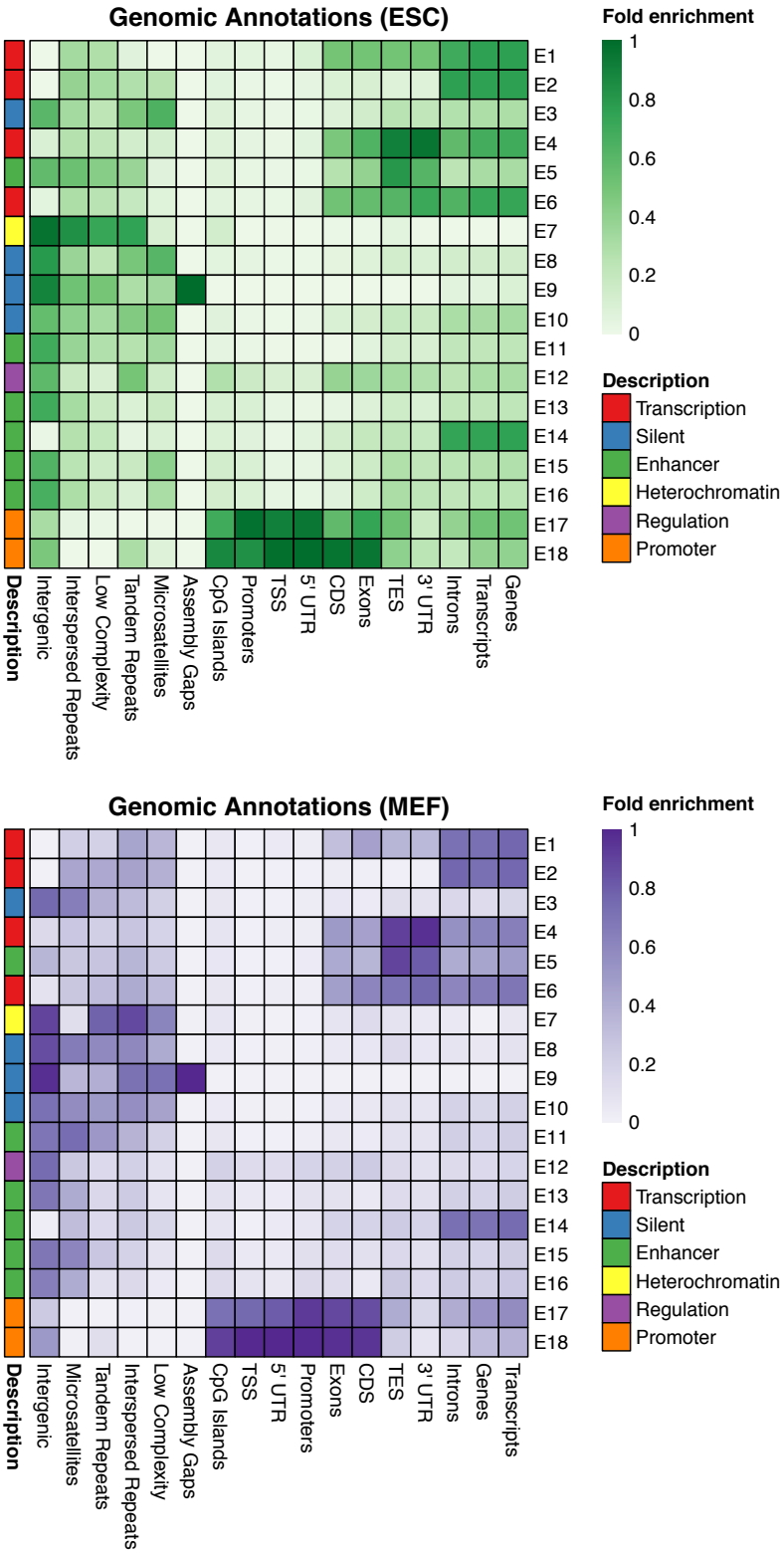

FIGURE 3.34. Heatmap of chromatin state enrichment over genomic features.

Heatmap of ESC and MEF chromatin state enrichment over genomic feature annotations. A darker colour corresponds to a greater enrichment, and there is a column-specific colour scale for the entire heatmap.

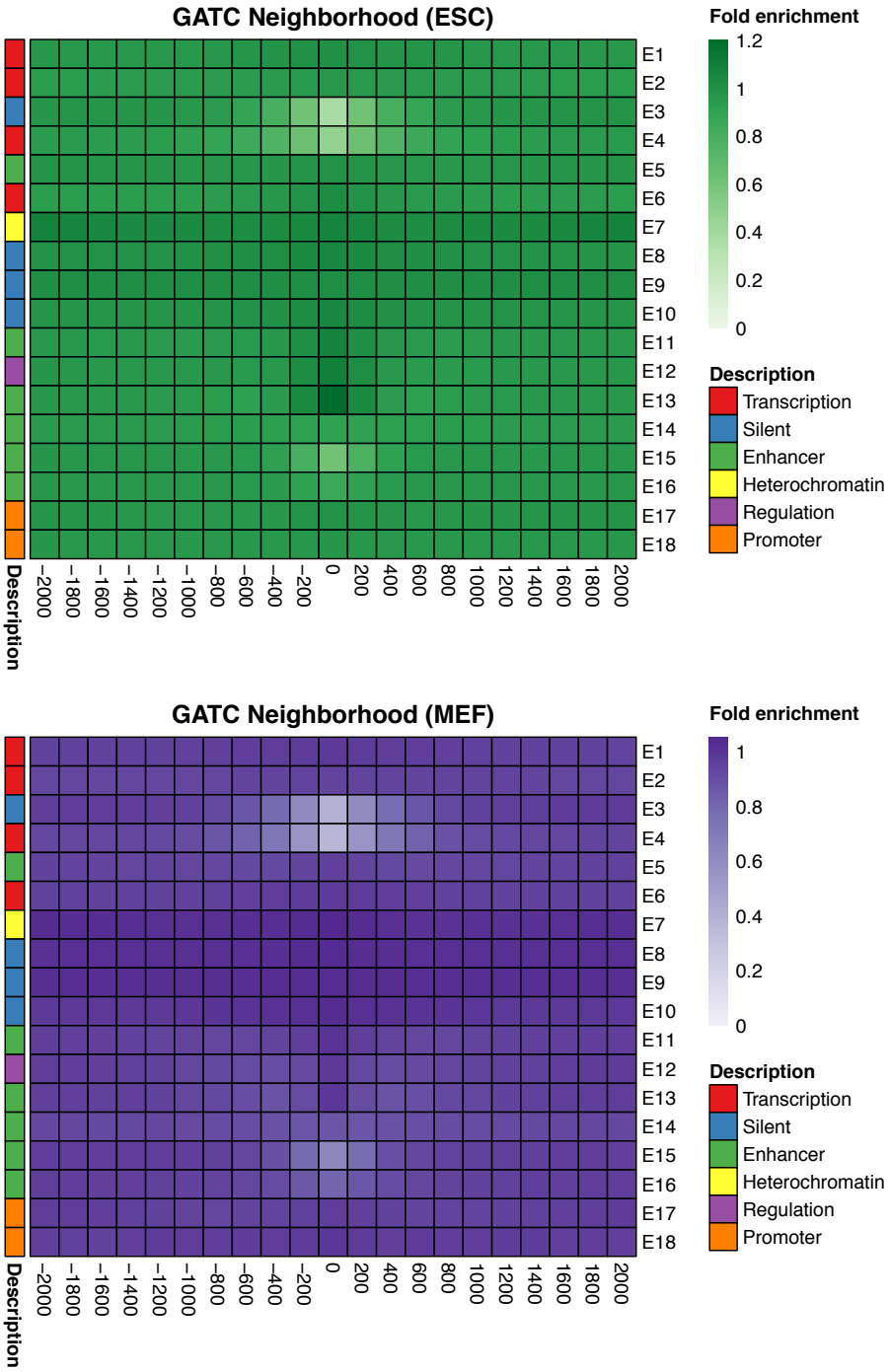

FIGURE 3.35. Heatmap of chromatin state enrichment around GATC sequences.

Heatmap of ESC and MEF chromatin state enrichment 2 kb around GATC sequences. A darker colour corresponds to a greater enrichment, and there is a column-specific colour scale for the entire heatmap.

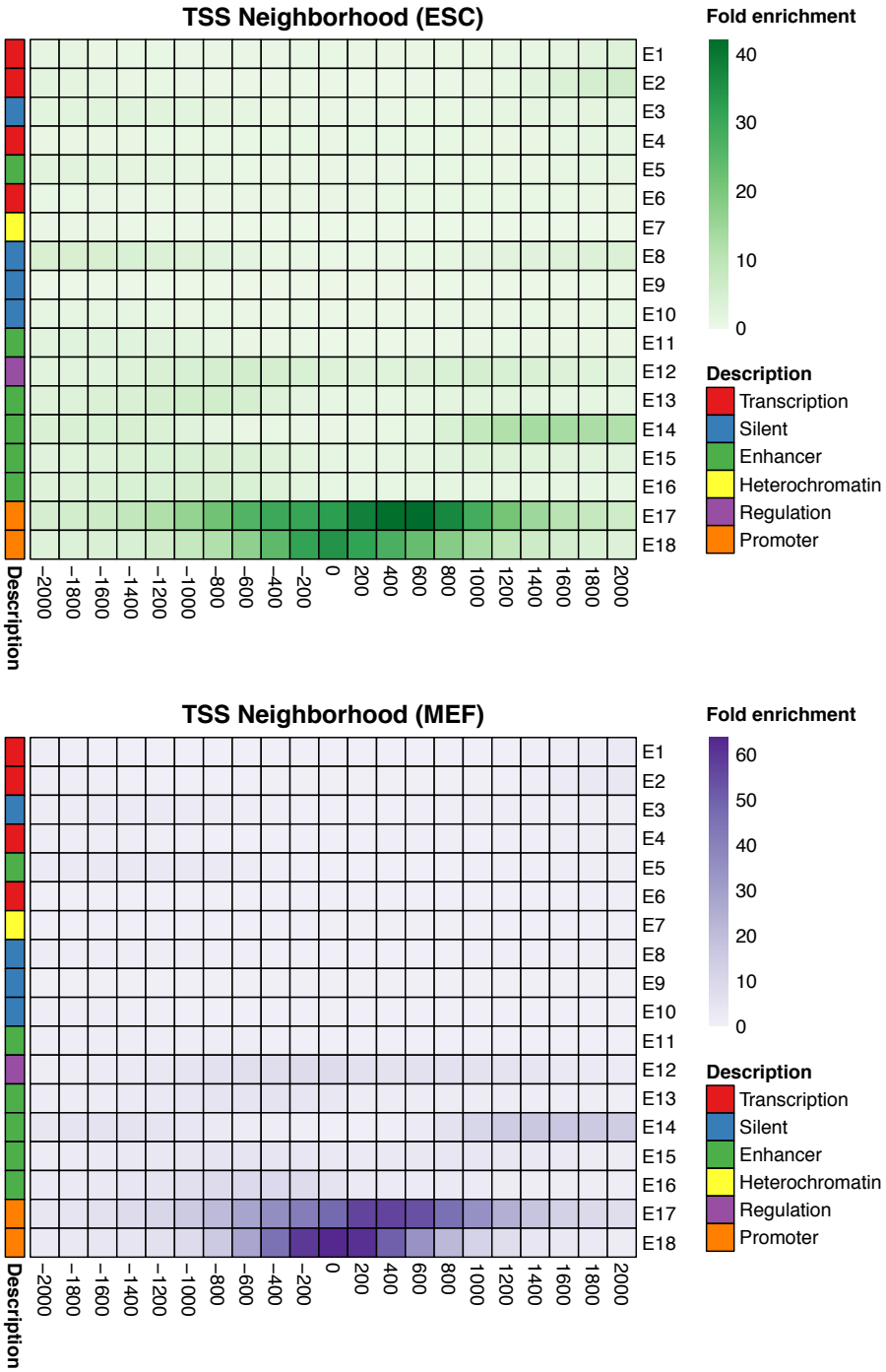

FIGURE 3.36. Heatmap of chromatin state enrichment around TSS sites.

Heatmap of ESC and MEF chromatin state enrichment 2 kb around transcription start sites (abbreviated TSS). A darker colour corresponds to a greater enrichment, and there is a column-specific colour scale for the entire heatmap.

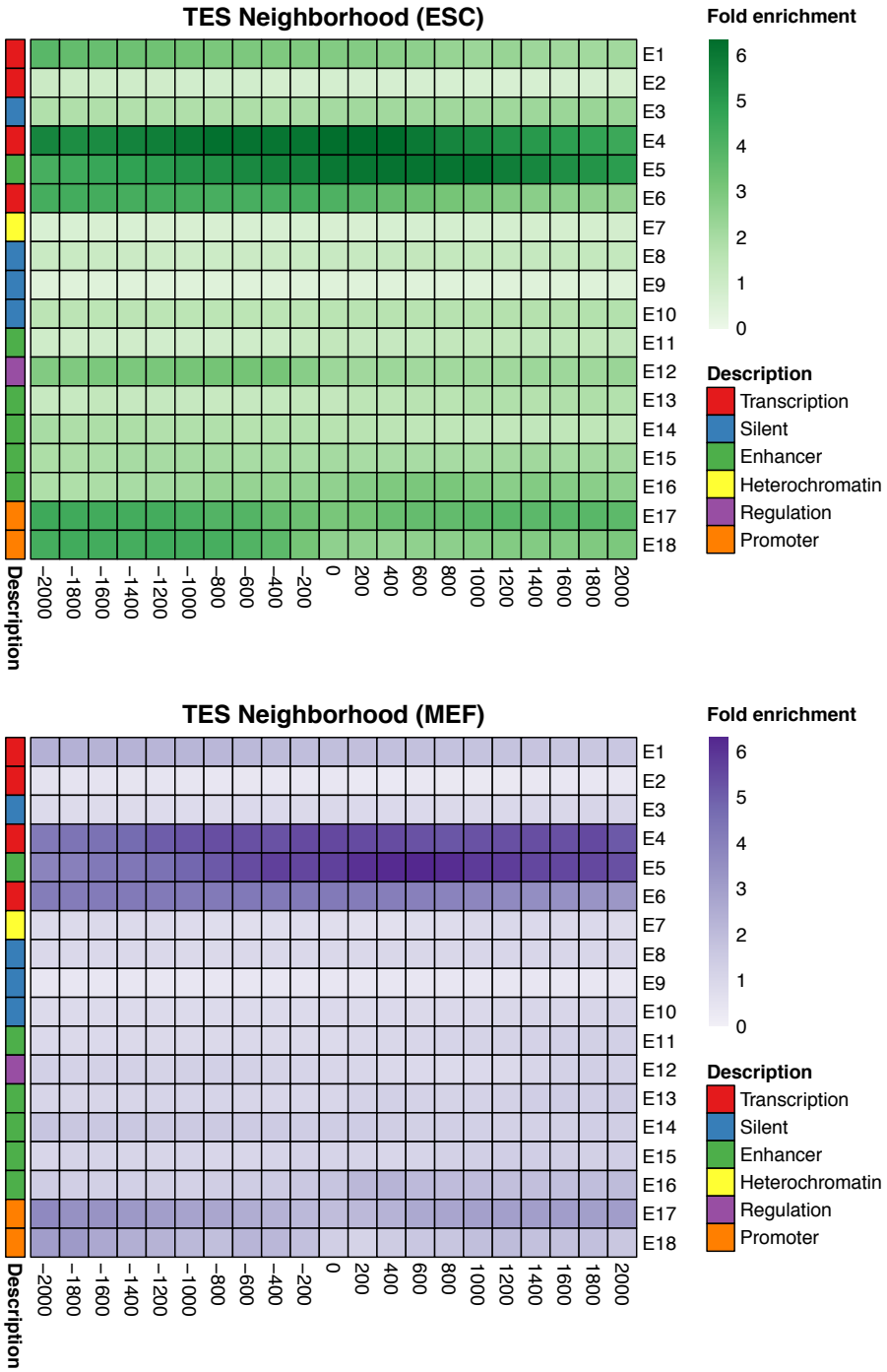

FIGURE 3.37. Heatmap of chromatin state enrichment around TES sites.

Heatmap of ESC and MEF chromatin state enrichment 2 kb around transcription end sites (abbreviated TES). A darker colour corresponds to a greater enrichment, and there is a column-specific colour scale for the entire heatmap.

3.4.10 Impact of m6A methylation in mouse embryonic stem cells

In DamID-seq experiments, binding sites are marked by the addition of a methyl group to the N6 position of adenine (abbreviated m6A) in nearby GATC sequences using an *Escherichia coli* DNA adenine methyltransferase. It is therefore required that cells from the experimental organism lack both an endogenous m6A methyltransferase and de-methyltransferase. In recent years, m6A methylation has been observed at very low levels in multiple eukaryotes: *Drosophila melanogaster* (Zhang et al., 2015), *Caenorhabditis elegans* (Greer et al., 2015), *Homo sapiens* (Xiao et al., 2018), and *Mus musculus* (Wu et al., 2016). These reports cast doubt on the utility of DamID-seq in eukaryotes because DNA may be marked by Dam and the endogenous methyltransferase, leading to false positive DNA binding sites being identified. In addition, true positive DNA binding sites may not be identified because methylation is lost by the as yet unidentified mechanisms that remove endogenous m6A methyltransferase.

To assess the impact of endogenous m6A methylation on DamID-seq data, the m6A methylome from mouse embryonic stem cells was surveyed (see Figure 3.38). The N6-methyladenines were identified by single molecule real time sequencing combined with chromatin immunoprecipitation (SMRT-ChIP) from a published study (Wu et al., 2016). Compared to all adenines in the mouse genome (~739.08 M), only a small proportion (0.046%) of m6A adenines were located within GATC sequences. Endogenous m6A methylation at non GATC sequences cannot interfere with the DamID-seq experiment because the restriction enzyme DpnI used to fragment the DNA only recognises methylated GATC sequences. To check whether the small proportion of endogenous m6A adenines within GATC sequences affected the levels of methylation measured by DamID-seq, read coverage from Oct4 DamID-seq data was plotted over Oct4 ChIP-seq and m6A SMRT-ChIP peak regions (see Figure 3.39). The plots showed that Dam-Oct4 coverage was consistently higher than Dam at Oct4 binding sites, regardless of

whether the binding site was in an m6A methylated region. In addition, Dam-Oct4 coverage was very low at regions which did not contain Oct4 binding but were in an m6A methylated region. In fact, the m6A levels observed here are likely to be even lower due to non-specific background in DNA immunoprecipitation followed by sequencing (DIP-seq) assays, originating primarily due to the intrinsic affinity of IgG for short unmodified DNA repeats (Lentini et al., 2018). In this specific case, it appears that the low level of endogenous m6A methylation does not affect the DamID-seq data. However, more DamID-seq and m6A methylome data from a range of cell types is required to fully investigate this potential fault with DamID-seq technology.

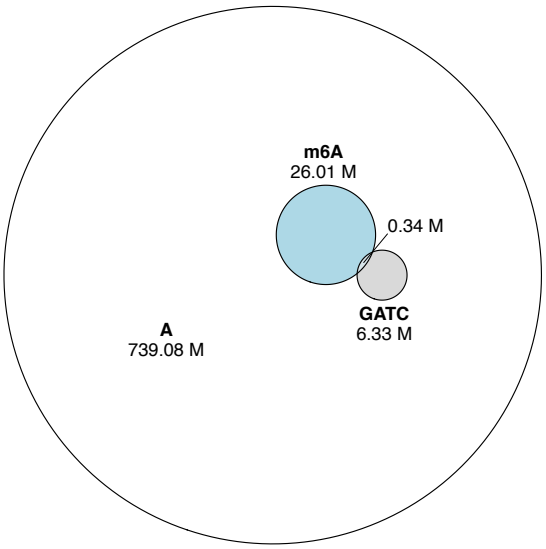

FIGURE 3.38. Proportion of N6-methyladenines within GATC sequences.

The Euler diagram represents the overlap between three sets of adenines in mouse embryonic stem cells: all adenines in the genome (abbreviated A), N6-methyladenines (abbreviated m6A), and adenines in a GATC sequence (abbreviated GATC). The N6-methyladenines were identified by single molecule real time sequencing combined with chromatin immunoprecipitation (Wu et al., 2016). The diagram highlights the small proportion of N6-methyladenines within GATC sequences compared to all adenines in the genome.

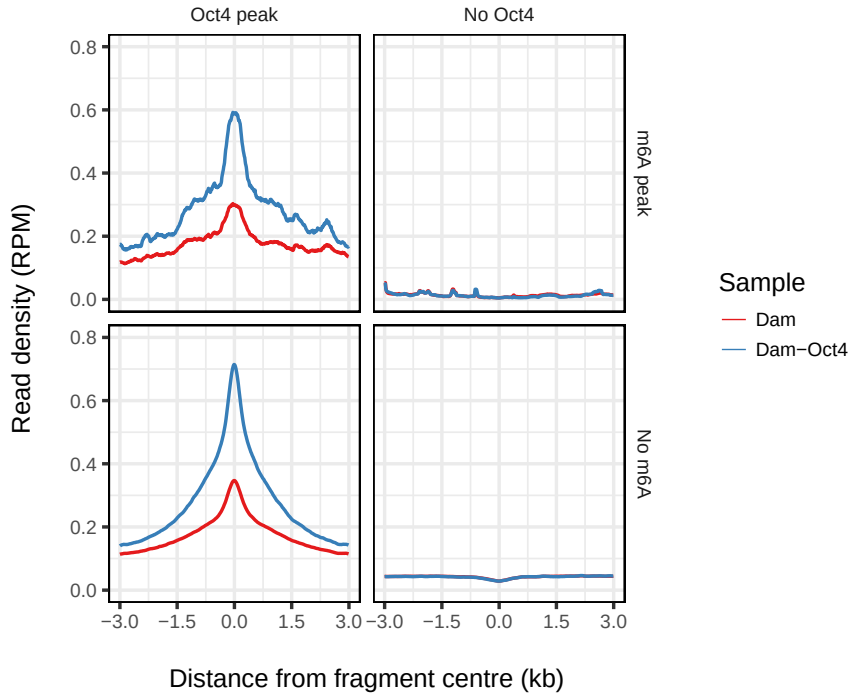

FIGURE 3.39. Graphs of Oct4 ESC DamID-seq read coverage at Oct4 and m6A peaks.

Graphs of Oct4 ESC DamID-seq read coverage at Oct4 ChIP-seq (n = 37,925) and m6A (n = 37,581) SMRT-CHIP peaks (Chronis et al., 2017; Wu et al., 2016).

3.5 Discussion

The results from this chapter showed that the accuracy of DamID-seq experiments can be influenced by differences in the experimental procedure (e.g. polymerase choice and DpnII digestion) and systematic biases in the sequencing data (e.g. restriction fragment length and GC content). In addition, it was demonstrated that Dam binds principally to euchromatin (i.e. loosely packaged DNA) which suggests the sequencing data could be repurposed to identify DNA accessibility sites. Equally important, these results have also demonstrated that Dam binding is not influenced by the local nucleotide composition of restriction sites and that endogenous m6A methylation can

not be detected in the ESC sequencing data analysed. In summary, these results provide a comprehensive overview of the factors which influence the quality of DamID-seq data and highlight the need for rigorous analysis methods to account for the different sources of biases discussed.

Chapter 4

Identification of transcription factor binding from DamID-seq data

4.1 Introduction

Gene expression can be regulated by controlling the rate at which DNA is transcribed into mRNA via the enzyme RNA polymerase. Before transcription occurs, RNA polymerase must first bind to the promoter sequence of a gene alongside other necessary proteins. Transcription factors promote or block the recruitment of RNA polymerase by binding to cis-regulatory elements which influence the structure of DNA at the promoter making it easier or harder for RNA polymerase to initiate transcription (Latchman, 1993). There are approximately 1,500 transcription factors encoded in the mammalian genome, all of which define when and where genes are expressed throughout the development of an organism (Zhou et al., 2017). It is therefore crucial that such important proteins are studied, and their individual contributions catalogued.

Currently, the most popular method to identify genome-wide DNA binding sites for transcription factors and other protein is ChIP-seq (Furey, 2012). It has been used extensively over the last decade to understand the role of transcription factors in gene

expression, especially during differentiation and reprogramming (Takahashi and Yamanaka, 2016). However, conventional ChIP-seq protocols require a clinical grade antibody and a minimum of 10^7 cells to produce a sufficient DNA yield (Gilfillan et al., 2012). This limits the application of ChIP-seq to already validated antibodies (expensive and temperamental) and cells grown *in vitro* under culture (under-represent *in vivo* complexity). Although recent advances such as CUT&RUN (Skene and Henikoff, 2017) and high intensity UV ChIP-seq (Steube et al., 2017) have managed to reduce the number of cells significantly, both of these techniques still require a highly specific antibody.

The limitations of ChIP-seq can be avoided altogether using DamID-seq because no antibodies or inefficient enzymes are required to enrich for factor-bound chromatin and genetic engineering can be used to measure *in vivo* binding with tissue-specific promoter driven Dam expression (Marshall et al., 2016). Although DamID-seq has been used extensively in *Drosophila* and other model systems, only recently have protocols for mammalian cells combined with high-throughput sequencing been published (Tosti et al., 2018). As a consequence, there has been no inspection of the accuracy and sensitivity of the sequencing data, including proper methods to analyse differential methylation. This eventually could lead to artefactual results and inappropriate conclusions being made which pollute our ideas about transcription factor binding. Therein lies the impetus to perform a comprehensive assessment of DamID-seq data by comparison with ChIP-seq data, including development of appropriate analysis methods.

4.2 Aims

The aims of this chapter are to evaluate different normalisation strategies for DamID-seq data, develop an accurate and sensitive peak calling method for transcription factor binding, and investigate systematic differences between ChIP-seq and DamID-seq assays. To accomplish these aims, normalisation strategies were compared with corresponding qDamID data from roughly two hundred restriction fragments, a software package for the comprehensive analysis of DamID-seq data was developed, and DamID-seq data of multiple transcription factors (Oct4 and Sox2) from multiple cell types (ESC and NSC) with multiple cell numbers (10^6, 10^4, 10^3) was compared with a large collection of corresponding ChIP-seq data from public experiments.

4.3 Attribution

The DamID-seq libraries were generated by Dr Luca Tosti - a previous PhD student in Prof Keisuke Kaji's research group - and the sequencing data was analysed by the author, James Ashmore. The ChIP-seq libraries were generated by the relevant research groups and the public sequencing data was re-analysed. The qDamID experiments were performed by the author, under supervision from Prof Keisuke Kaji.

4.4 Results

4.4.1 Analysis of published Oct4 ESC ChIP-seq experiments

In order to determine the accuracy and sensitivity of DamID-seq data, a comprehensive set of transcription factor binding sites for comparative analysis was first required. There is an exponential wealth of ChIP-seq data in public databases which can be downloaded and re-analysed for this objective, and additionally the variability

between experiments can be investigated. A comprehensive search for Oct4 ESC ChIP-seq data from multiple databases (including ArrayExpress, ENA, ENCODE, GEO, and SRA) was performed and then manually checked to ensure wild type experimental conditions were followed. Wild type was defined as any combination of antibody and cell line without deleterious experimental treatment or genetic modification. Approximately 32 ChIP-seq experiments met these criteria (see Table 4.1) and the raw sequencing data was analysed using the same computational pipeline to ensure technical consistency (see Figure 4.1).

Reference	Condition	Cell	Strain	Antibody	Media
Aksoy I (2013)	TgSox2-V5	KH2	C57BL/6 x 129/Sv	sc-8628	Serum
Ang YS (2011)	WT	CCE	129S/SvEv-Gpi1	sc-8628	Serum
Buecker C (2014)	WT	R1	129X1/SvJ x 129S1/Sv	sc-8628	2i
Chen X (2008)	WT	E14TG2a	129P2/Ola	sc-8628	Serum
Chronis C (2017)	WT	v6.5	C57BL/6 x 129S4/SvJae	AF1759	Serum
Das PP (2014)	GFP shRNA	J1	129S4/SvJae	sc-8628	Serum
Flynn RA (2016)	Non-targeting ASO	v6.5	C57BL/6 x 129S4/SvJae	sc-8629	Serum
Galonska C 1 (2015)	WT	KH2	C57BL/6 x 129/Sv	sc-8628	2i
Galonska C 2 (2015)	WT	KH2	C57BL/6 x 129/Sv	sc-8628	Serum
Hardison R (2014)	WT	E14TG2a	129P2/Ola	ab19857	Serum
Hu G (2013)	Luc shRNA	CMTI-1	129S6/SvEvTac	ab19857	Serum
Jacinto FV (2015)	Scrambled shRNA	E14TG2a	129P2/Ola	sc-5279	Serum
Jang H (2012)	TgOct4-Flag (-Dox)	ZHBTc4	129P2/Ola	F3165	Serum
Karwacki-Neisius V 1 (2013)	WT	E14TG2a	129P2/Ola	sc-8628	2i
Karwacki-Neisius V 2 (2013)	Oct4 (+/-)	OKO160	129P2/Ola	sc-8628	2i
King HW 1 (2017)	TgOct4 (+Dox)	ZHBTc4	129P2/Ola	C30A3C1	Serum
King HW 2 (2017)	TgOct4 (-Dox)	ZHBTc4	129P2/Ola	C30A3C1	Serum
Krishnakumar R (2016)	miR-290/302	v6.5	C57BL/6 x 129S4/SvJae	sc-9081	2i
Liu Y 1 (2017)	Asynchronous (-H3Ser10p)	v6.5	C57BL/6 x 129S4/SvJae	sc-8628	Serum
Liu Y 2 (2017)	Mitotic (+H3Ser10p)	v6.5	C57BL/6 x 129S4/SvJae	sc-8628	Serum
Liu Z (2017)	WT	129/Sv	129/Sv	sc-8628	Serum
Marson A (2008)	WT	v6.5	C57BL/6 x 129S4/SvJae	sc-8628	Serum
Miller A (2016)	Mbd3 (+/-)	E14TG2a	129P2/Ola	sc-8628	2i
Okashita N (2016)	WT	E14TG2a	129P2/Ola	sc-8629	Serum
Shen Z (2017)	WT	C57BL/6	C57BL/6	sc-8629	2i
Shin J 1 (2016)	G1 (-Noc)	E14TG2a	129P2/Ola	sc-5279	Serum
Shin J 2 (2016)	G2/M (+Noc)	E14TG2a	129P2/Ola	sc-5279	Serum
Tu S (2016)	WT	KH2	C57BL/6 x 129/Sv	sc-5279	2i
Wang L (2014)	Non-targeting ASO	J1	129S4/SvJae	sc-8628	Serum
Whyte WA (2013)	WT	v6.5	C57BL/6 x 129S4/SvJae	sc-8628	Serum
Xu T (2015)	WT	E14TG2a	129P2/Ola	ab19857	Serum
Yang SH (2014)	Rex1-GFP	E14TG2a	129P2/Ola	sc-8628	2i

TABLE 4.1. Collection of published Oct4 ESC ChIP-seq experiments.

This table contains the metadata for 32 published Oct4 ESC ChIP-seq experiments. They were found by automating queries to the ArrayExpress, ENA, ENCODE, GEO, and SRA biological sequence databases. Each experiment is referenced using the first author's last name and year of publication (see Chapter 2 for a full list of accession numbers).

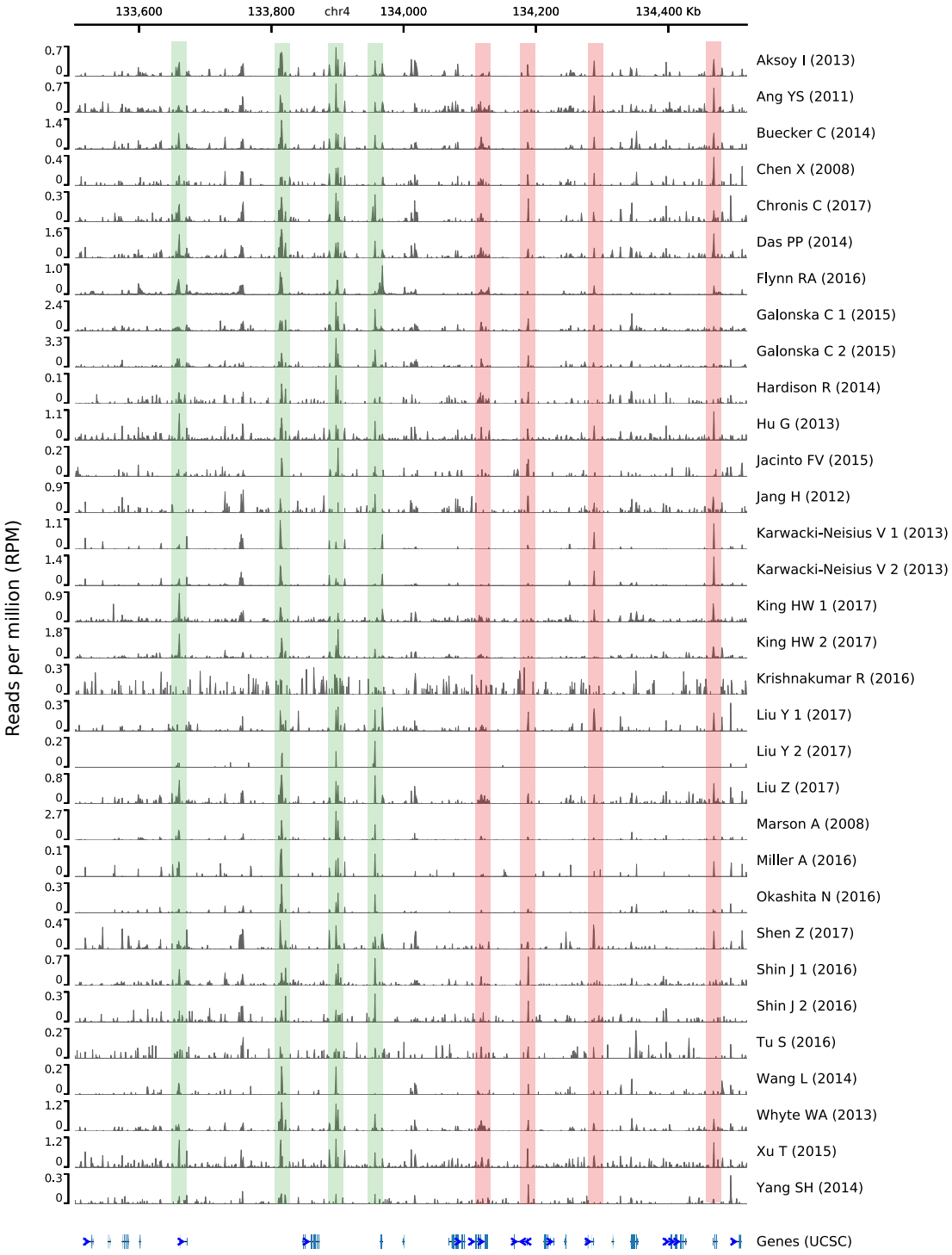

FIGURE 4.1. Genomic snapshot of published Oct4 ESC ChIP-seq experiments at a representative locus (chr4:133,500,000-134,500,000).

The genome browser tracks display the input-subtracted read coverage from published Oct4 ESC ChIP-seq experiments at a representative locus (chr4:133,500,000-134,500,000). Each track is scaled independently to mitigate large differences in the signal-to-noise ratio from different experiments. The green and red bars indicate reproducible and variable regions of enrichment, respectively.

The quality of the sequencing data was checked using a variety of metrics proposed by the ENCODE and CISTROME consortia based upon sequence quality, mapping quality, library complexity, and enrichment of immunoprecipitated chromatin (see Figure 4.2). None of the experiments passed all of the quality control metrics and many underperformed according to library complexity. The library size was still sufficient according to previous guidelines for transcription factor binding in mammalian cells, and enrichment of immunoprecipitated chromatin was generally successful across most of the experiments. Despite their quality, meaningful biological information has already been extracted from these experiments so not much consideration was given to interpreting these results at this stage of the investigation.

FIGURE 4.2. Quality metrics for published Oct4 ESC ChIP-seq experiments.

The quality of each ChIP-seq experiment was evaluated using quality metrics proposed by the ENCODE consortia and the CISTROME project. The green and red circles indicate whether or not an experiment passed the recommended guidelines for transcription-factor ChIP-seq experiments (Landt et al., 2012). The sequence quality metrics include the average read length (Length), the total number of reads (Reads), and the average read quality (Quality). The mapping quality metrics include the number of mapped reads (Mapped), the number of uniquely mapped reads (Unique), and the number of uniquely mapped reads after PCR duplicate removal (Usable). The library complexity metrics include the non-redundant fraction (NRF), the PCR bottleneck coefficient 1 (PBC1), and the PCR bottleneck coefficient 2 (PBC2). The ChIP enrichment metrics include the normalised-strand cross-correlation (NSC), the relative strand cross-correlation (RSC), and the fraction of reads in peaks (FRiP).

The similarity between experiments was measured by calculating Spearman's rank correlation coefficient on the number of aligned reads within 500 bp windows along the genome. The correlation coefficients ranged between -0.20 and 0.67 but on average experiments were only 0.22 correlated (see Figure 4.3). This was concerning as it implied that the ChIP-seq experiments were not particularly reproducible, however this result could change substantially based on the size of the window because larger

windows average out fluctuations whilst smaller ones achieve greater scrutiny. Transcription factor binding sites were detected by comparing the pileup of sequencing reads along the genome between the immunoprecipitated library and the naked DNA library in a process otherwise known as peak calling. The number of peaks ranged between 0 and 95,678 but on average 19,636 were called (see Figure 4.4). The discrepancy between the number of peak calls was concerning given the experiments were all supposed to be measuring the same biological phenomenon. To identify technical aspects which may be responsible, regression models were used to evaluate the relationship between the quality metrics and the number of peak calls (see Figure 4.5). None of the regression models showed a strong linear relationship, even the number of mapped reads - a commonly relied upon measure of appraisal for sequencing experiments was not predicative of the number of peaks called.

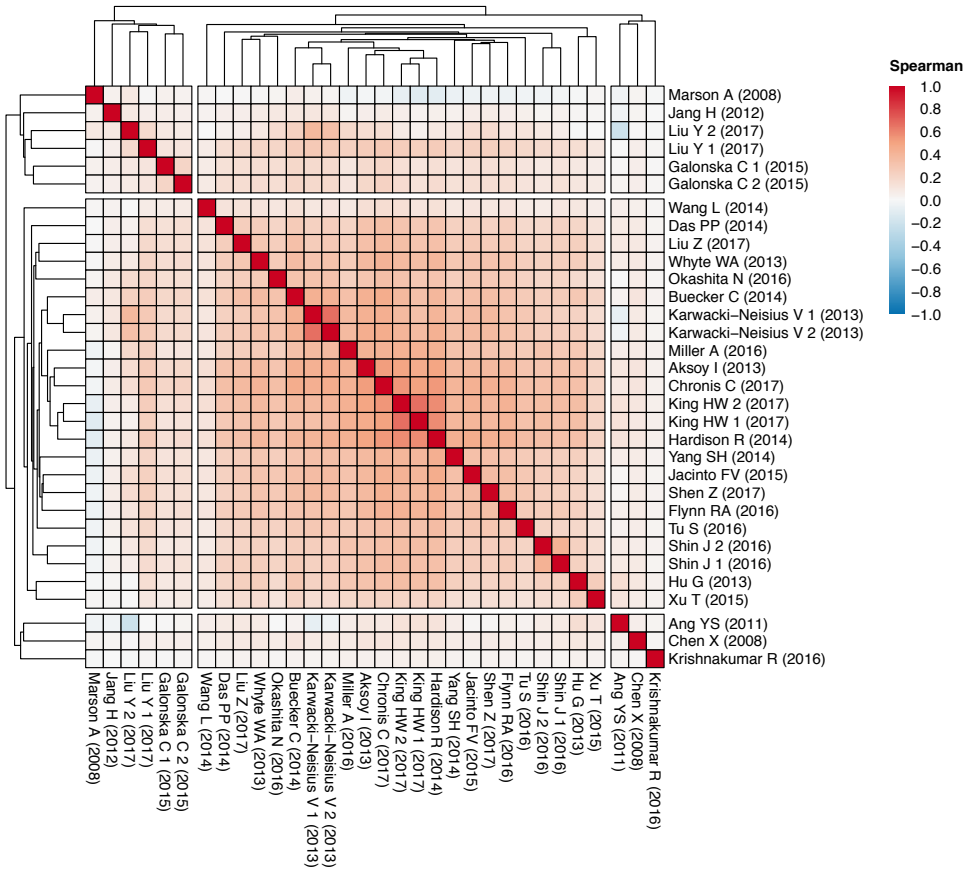

FIGURE 4.3. Spearman correlation of published Oct4 ESC ChIP-seq experiments.

Spearman's rank correlation coefficient was used to measure the similarity in read coverage between experiments. The coefficients ranged between -0.20 and 0.67 but on average experiments were only 0.22 correlated.

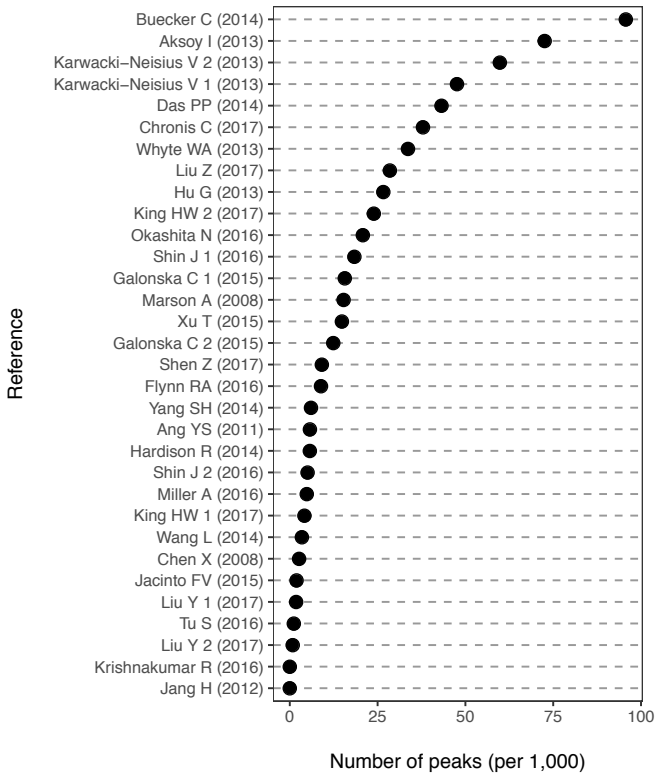

FIGURE 4.4. Number of peaks from published Oct4 ESC ChIP-seq experiments.

Transcription factor binding sites were detected from published Oct4 ESC ChIP-seq experiments by peak calling using MACS2 (FDR < 0.05). The number of peaks ranged between 0 and 95,678 indicating a large variability between experiments.

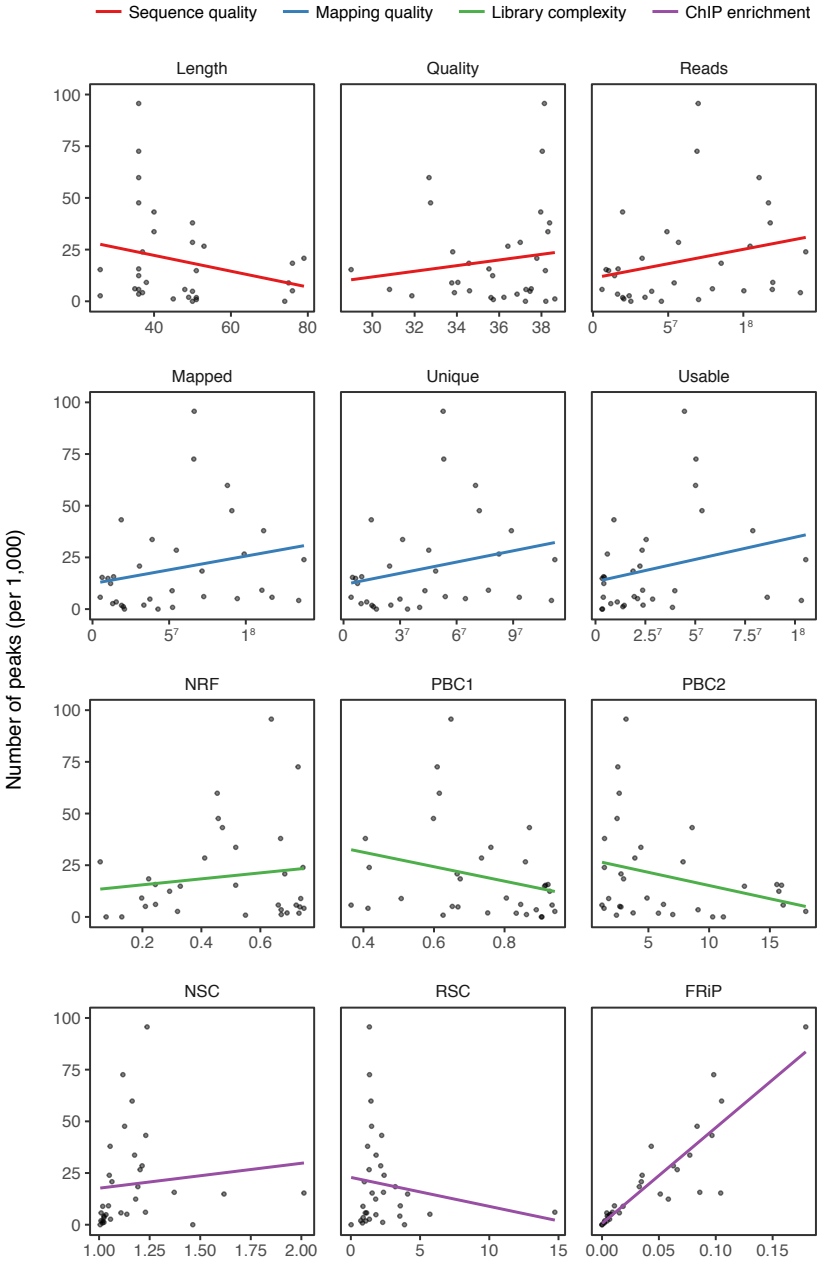

FIGURE 4.5. Relationship between quality metrics and number of peaks from published Oct4 ESC ChIP-seq experiments.

The panel of graphs display the relationship between a single quality metric and the number of peaks called from published Oct4 ESC ChIP-seq experiments. The coloured lines represent a linear regression model used to evaluate the relationship between the two variables. The sequence quality metrics include the average read length (Length), the total number of reads (Reads), and the average read quality (Quality). The mapping quality metrics include the number of mapped reads (Mapped), the number of uniquely mapped reads (Unique), and the number of uniquely mapped reads after PCR duplicate removal (Usable). The library complexity metrics include the non-redundant fraction (NRF), the PCR bottleneck coefficient 1 (PBC1), and the PCR bottleneck coefficient 2 (PBC2). The ChIP enrichment metrics include the normalised-strand cross-correlation (NSC), the relative strand cross-correlation (RSC), and the fraction of reads in peaks (FRiP).

The similarity between experiments was then measured by calculating Jaccard's similarity coefficient based on the size of the intersection and union of peak calls along the genome. The correlation coefficients ranged between 0 and 0.50 but on average experiments were only 0.11 correlated (see Figure 4.6). Under the relatively naïve assumption that peaks between experiments should be highly correlated given they are supposed to be measuring similar binding sites along the genome, this result was surprising. Perhaps a more appropriate null hypothesis for this comparison would be that the correlation between studies is the same as the correlation between replicates within a study. However, technical variation between experiments makes the null hypothesis unlikely, therefore the 0.1 correlation reported was relatively unsurprising given that on average replicates within the same study were only 0.27 correlated. It is also difficult to assess whether this correlation is surprising given, to the best of our knowledge, there are few similar comparisons in the literature. The closest examples either measure reproducibility between replicates within multiple studies (Devailly et al., 2015) measure correlation across hundreds of datasets in order to pull out general information regarding the properties of ChIP-seq data (Landt et al., 2012; Liu et al., 2011). Additionally, engagement of the DNA-binding protein with the genome is another factor which may influence the expected correlation between experiments. For example, it has been shown that engagement of Oct4 with the genome is dynamic and context-dependent (i.e. "naïve" versus "ground state" versus "primed" stem cell states) in early differentiation (Simandi et al., 2016). Overall, this dynamism could contribute to highly variable binding sites and a low correlation between the ChIP-seq experiments. The experiments were also significantly clustered (AU $\geq$ 95) using multiscale bootstrap resampling into three groups, which was indicative of some unobserved variable (see Figure 4.7).

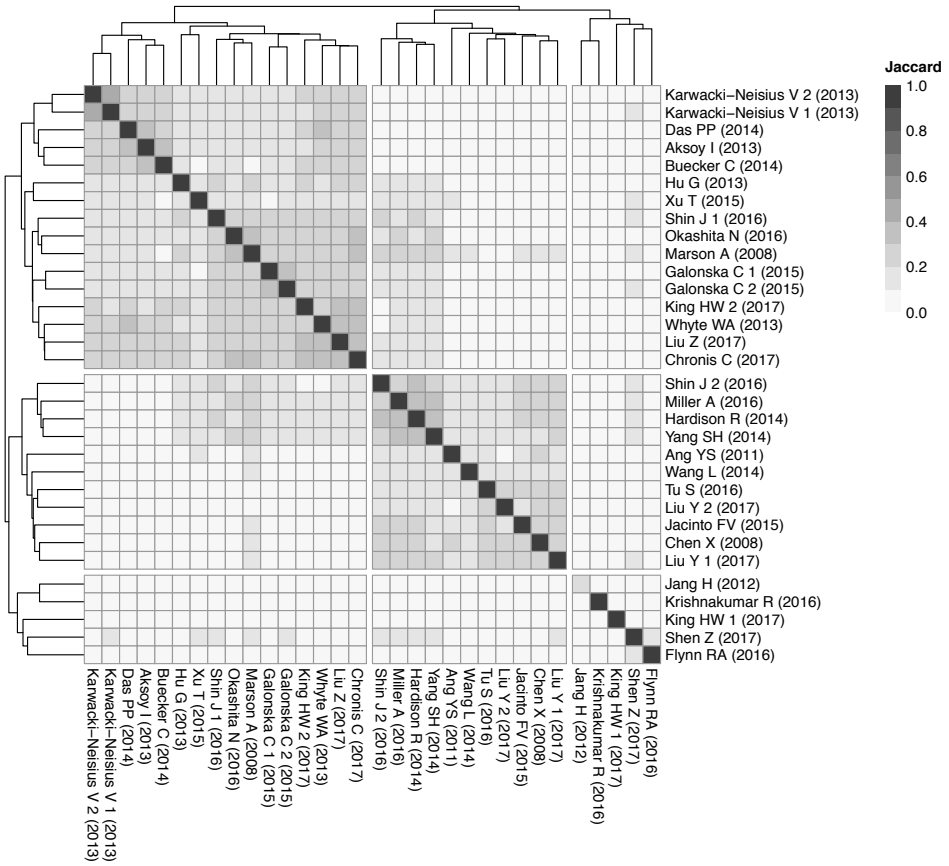

FIGURE 4.6. Jaccard correlation of published Oct4 ESC ChIP-seq experiments.

Jaccard's correlation coefficient was used to measure the similarity in peak calls between experiments. The correlation coefficients ranged between 0 and 0.50 but on average experiments were only 0.11 correlated.

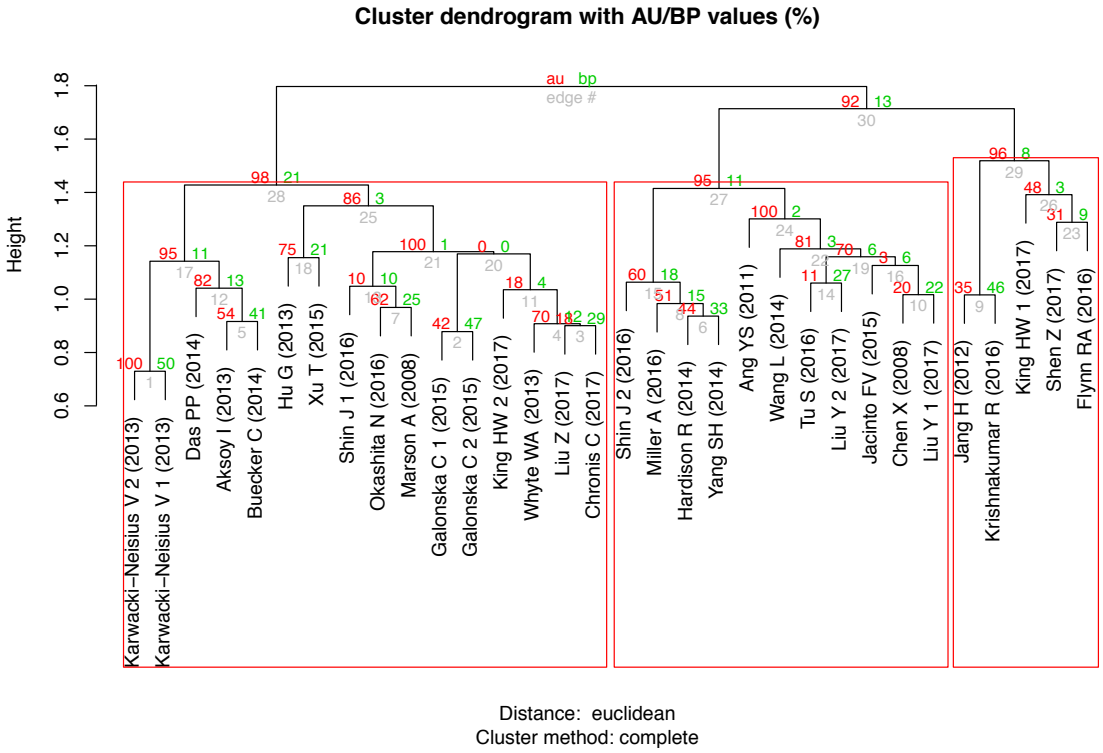

FIGURE 4.7. Bootstrap analysis of Jaccard correlation clustering.

Hierarchical clustering of 32 ChIP-seq experiments using Jaccard's correlation coefficient. Values at branches are the approximately unbiased (AU) p-values (left), bootstrap probability (BP) values (right), and cluster labels (bottom). Clusters with AU $\geq$ 95 are indicated by the red rectangles.

It was possible that independent experimental procedures may have contributed to the clustering pattern, so the similarity between peak calls was cross-referenced against different experimental variables: First, the media used to culture the cells (serum or 2i) was inspected because these have been shown to induce distinct expression profiles known as naïve and ground state pluripotency (Kolodziejczyk et al., 2015). There was no obvious clustering of the experiments based upon this variable which was surprising given the experimental evidence which shows naïve and ground state pluripotent cells have different characteristics (see Figure 4.8).

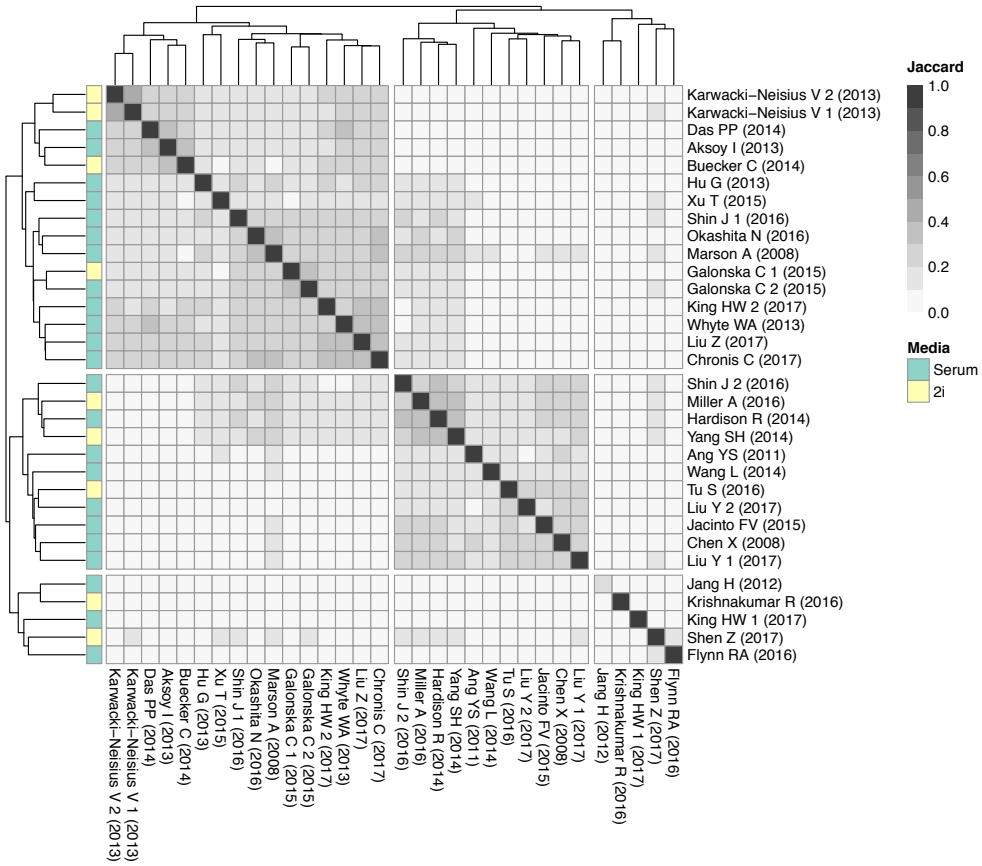

FIGURE 4.8. Cell culture media of published Oct4 ESC ChIP-seq experiments.

The cell culture media used in each published Oct4 ESC ChIP-seq experiment is displayed alongside the heatmap of the Jaccard correlation coefficients.

Second, the cell line used to generate the population was examined because it has been reported that different lines exhibit varying levels of pluripotency markers (Ginis et al., 2004). However, no apparent clustering of the experiments based upon this variable was observed (see Figure 4.9).

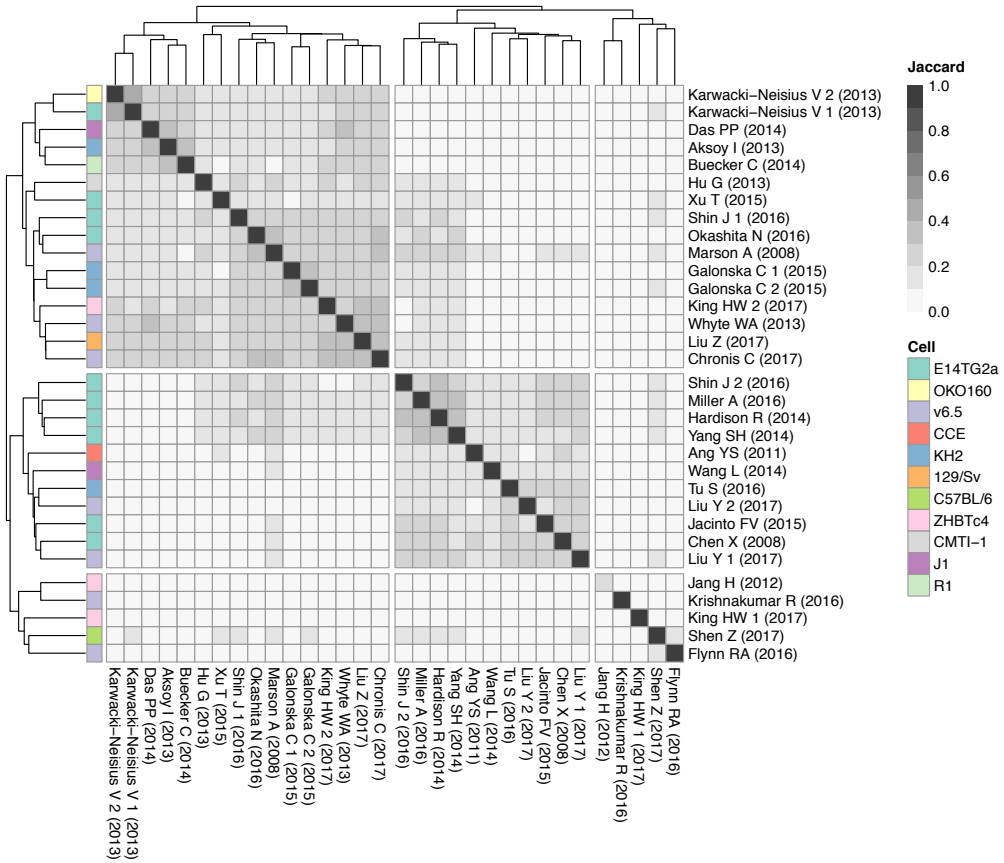

FIGURE 4.9. Cell lines of published Oct4 ESC ChIP-seq experiments.

The cell lines used in each published Oct4 ESC ChIP-seq experiment is displayed alongside the heatmap of the Jaccard correlation coefficients.

Third, the mouse strain used to generate the cell lines was tested because it has been demonstrated that different strains require slightly different procedures for the establish of pluripotency (Kawase et al., 1994). Again, no apparent clustering of the experiments based upon this variable was observed (see Figure 4.10).

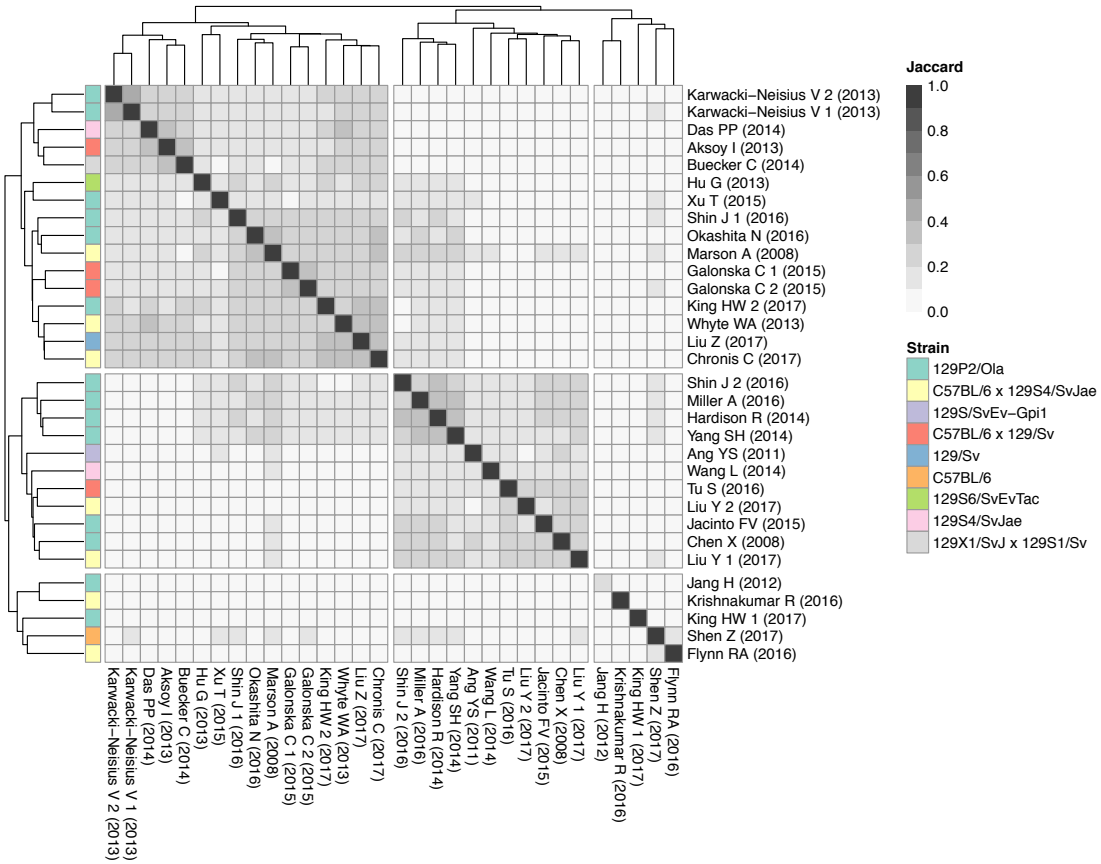

FIGURE 4.10. Mouse strains of published Oct4 ESC ChIP-seq experiments.

The mouse strains used in each published Oct4 ESC ChIP-seq experiment is displayed alongside the heatmap of the Jaccard correlation coefficients.

Lastly, the antibody used to immunoprecipitate the chromatin was checked because it is well known that different antibodies have different binding characteristics that can bias which binding sites are detected (Jager and Vaegter, 2016). Just as before, no apparent clustering of the experiments based upon this variable was detected (see Figure 4.11).

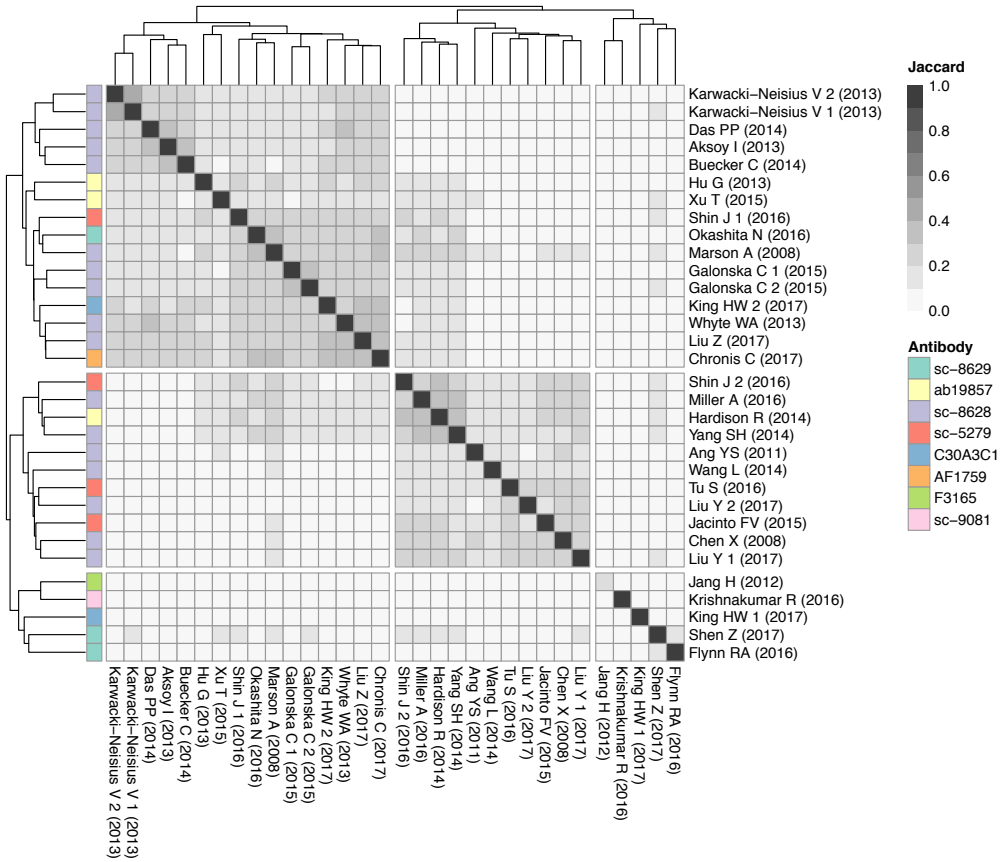

FIGURE 4.11. ChIP antibodies of published Oct4 ESC ChIP-seq experiments.

The ChIP antibodies used in each published Oct4 ESC ChIP-seq experiment is displayed alongside the heatmap of the Jaccard correlation coefficients.

The inability to identify a causative variable for the clustering pattern was surprising and demonstrated that experiments should be carefully cross-validated, particularly if a novel cell line or antibody is employed. To investigate the differences at a higher resolution, binding sites which were bound or unbound across all experiments were visualised (see Figure 4.12). The occupancy map demonstrated that a large number of sites were either unique or only bound in a small handful of experiments. These sites would have greatly reduced the correlation coefficients previously calculated and indicate a high level of variability. Interestingly, a small group of sites near the top and

bottom of the map appeared to be bound in just over half of the experiments. These sites also exhibited high read coverage (see Figure 4.13) and were enriched for the Oct4 binding motif (see Figure 4.14).

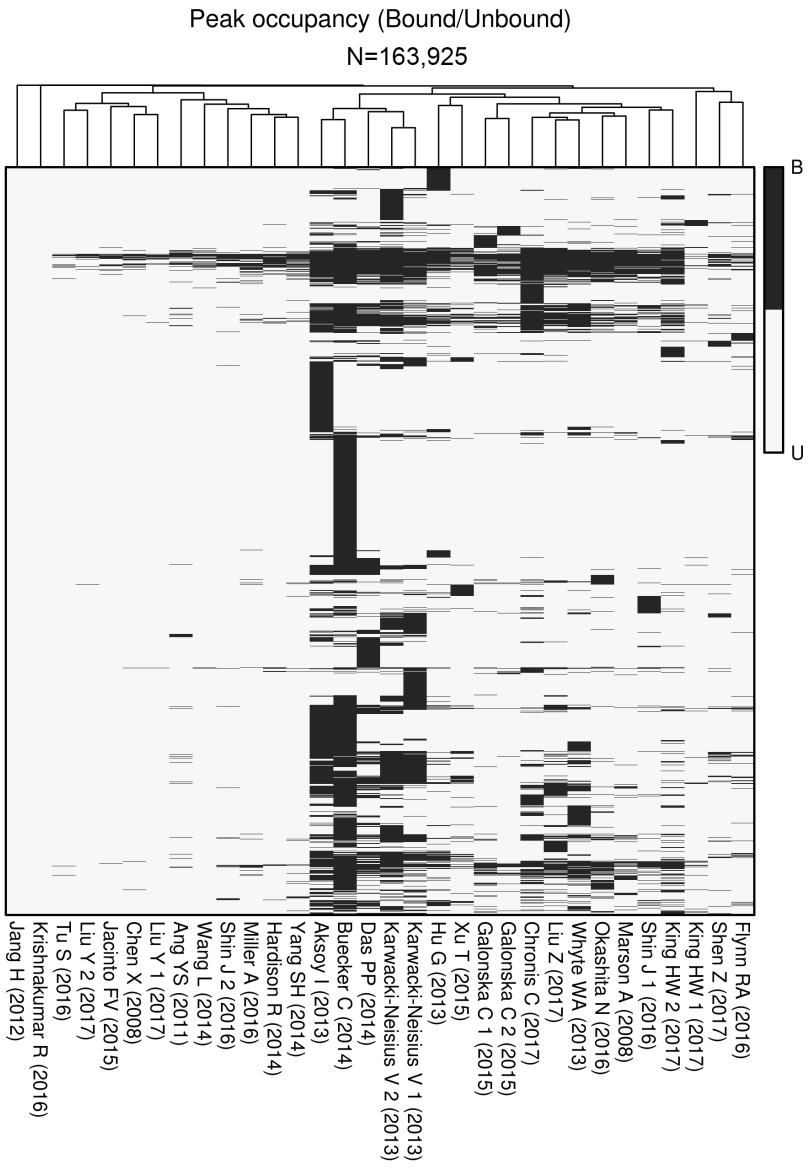

FIGURE 4.12. Peak occupancy from published Oct4 ESC ChIP-seq experiments.

Sites which are bound or unbound in each ChIP-seq experiment are coloured black and grey, respectively. Each row of the heatmap represents a peak region within the genome called from at least one of the analysed ChIP-seq experiments. Each column of the heatmap represents a single experiment whose ChIP-seq data was analysed. The heatmap shows which peaks were called across multiple experiments, revealing that a large number of sites are bound in only a handful of experiments.

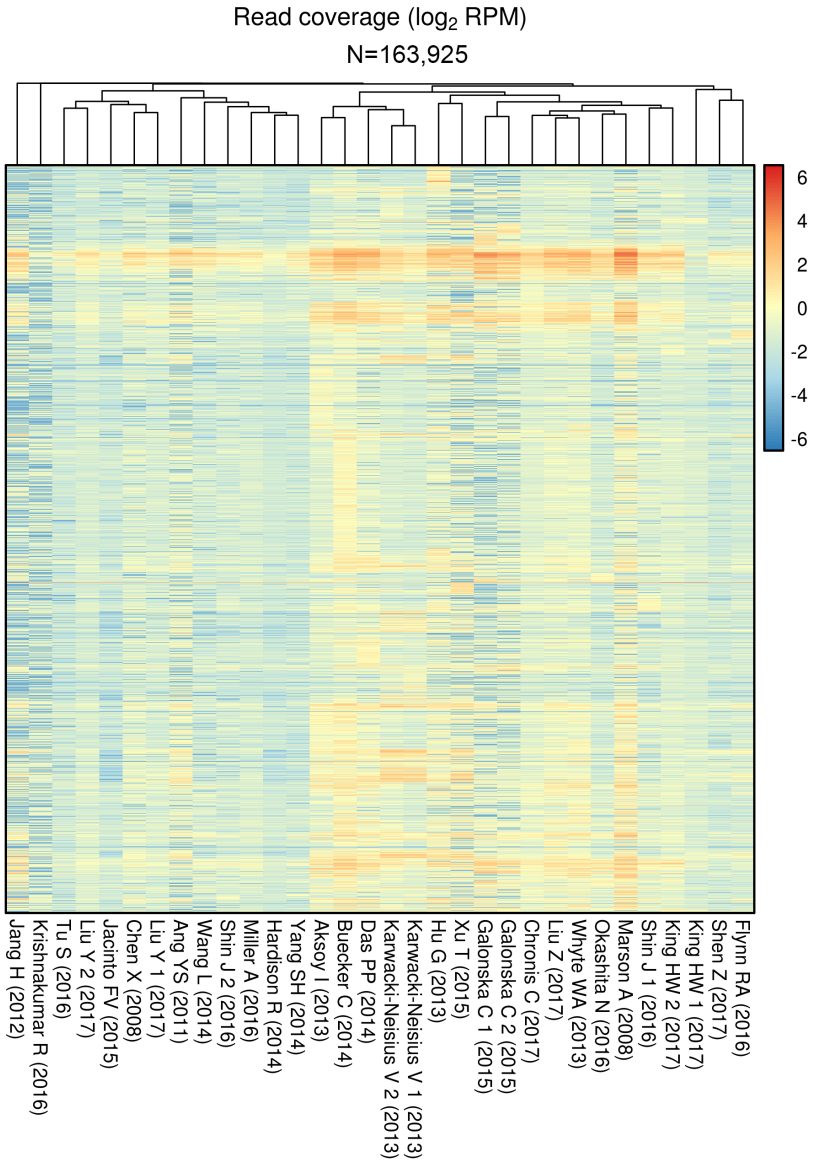

FIGURE 4.13. Peak occupancy and read coverage from published Oct4 ESC ChIP-seq experiments.

Sites which are bound or unbound in each ChIP-seq experiment are coloured by the input-subtracted read coverage. The heatmap shows that conserved sites exhibit high read coverage, which indicates strong binding.

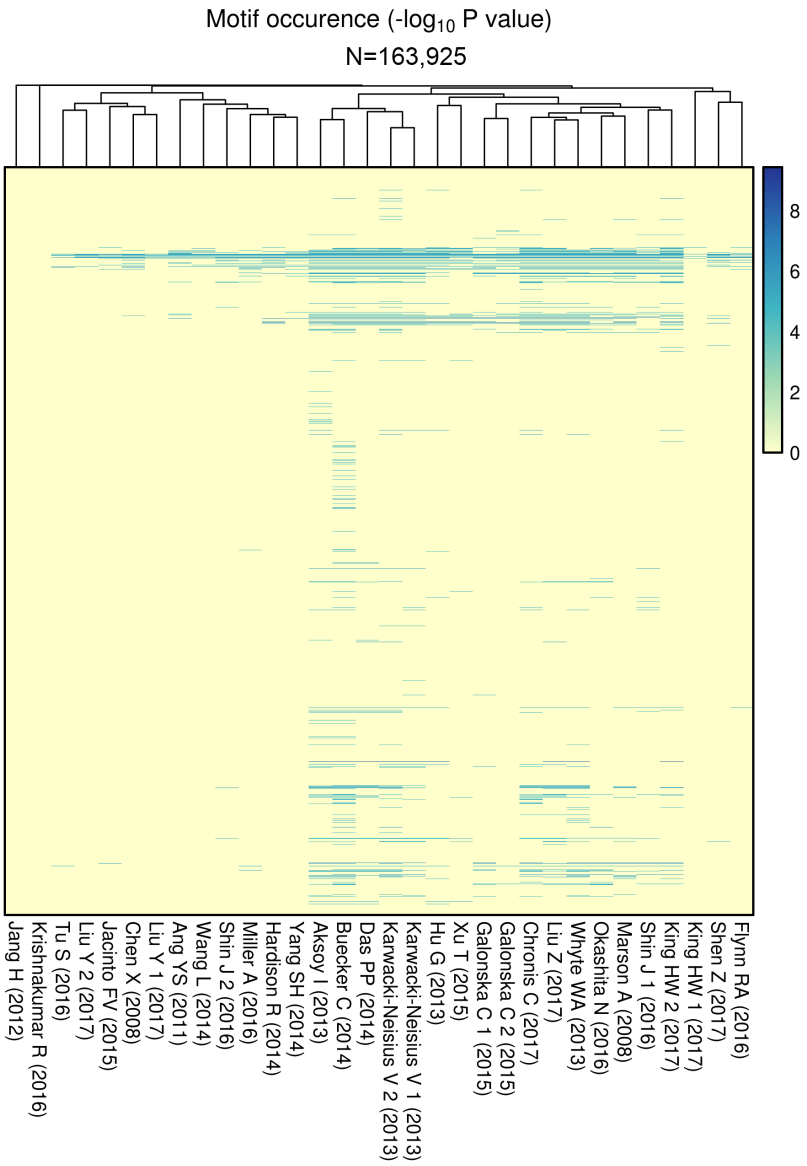

FIGURE 4.14. Peak occupancy and motif enrichment from published Oct4 ESC ChIP-seq experiments.

Sites which are bound or unbound in each ChIP-seq experiment are coloured by the likelihood of Oct4 motif enrichment. The heatmap shows that conserved sites exhibit higher Oct4 motif enrichment than un-conserved sites.

For the comparison with DamID-seq it was necessary to filter the data to obtain a comprehensive yet reliable set of Oct4 binding sites. An unsupervised filtering of the data was performed by selecting only peaks in greater than a chosen number of experiments. This number was determined by measuring how many peaks were in greater than 1 to 32 experiments then identifying at which number the rate of change started to plateau (see Figure 4.15). Although there was no obvious elbow, a distinct rate of change was visible from five to seven experiments. The optimal number of experiments chosen therefore was six (median value) , leaving approximately 34,891 peaks for downstream comparison with DamID-seq data (see Figure 4.16). Surprisingly, less than 5,000 peaks were present in more than 16 experiments which suggested either that at least half of experiments were of particularly low quality - which was hard to assess given that all of the usual quality metrics were not informative - or that there is a substantial amount of inherent variability in the ChIP-seq experimental procedure. It could also be argued that given 79% of sites are not reproducible across the experiments, then much of the signal underlying these sites may simply be random noise. However, this is unlikely given that sites are identified by calling significant peaks against an input sample representing the chromatin background (i.e. random noise). Regardless, the genes which were associated with the conserved set of peaks (by nearest TSS to the centre of the peak region) were appropriately enriched for multiple biological processes and pathways related to pluripotency and the embryonic stem cell niche (see Figure 4.17).

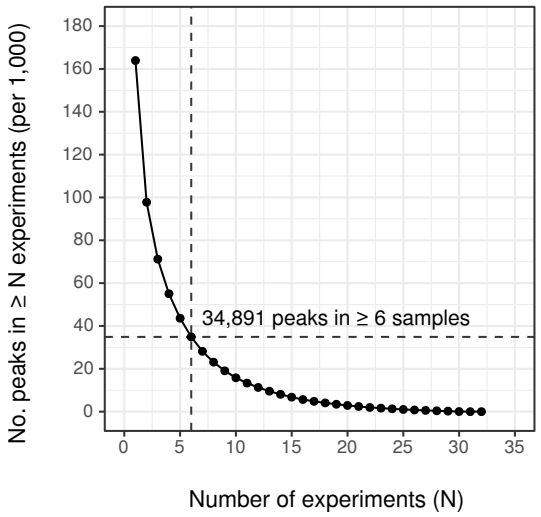

FIGURE 4.15. Filtering of peaks from published Oct4 ESC ChIP-seq experiments.

A comprehensive yet reliable set of Oct4 ESC ChIP-seq peaks was selected by first counting how many peaks were in more than N experiments. The number of peaks in more than 1 to 32 experiments was then plotted and the value of N which occurred before the number of peaks began to plateau was used.

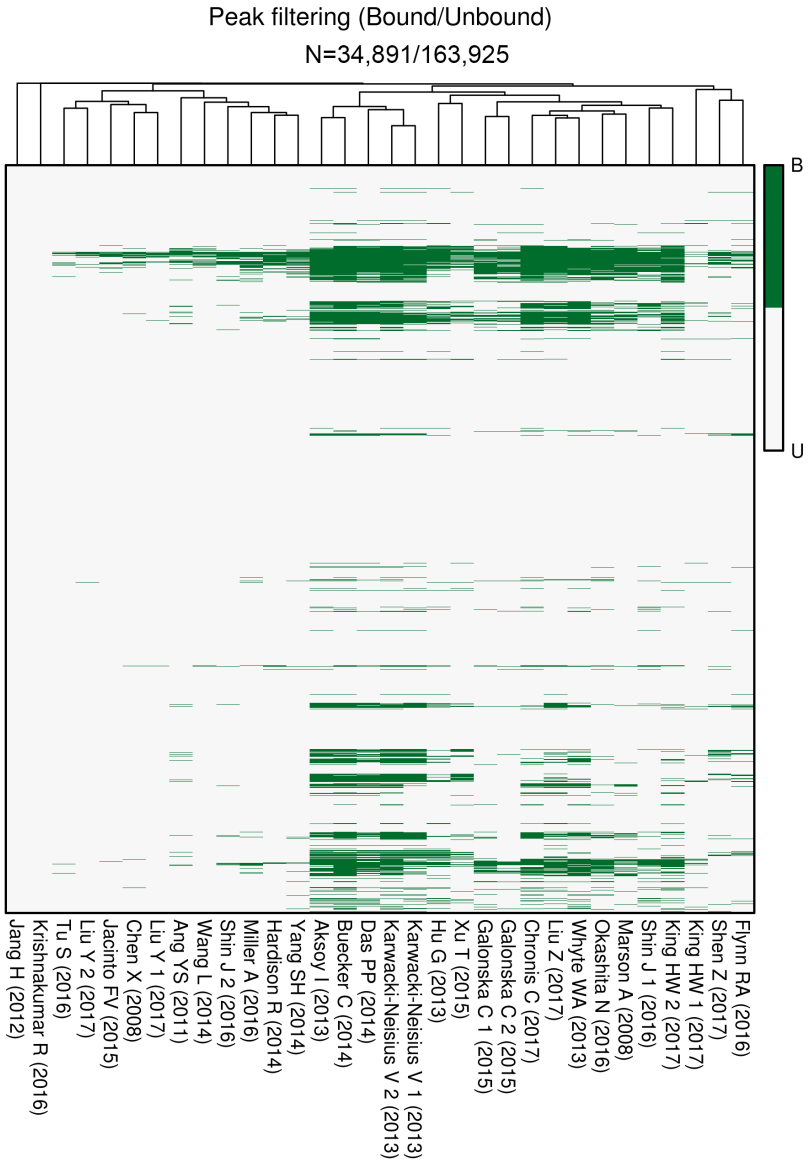

FIGURE 4.16. Filtered peak occupancy from published Oct4 ESC ChIP-seq experiments.

Sites which are bound or unbound in each ChIP-seq experiment after filtering for reproducible peaks are coloured green and grey, respectively. Around 79% of sites were filtered, leaving just 34,891 peaks for downstream comparison.

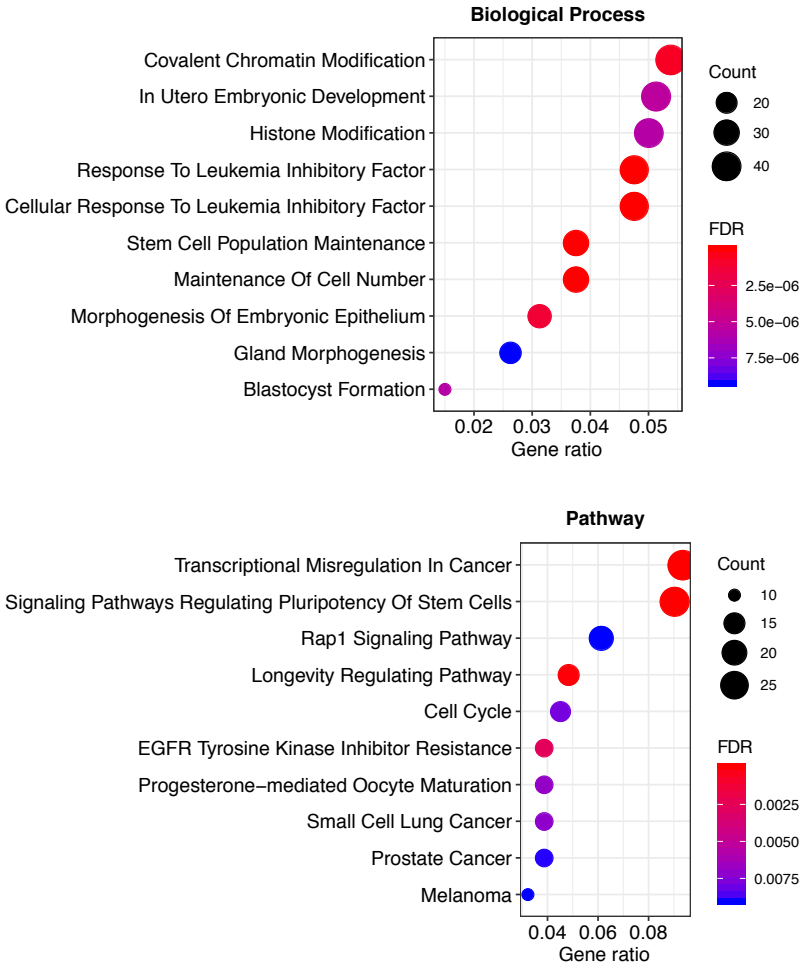

FIGURE 4.17. Gene ontology and molecular pathway analysis on conserved Oct4 ESC ChIP-seq peaks.

Gene ontology and molecular pathway analysis was used to profile the binding sites of the conserved Oct4 ESC ChIP-seq peaks. The peaks were enriched for categories related to pluripotency and the stem cell niche, which were indicative of the known function of Oct4 in embryonic stem cells.

After filtering, the similarity between experiments was measured again by calculating Spearman's rank correlation coefficient on the number of aligned reads within the 34,891 peak regions along the genome. As expected, the range of correlation coefficients increased up to 0.80, however on average they were still only 0.26 correlated

(see Figure 4.18). Similar to past observations, the experiments clustered into groups but neither of these could be explained by cross-referencing with either quality metrics or experimental variables. Considering that many experiments were on average only 0.21 correlated and that one cluster of experiments negatively correlated with the other cluster, this suggests that a particular fraction of Oct4 binding sites exhibit great variability between experiments possibly because they are bound transiently. The transience of transcription factor binding is well known, given single-molecule tracking and photobleaching studies have repeatedly shown that the time a factor spends residing at its binding site lasts for only a few seconds (Swift and Coruzzi, 2017). For example, Chen and colleagues show through live-cell single-molecule imaging experiments that Oct4 and Sox2 spend 98% of their search time sampling non-specific sites in the nucleus of ES cells before acquiring a cognate binding site (Chen et al., 2014). Taking into account the molecular concentration of both factors, they further show that a single Oct4-Sox2 site is sampled roughly every 24 seconds and the residence time is usually between 12 and 16 seconds. These dynamics are consistent with the sliding and sampling model for transcription factors which has previously been reported, whereby factors alternate between diffusing and non-specifically sliding along naked DNA in search for a cognate binding site (Hammar et al., 2012). This model may explain the variability in ChIP-seq experiments given cross-linking produces a snapshot of binding at the moment of fixation, capturing some fraction of specific binding and non-specific sliding. Filtering sites by some measure of binding strength, such as fraction of reads in peaks, is also unlikely to select specific binding as residence time appears to be unrelated to the strength of binding (Swift and Coruzzi, 2017). To uncover the reason behind the two clusters of ChIP-seq experiments, additional information such as chromatin accessibility and conformation plus gene expression from the same studies (which unfortunately is unavailable in almost all cases) would

have to be integrated to better define these high variability binding sites as either genuine regulatory elements with biological influence or simply non-specific yet strongly bound interactions.

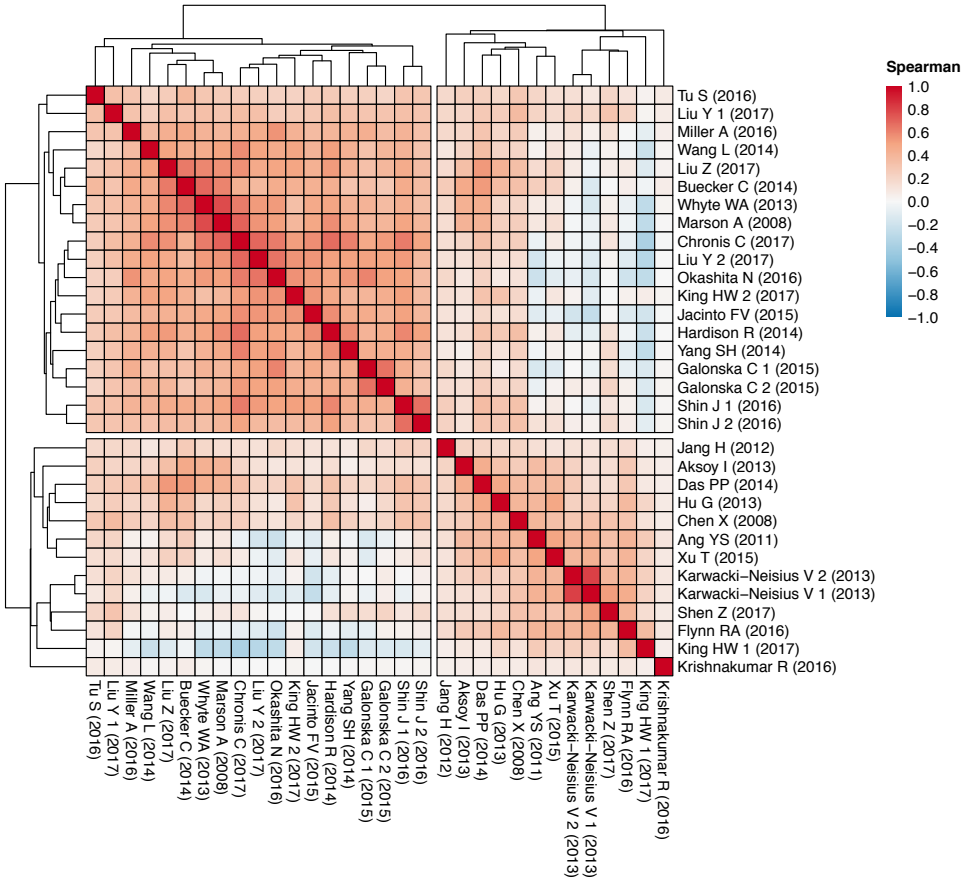

FIGURE 4.18. Spearman correlation heatmap of Oct4 ESC ChIP-seq experiments.

Spearman's rank correlation coefficient was used to measure the similarity in read coverage inside peak regions between experiments. The coefficients ranged between -0.30 and 0.80 but on average experiments were only 0.26 correlated.

Overall, these results were surprising for a number of reasons: none of the commonly used quality metrics were able to explain the disparity in the number of peak calls; the correlation between experiments was consistently low even though they were supposed to be measuring the same biological phenomenon, and many of the peaks called

were unique to only a handful of experiments which also could not be explained by any experimental variables reported. Despite these concerns, a comprehensive yet reproducible set of 34,891 binding sites were identified which exhibited high read coverage, Oct4 motif enrichment, and were implicated in relevant biological process and molecular pathways. This set of peaks was used to measure the accuracy and sensitivity of transcription factor binding site identification using Oct4 ESC DamID-seq data from a range of cell numbers.

4.4.2 Identification of binding sites from DamID-seq data

Having generated a comprehensive yet reliable set of Oct4 binding sites from ChIP-seq data, a method to identify binding sites from DamID-seq data was then developed. The analysis of most quantitative sequencing experiments involves aligning reads to the genome, counting reads within genomic features, and comparing the means between conditions. The read counts usually undergo pre-processing to remove technical biases and make the comparison between conditions more accurate and fairer. For example, if the number of reads sequenced in one condition is twice as much as the other condition, the read counts from the former would on average be double those from the latter. Comparing the means would therefore be inaccurate because technical variation is misinterpreted as biological variation. In this case, simply dividing by the total number of reads can remove this discrepancy. There are however many other sources of technical variation which need to be considered, and a range of normalisation strategies for between and within sample comparisons have been developed for sequence count data (Aleksic, Carl, and Frye, 2014). Read counts can be normalised with respect to library size (counts per million), to the length of the feature (reads per kilobase per million and transcripts per million), and to library composition (Trimmed mean of M-values, relative log expression, and upper quartile) (Li et al., 2015). Each of these normalisations assume a statistical property about the data which

must be carefully checked to ensure the transformed counts accurately reflect the biology. Unfortunately, when novel technologies are developed it is sometimes unclear what assumptions are appropriate and instead additional evidence to help decide can be generated.

In order to evaluate appropriate normalisation strategies for DamID-seq data, a reference standard was generated using corresponding qDamID data of roughly two hundred restriction fragments (see Figure 4.19). To generate Dam/Dam-Oct4 fold change values, qDamID experiments were performed on both Dam and Dam-Oct4 expressing cells with and without DpnII digestion. For each restriction fragment, the level of methylation in the Dam and Dam-Oct4 expressing cells was calculated by measuring the difference in cycle threshold values between the DpnII digested and undigested samples. The Dam/Dam-Oct4 fold changes values for each restriction fragment were then calculated based on the ratio of the DpnII values previously calculated. The restriction fragments were chosen based on DamID-seq data to generate the largest range of fold-change values at multiple levels of methylation. The qDamID fold-change values were then compared to Oct4 ESC DamID-seq fold-change values generated after applying different normalisations to the restriction fragment read counts (see Figure 4.20). Surprisingly, the correlation plots showed very little difference between different normalisations and in some cases, for example transcripts per million (TPM) and the geometric mean implemented in DESeq2 (DE), the indicated correlation coefficient did not accurately represent the relationship observed. For example, the pattern observed in the DE panel shows almost no relationship even though the calculated correlation coefficient is similar to the other panels, which tend to show a more linear relationship as expected. This outcome can occur due to single outliers in the data, as effectively shown by Anscombe's quartet - a dataset comprising of four different sets all of which have the same descriptivie statistics but ultimately show

very different patterns when plotted. One explanation for these results is that DamID-seq data may not require a specialised normalisation, and instead simply scaling by library size is sufficient to measure differential methylation. A more convincing explanation is that the number of restriction fragments assayed (i.e. 192 out of a possible 6,684,545 in the mouse genome) was too few to demonstrate any effect size between different normalisations. Overall, the results from the qDamID data were difficult to interpret so alternative approaches to evaluate appropriate normalisation strategies were investigated.

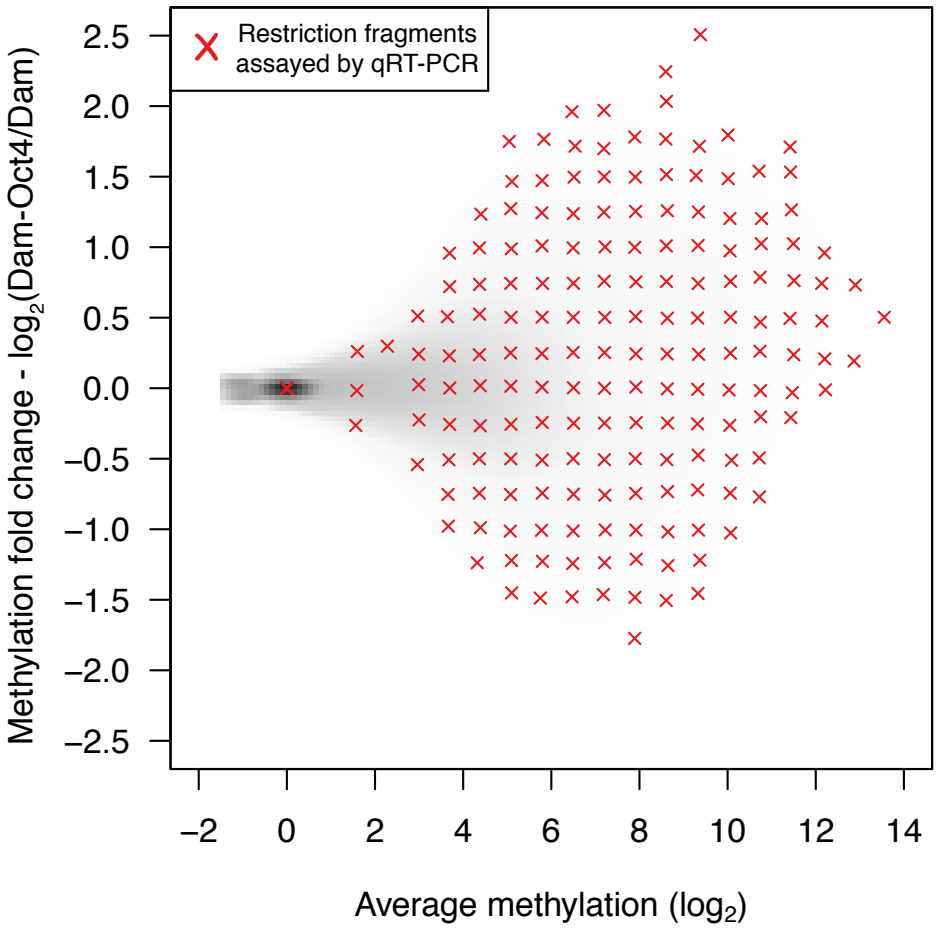

FIGURE 4.19. Mean-difference plot of restriction fragments assayed by qDamID.

Plot shows the average methylation and fold change for each restriction fragment from Oct4 ESC DamID-seq data. The red crosses indicate the location of 192 restriction fragments assayed for the qDamID experiment. Restriction fragments were chosen to achieve the largest range of fold change values across multiple levels of methylation.

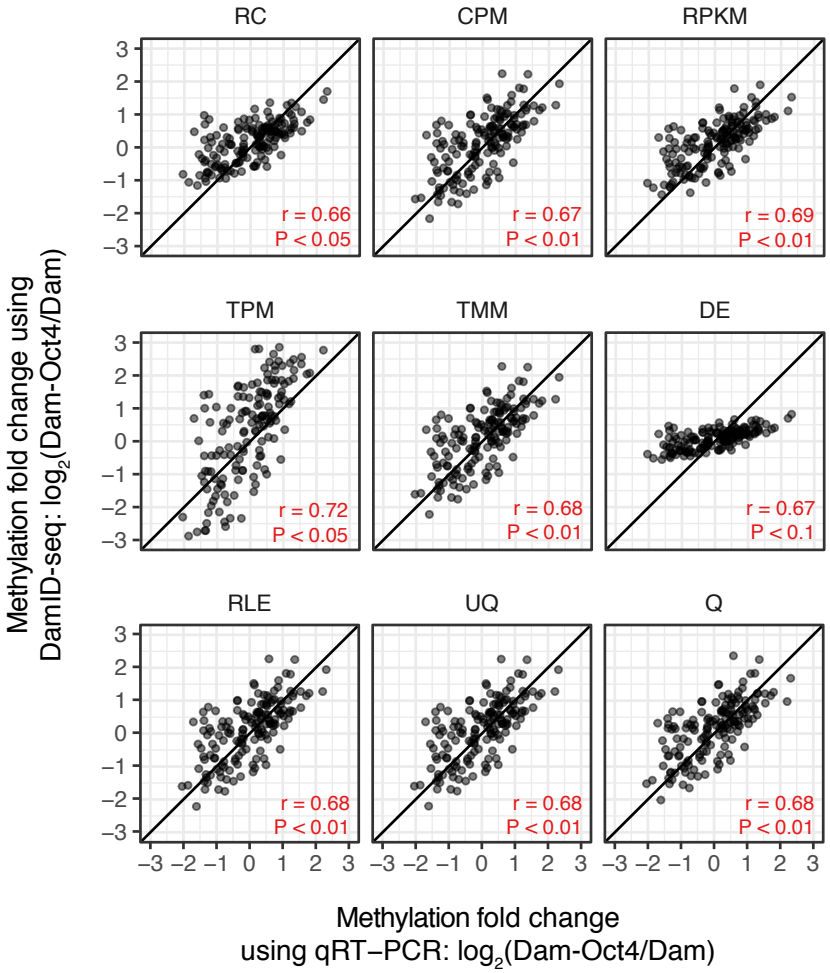

FIGURE 4.20. Correlation between qDamID and DamID-seq fold change values using different normalisations.

Graphs show the correlation between fold change values for 192 restriction fragments measured using qDamID and DamID-seq after applying different normalisations to the read counts. The normalisations evaluated are raw read count (RC), counts per million (CPM), reads per kilobase per million (RPKM), transcripts per million (TPM), trimmed mean of M-values (TMM), DESeq2 (DE), relative log expression (RLE), upper quartile (UQ), and quantile (Q). The relationship between fold change values was calculated using Pearson's correlation coefficient, which is indicated in the bottom right hand corner of each graph. A two-sided t-test was used to measure the signifiance of the relationship. The P value tests the null hypothesis that there is no correlation between DamID-seq and qDamID fold change values. The two-tailed P value answers the question, if the null hypothesis were true, what is the chance that 192 randomly picked restriction fragments would have an r value greater or less than the one reported.

Many normalisations assume that only a minority of features are significantly different between the conditions being compared. These types of normalisation are referred to as global adjustment methods because they assume that any global variability between conditions is due to technical variation and should be removed. However, if a substantial number of features are different then these normalisations would be inappropriate and have been shown to remove genuine differences between the conditions (Hicks and Irizarry, 2015). To check whether global adjustment methods were appropriate for DamID-seq data, the distribution of read counts between Dam and Dam-Oct4 samples were compared (see Figure 4.21). The density plots showed that the Dam and Dam-Oct4 samples were globally quite different, particularly at lower cell numbers. The Dam-Oct4 libraries also tended to generate a longer tail at higher read counts which was indicative of genuine binding. This pattern was also observed using a different transcription factor and cell line with Sox2 NSC DamID-seq experiments (see Figure 4.22).

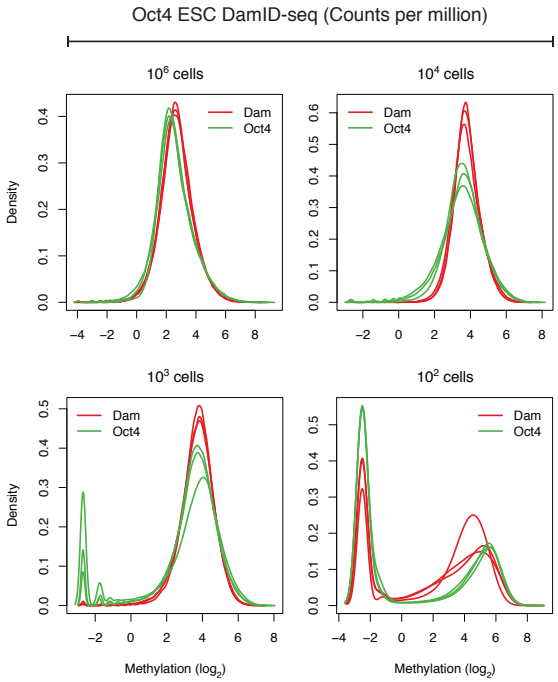

FIGURE 4.21. Distribution of read counts from Oct4 ESC DamID-seq data.

The panel of graphs show the distribution of restriction fragment read counts for Dam and Dam-Oct4 samples from DamID-seq experiments using a range of cell numbers. The read counts were normalised using $\log_2$ counts per million. The distributions show global differences between the Dam and Dam-Oct4 samples which indicate that global adjustment methods are inappropriate for normalisation.

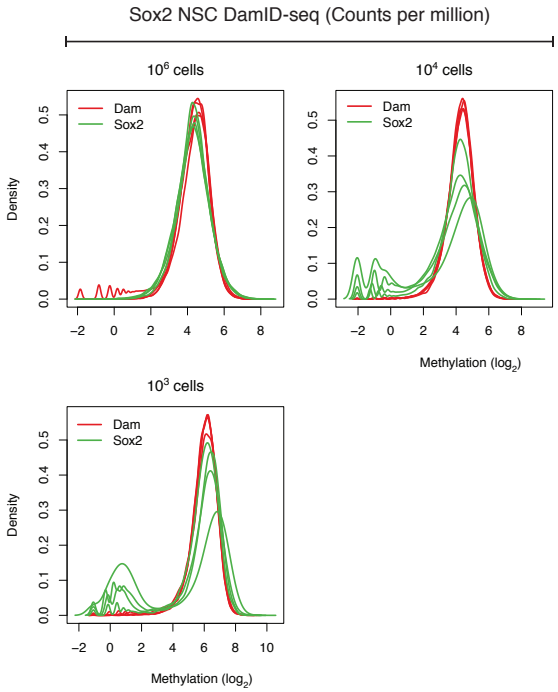

FIGURE 4.22. Distribution of read counts from Sox2 NSC DamID-seq data.

The panel of graphs show the distribution of restriction fragment read counts for Dam and Dam-Oct4 samples from multiple DamID-seq experiments using a range of cell numbers. The read counts were normalised using $\log_2$ counts per million. The distributions show global differences between the Dam and Dam-Oct4 samples which indicate that global adjustment methods are inappropriate for normalisation.

To confirm these observations a statistical test called Quantro used to detect significantly global differences in the read count distributions between the Dam and Dam-Oct4 samples (see Figure 4.23). The test measures whether the median of the distributions are different across groups, then permutes the data to assess how likely by chance this difference would have occurred (Hicks and Irizarry, 2015). The test statistics were significant using 1,000 permutations which indicated that there were statistically significant global differences in the distribution of read counts between the Dam and Dam-Oct4 samples, therefore global adjustment methods should be avoided. This result was also repeated using the Sox2 NSC DamID-seq data 4.24.

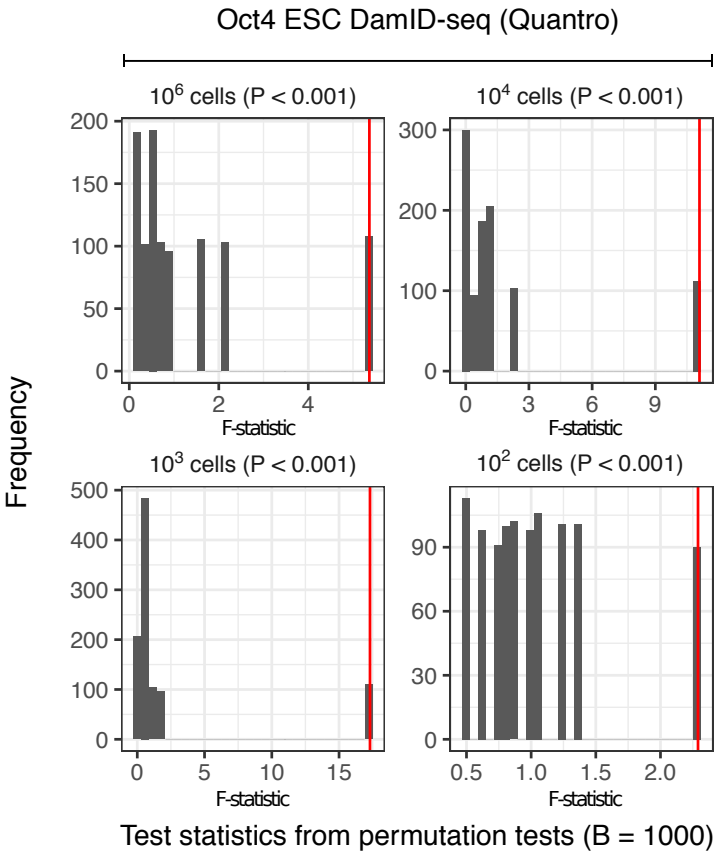

FIGURE 4.23. Quantro test statistics for Oct4 ESC DamID-seq data.

The test statistic Quantro was used to test for global differences between and within the distributions of Dam and Dam-Oct4 read counts. Each graph contains a histogram of the null test statistics from sample permutations (B = 1,000). The red line is the observed test statistic.

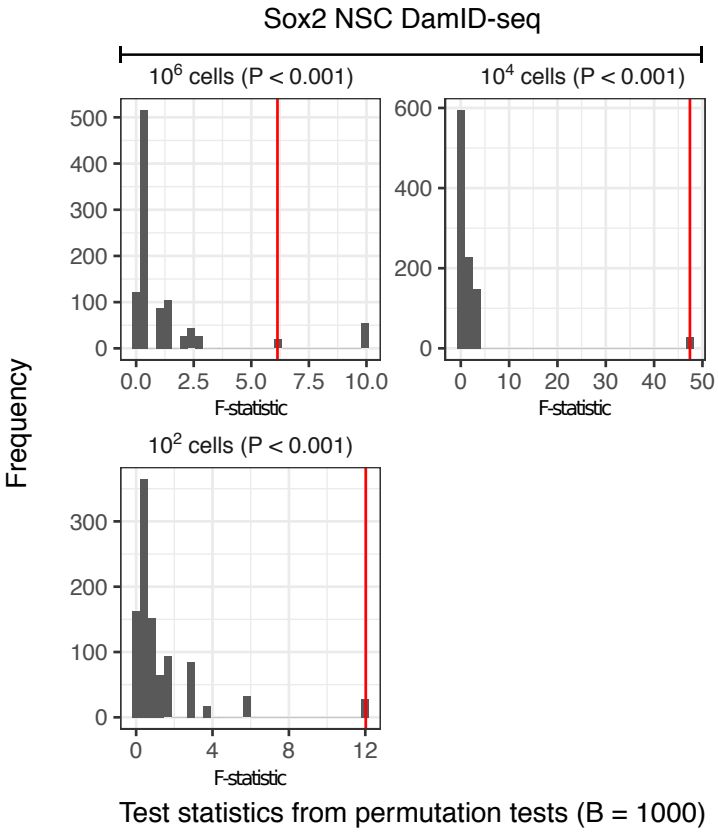

FIGURE 4.24. Quantro test statistics for Sox2 NSC DamID-seq data.

The test statistic Quantro was used to test for global differences between and within the distributions of Dam and Dam-Sox2 read counts. Each graph contains a histogram of the null test statistics from sample permutations (B = 1,000). The red line is the observed test statistic.

A recently proposed method called smooth quantile normalisation – a generalization of quantile normalisation - was specifically developed to address this type of scenario (Hicks et al., 2018). It normalises the data so that samples within a condition have the same distribution, but the distribution between conditions can still globally differ. To examine the effect of this normalisation on the DamID-seq data, the distribution of the normalised read counts were examined (see Figure 4.25). As expected the global differences between the Dam and Dam-Oct4 samples were retained and in some cases became more pronounced. At moderate to higher levels of methylation, the read count

distributions from the Dam and Dam-Oct4 samples were perfectly overlapping within each condition, which suggested all technical variation had been removed. The read counts at the lower levels of methylation however became more variable, but this was of less concern because low counts are already variable due to random sampling (Anders and Huber, 2010). Again, the same effects were observed in the Sox2 NSC DamID-seq data (see Figure 4.26). The impact of normalisation was then visualised on a genome-wide scale by plotting raw and processed coverage tracks at restriction fragment resolution (see Figures 4.27 and 4.28). The raw coverage tracks (normalised using reads per million to account for sequencing depth) show large variation between fragments due to differences in length and GC content. However, the processed coverage tracks (using conditional quantile modelling to remove fragment biases and smooth quantile normalisation to retain global differences in methylation) exhibit a flatter distribution showing that this variation is greatly reduced. This reduction allows any downstream peak regions to be correctly ranked because their coverage is no longer associated with factors unrelated to the location of the DNA-binding site of the Dam-fusion protein. Additionally, the processed coverage tracks show less variation between replicate samples allowing previously highly variable restriction fragments to be identified as significantly differentially methylated regions.

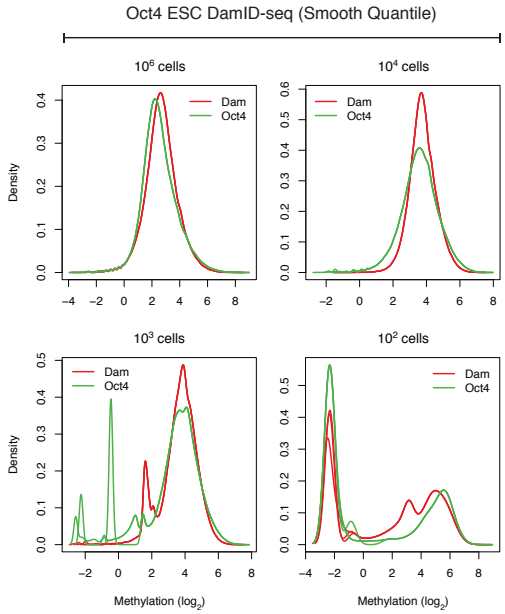

FIGURE 4.25. Distribution of smooth quantile normalised read counts from Oct4 ESC DamID-seq data.

The graphs show the distribution of restriction fragment read counts for Dam and Dam-Oct4 samples from multiple DamID-seq experiments using a range of cell numbers. The read counts were normalised using smooth quantile normalisation. The distributions show that technical variation between replicates is removed, but global differences between the Dam and Dam-Oct4 samples are retained.

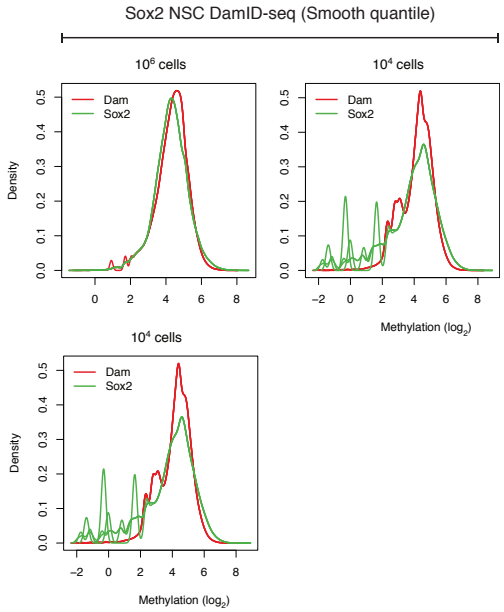

FIGURE 4.26. Distribution of smooth quantile normalised read counts from Sox2 NSC DamID-seq data.

The graphs show the distribution of restriction fragment read counts for Dam and Dam-Sox2 samples from multiple DamID-seq experiments using a range of cell numbers. The read counts were normalised using smooth quantile normalisation. The distributions show that technical variation between replicates is removed, but global differences between the Dam and Dam-Sox2 samples are retained.

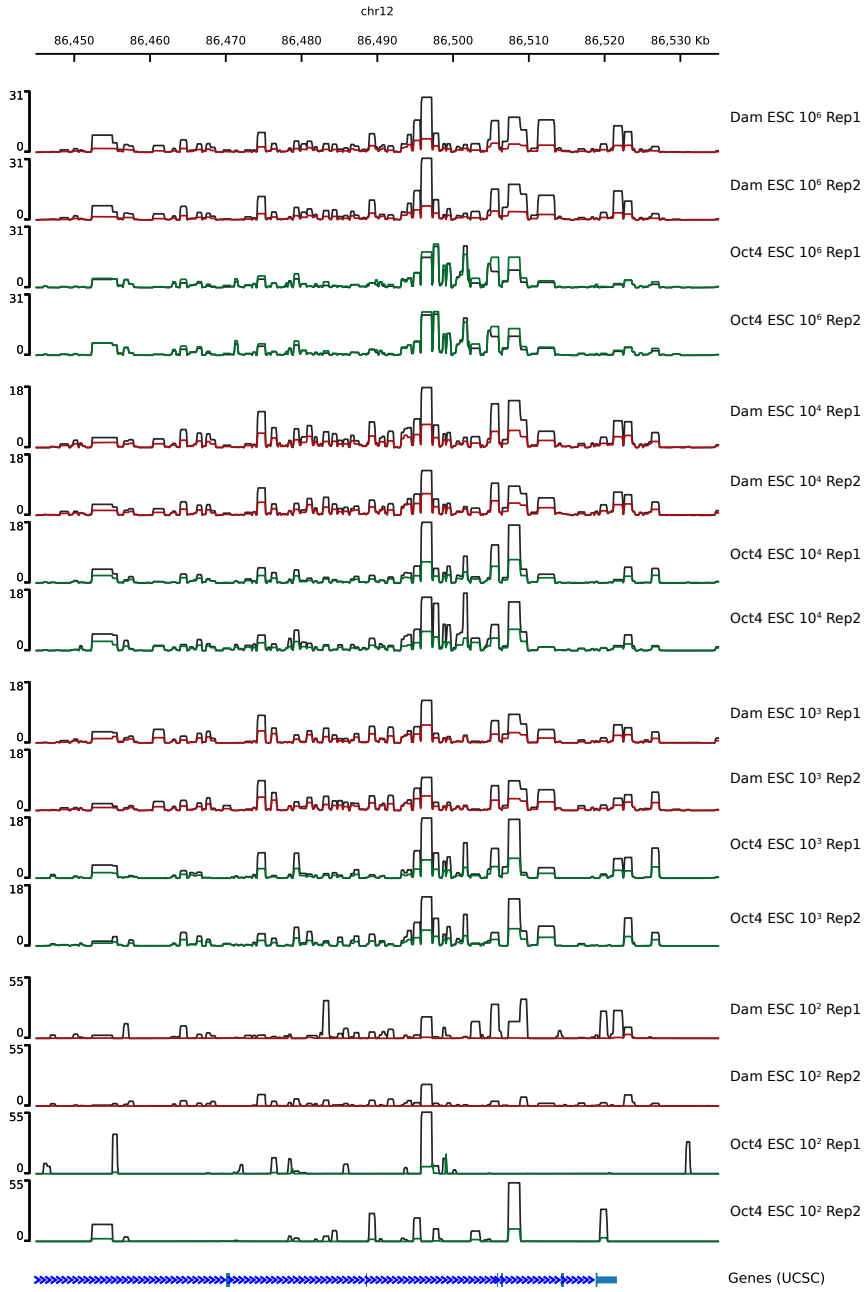

FIGURE 4.27. Genomic snapshot of Oct4 ESC DamID-seq data before and after normalisation.

Genome browser tracks show Oct4 ESC DamID-seq data before and after normalisation: The 'before' tracks (coloured in black) were transformed using a simple reads per million normalisation (RPM) to account for differences in library size. The 'after' tracks were first transformed using conditional quantile modelling (CQ) to remove fragment length and GC content biases, and then transformed using smooth quantile normalisation (SQ) to retain differences in the global methylation patterns between the Dam-Oct4 (coloured in green) and Dam (coloured in red) samples.

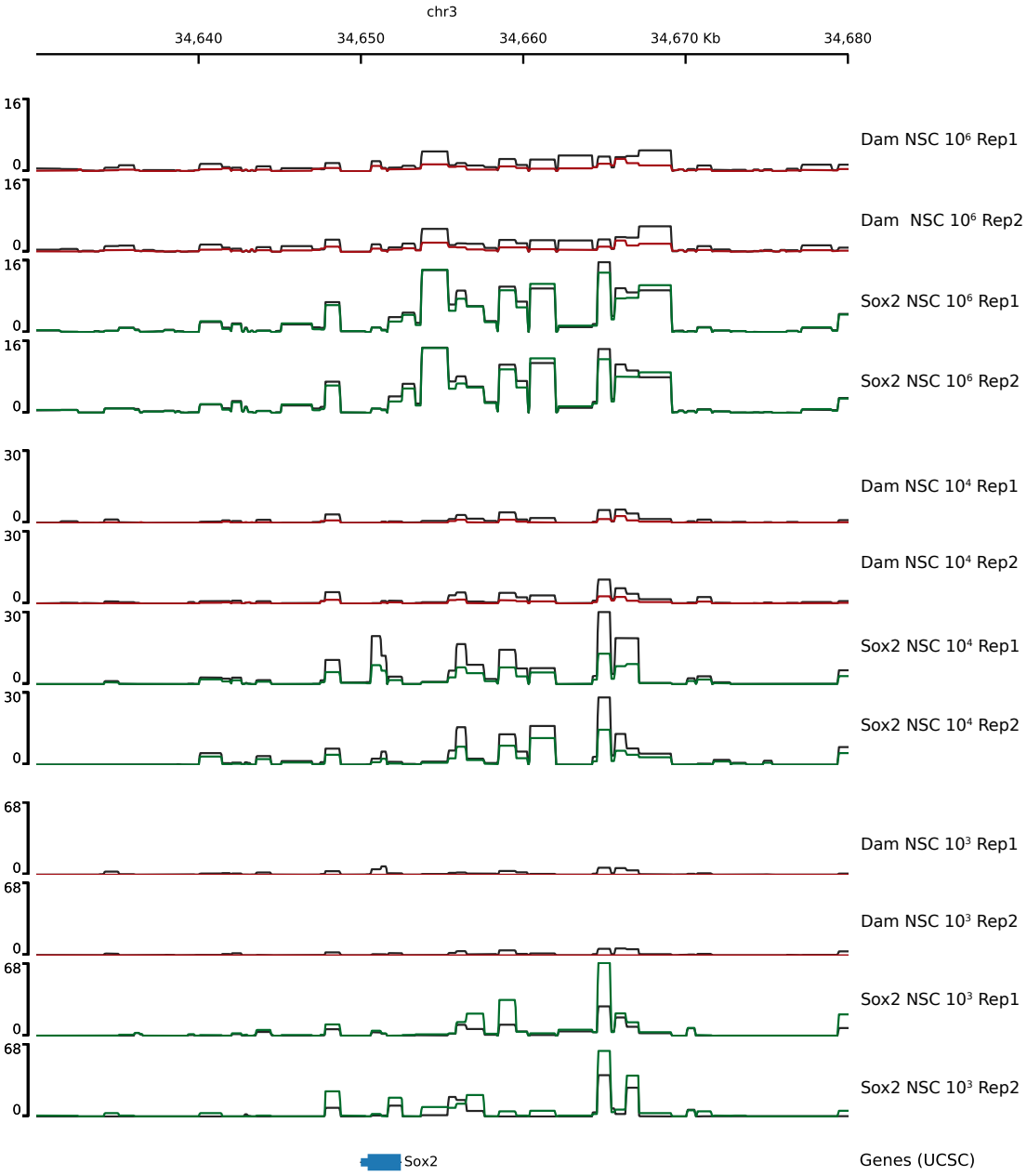

FIGURE 4.28. Genomic snapshot of Sox2 NSC DamID-seq data before and after normalisation.

Genome browser tracks show Sox2 NSC DamID-seq data before and after normalisation: The 'before' tracks (coloured in black) are transformed using a simple reads per million normalisation (RPM) to account for differences in library size. The 'after' tracks are first transformed using conditional quantile modelling (CQ) to remove fragment length and GC content biases, and then transformed using smooth quantile normalisation (SQ) to retain differences in the global methylation patterns between the Dam-Sox2 (coloured in green) and Dam (coloured in red) samples.

Having established a suitable normalisation strategy, differential methylation analysis between the Dam and Dam-fusion samples using linear modelling and empirical Bayes methods from the limma package was performed (Ritchie et al., 2015). An important feature of these methods are that they share information between features in order to better model the mean-variance trend in the sequence count data. The variance across all samples is typically used to estimate this trend because it is assumed that dispersion is similar between conditions. To check whether this assumption holds for DamID-seq data the similarity between Dam and Dam-fusion replicates in both Oct4 ESC and Sox2 NSC DamID-seq data was examined (see Figures 4.29 and 4.30). Multidimensional scaling plots revealed that the Dam-fusion samples were consistently more variable than the Dam samples, which indicated that dispersion should be modelled for each group independently. This was achieved by proving group information to the array weights function in the limma package. After differential methylation analysis, restriction fragments within 1 kb of each other were then stitched together and a Fisher's combined P value and multiple testing corrected P value were calculated (Lun and Smyth, 2016). Only regions with a significant positive fold change above zero (i.e. higher in the Dam-fusion samples) were retained.

The observation that Dam and Dam-fusion proteins exhibit significantly different methylation distributions reflects the fact that Dam methylation is observed across the entirety of the genome whilst Dam-fusion methylation is primarily at the DNA-binding sites of the target protein. These differences are in agreement with the theory behind the methodology of the DamID technology, whereby we expect the Dam-fusion protein to be localised to the target sites and less often at non-target sites where Dam is regularly situated. It is therefore very important that any normalisation procedure used retains this key biological feature of the sequencing data to ensure peak calling accuracy. Surprisingly, the DamID-seq analysis pipeline produced by Marshall and

colleagues (introduced in Subsection 2.2.7) attempts to normalise away this inherent biological difference by transforming the data so that the log2 methylation ratio (Dam-fusion/Dam) across the majority of restriction fragments in the genome is exactly zero (i.e. the level of methylation across non-target sites is forced to be exactly the same in both the Dam and Dam-fusion proteins). Specifically, they calculate the kernel density distribution of log2 methylation ratios across the 40% to 90% deciles (the first free deciles are removed because they generate inconsistent normalisation factors and the last decile is removed because it is thought to contain genuine binding) and then calculate a numerical constant to transform the data such that the point of maximum kernel density of the log2 methylation ratios (Dam-fusion/Dam) is exactly zero. Theoretically, this procedure is unsuitable for DamID-seq data for a few reasons: First, it is likely to normalise away binding which does not occur in the highest decile by methylation. This is also problematic because a hard threshold does not accommodate proteins with different DNA-binding dynamics, such as lamin-binding and transcription factors, which markedly differ in their size and distribution along the genome. Second, by over-estimating the level of background noise (i.e. assuming everything below the last decile does not contain binding) the normalization factors will exaggerate the Dam-fusion signal which occurs in the last decile by methylation. While this may produce a cleaner visualization over genuine DNA-binding sites, it also will create peaks over non-target sites which are highly methylated in both the Dam and Dam-fusion proteins. Third, it is unclear how such a normalization procedure would accurately take into account information from multiple replicates given the ratios are calculated based on a pair of single replicate samples and not multiple replicate samples. The result of this single replicate normalization factor would be that the coverage values would not be comparable between multiple replicates processed separately by the pipeline, instead all of sequencing data would have to be merged first which removes any advantage of performing replicate experiments to measure

variability. Additionally, as both technical and biological variability increases with fewer cells it becomes more important to normalise and model the replicates together to ensure the false-positive rate of peak calling is controlled.

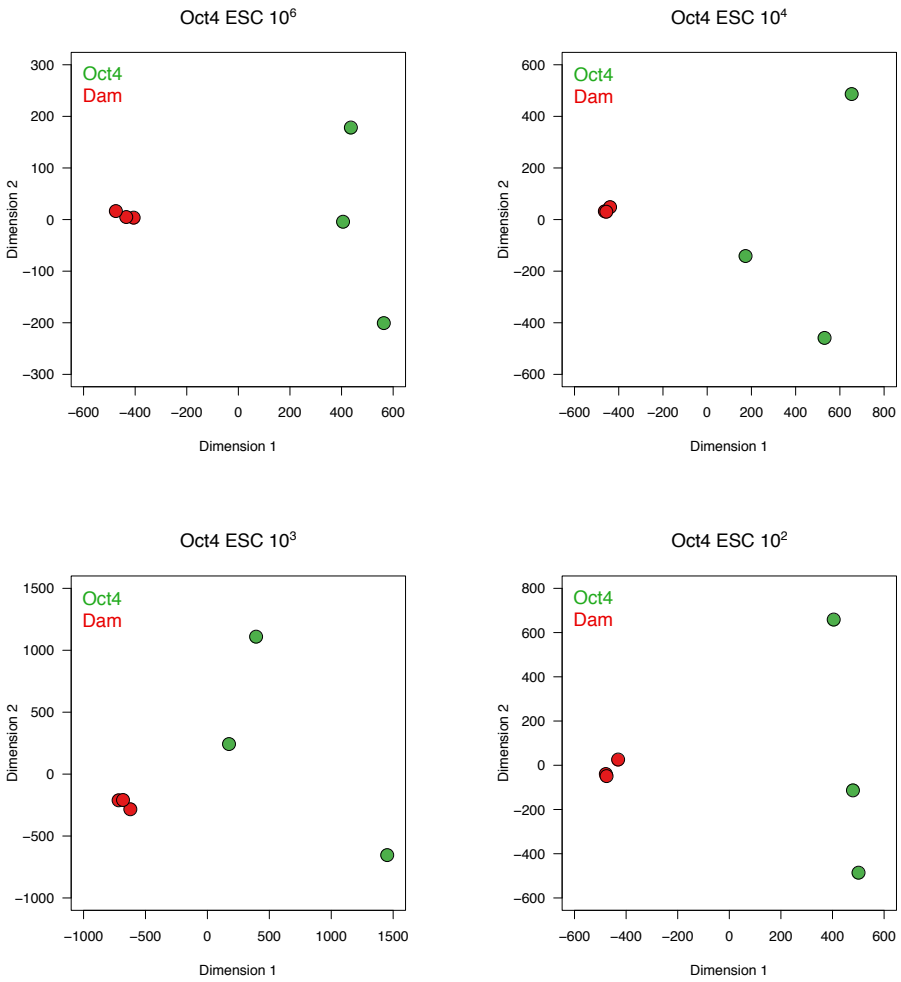

FIGURE 4.29. Multidimensional scaling analysis of Oct4 ESC DamID-seq data.

Multidimensional scaling was used to evaluate the distance between and within Dam and Dam-Oct4 sample groups. The plots show that the Dam-Oct4 samples are much more variable then the Dam samples, even at lower cell numbers. This difference is accounted for in the linear modelling strategy by providing group information to the array weights function in the limma package (Ritchie et al., 2015).

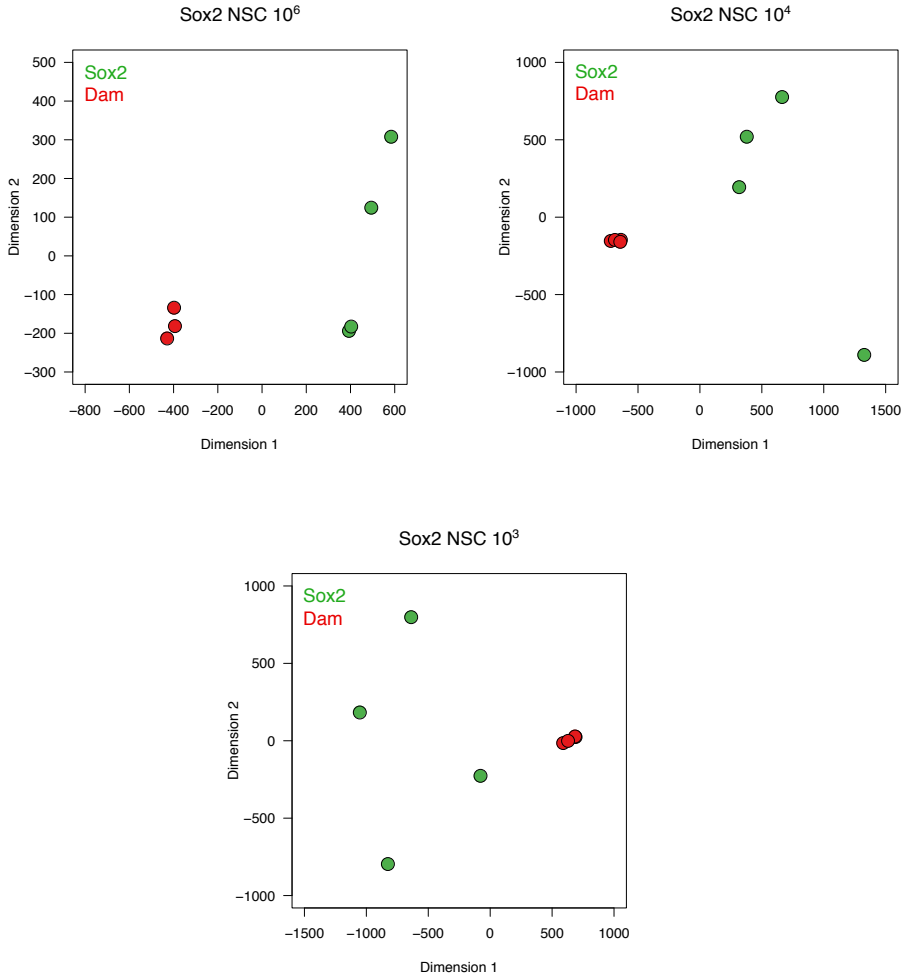

FIGURE 4.30. Multidimensional scaling analysis of Sox2 NSC DamID-seq data.

Multidimensional scaling was used to evaluate the distance between and within Dam and Dam-Sox2 sample groups. The plots show that the Dam-Oct4 samples are much more variable then the Dam samples, even at lower cell numbers. This difference is accounted for in the linear modelling strategy by providing group information to the array weights function in the limma package (Ritchie et al., 2015).

4.4.3 Comparison of binding sites from DamID-seq and ChIP-seq data

In order to evaluate the DamID-seq peak calling strategy, a comparison of Oct4 ESC binding sites identified from DamID-seq data and the previously analysed ChIP-seq

data was performed. The DamID-seq data was generated using 10^6, 10^4, and 10^3 cells whilst the public ChIP-seq data was generated using a minimum of 10^7 cells. Three ChIP-seq experiments with the highest numbers of peaks were used for visual and quantitative comparison with the DamID-seq data. For easier comprehension, different groups of overlapping and unique peaks were enumerated: Set 1 refers to peaks in DamID-seq and both the filtered and unfiltered ChIP-seq; Set 2 refers to peaks in DamID-seq and the unfiltered ChIP-seq; Set 3 refers to peaks in the filtered ChIP-seq only; Set 4 refers to peaks in DamID-seq only; and Set 5 refers to peaks in the unfiltered ChIP-seq only.

Genome browser images showed that DamID-seq and ChIP-seq reads accumulated at similar positions and the pileup of DamID-seq reads at predicted binding sites was reassuringly higher in the Dam-Oct4 library compared to the control Dam library (see Figure 4.31). The location of the DamID-seq and ChIP-seq peak calls was also comparable despite being called using different algorithms, this suggested that the DamID-seq peak calling strategy was suitable and that further investigation of the binding sites was justified. One noticeable difference however was that DamID-seq peak calls were much larger (ranging from 10^3 to 10^4 bp) than the ChIP-seq peak calls (ranging from 10^2 to 10^3 bp) which unfortunately lowers the resolution (see Figure 4.32). This is because DamID-seq peaks are limited by the size of the restriction fragments, unlike ChIP-seq which typically uses small overlapping windows to identify enrichment (Zhang et al., 2008).

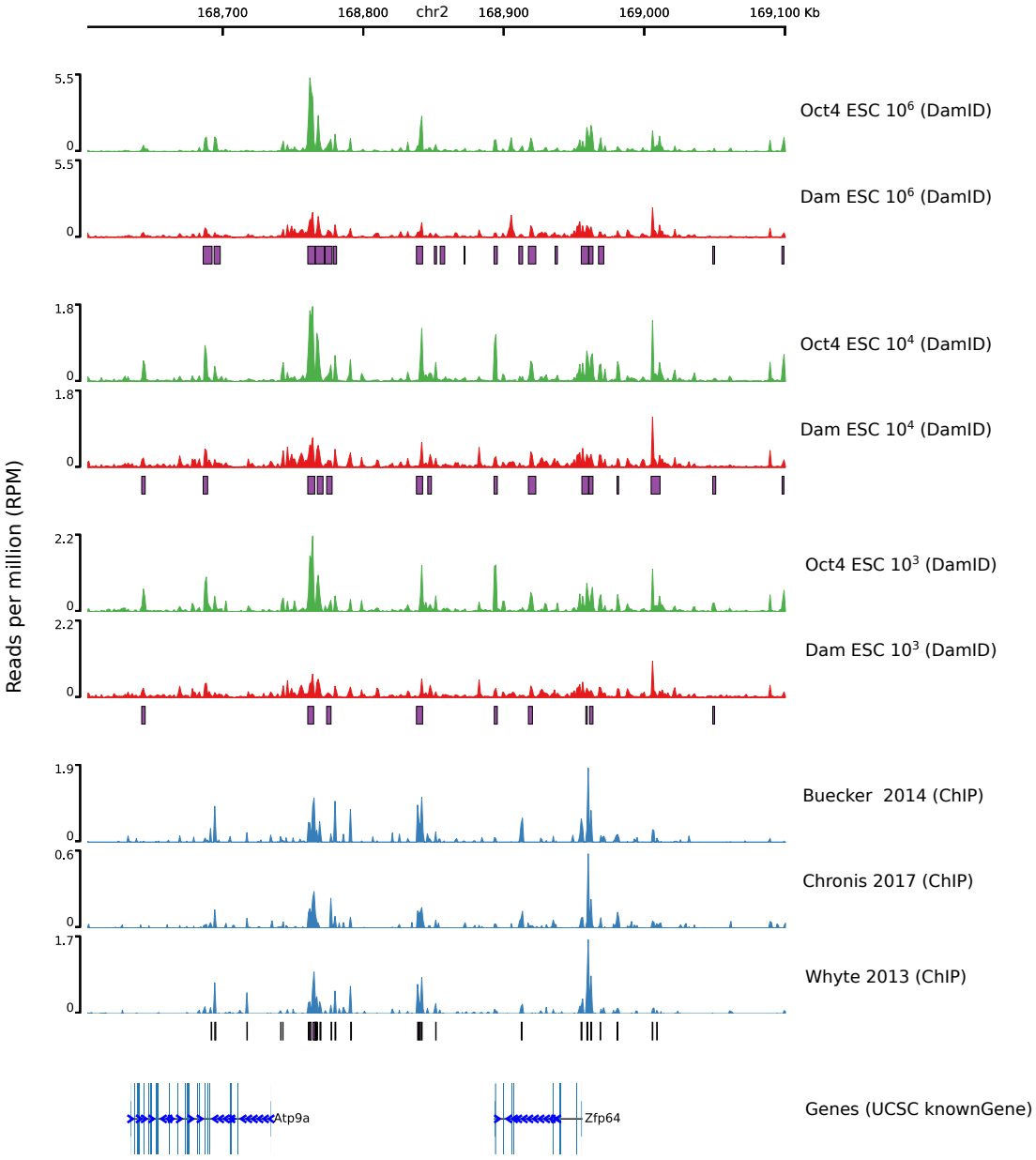

FIGURE 4.31. Genomic snapshot of Oct4 ESC DamID-seq and ChIP-seq peak calls.

Genome browser tracks show good agreement between Oct4 ESC DamID-seq and ChIP-seq read coverage and peak calls. As expected, the signal from the Dam-Oct4 samples are higher than the Dam samples at predicted transcription-factor binding sites.

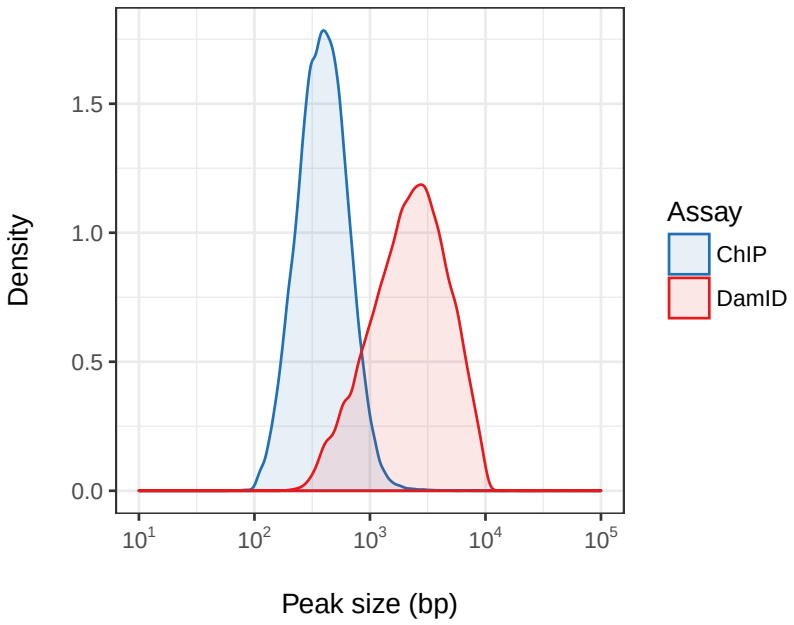

FIGURE 4.32. Distribution of peak sizes from Oct4 ESC DamID-seq and ChIP-seq data.

The distribution of peak sizes indicate that DamID-seq is a lower resolution assay than ChIP-seq. This is because DamID-seq peaks are called using restriction fragments, rather than small overlapping windows typically used in ChIP-seq algorithms (Zhang et al., 2008).

Approximately 72% of the DamID-seq peaks called using 10^6 cells overlapped with 23% of the unfiltered ChIP-seq peaks (Sets 1 and 2) identified from the published experiments analysed previously (see Figure 4.33). However, the latter percentage increased to 40% when the filtered set of ChIP-seq peaks (Set 1) was compared, which demonstrated that DamID-seq was able to identify a substantial proportion of reproducible binding sites despite using a tenth of the cells. The pileup of aligned reads also demonstrated that many of the overlapping peaks (Set 1) were strongly bound in both the DamID-seq and ChIP-seq experiments (see Figure 4.34). Nonetheless, there was a unique set of strongly bound ChIP-seq peaks (Set 3) which were not bound at all in the DamID-seq experiments that indicated a significant difference between the two assays. Histone modification and chromatin accessibility data revealed that DamID-seq

peaks which overlapped with the filtered set of ChIP-seq peaks (Set 1) were in highly accessible chromatin containing H3K4me1, H3K4me2, and H3K27ac modifications, indicative of enhancer and transcription factor binding regions (see Figure 4.35). Interestingly, peaks which were unique to ChIP-seq (Set 3) were also in highly accessible chromatin but contained H3K4me2, H3K4me3, and H3K9ac modifications. This combination of epigenetic marks suggested that DamID-seq was less able to detect binding sites within the promoters of actively transcribed genes. The annotation of peaks to genomic features additionally demonstrated that 35% of the ChIP-seq specific peaks (Set 1) were located within promoters, compared to less than 5% of the DamID-seq specific peaks (Set 4) (see Figure 4.37). One explanation for this deficiency is that Dam may be unable to methylate chromatin which is highly occupied by other DNA-binding proteins, particularly the RNA polymerase II complex. However, both ChIP-seq and DamID-seq are able to detect binding in enhancers, which are also highly occupied and often bound by RNA polymerase II complex. It has been shown in *Drosophila* that the frequency of GATC sites within promoter regions is decreased compared to the surrounding genomic regions, therefore it follows that methylation will be absent or reduced (Aughey et al., 2018). This depletion of GATC sites in promoter regions is also observed in the mouse genome (see Figure 4.36) and may explain why fewer binding sites were detected at promoters using DamID-seq data. Following peak annotation to the nearest TSS (using annotatePeaks from Homer), gene ontology and pathway analyses demonstrated that DamID-seq and ChIP-seq peaks were largely enriched for the same categories, including regulation of pluripotency, embryonic development, and stem cell population maintenance (see Figure 4.38).

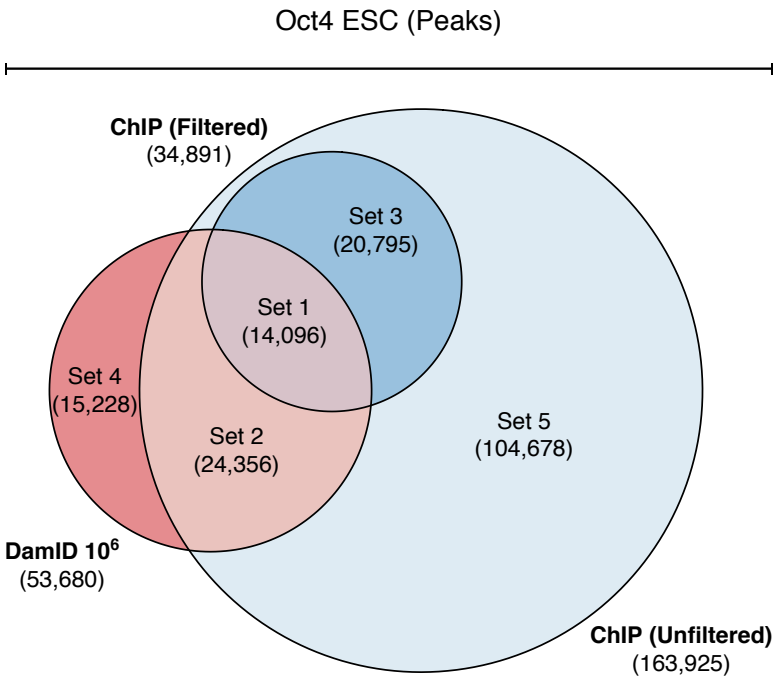

FIGURE 4.33. Comparison of Oct4 ESC peaks between DamID-seq and ChIP-seq data.

The Euler diagram represents the overlap between Oct4 ESC peaks from DamID-seq and ChIP-seq data. The filtered ChIP-seq peaks were derived from the overlap between multiple published experiments, as described previously. For easier comprehension, different groups of overlapping and unique peaks were enumerated: Set 1 refers to peaks in DamID-seq and both the filtered and unfiltered ChIP-seq; Set 2 refers to peaks in DamID-seq and the unfiltered ChIP-seq; Set 3 refers to peaks in the filtered ChIP-seq only; Set 4 refers to peaks in DamID-seq only; and Set 5 refers to peaks in the unfiltered ChIP-seq only.

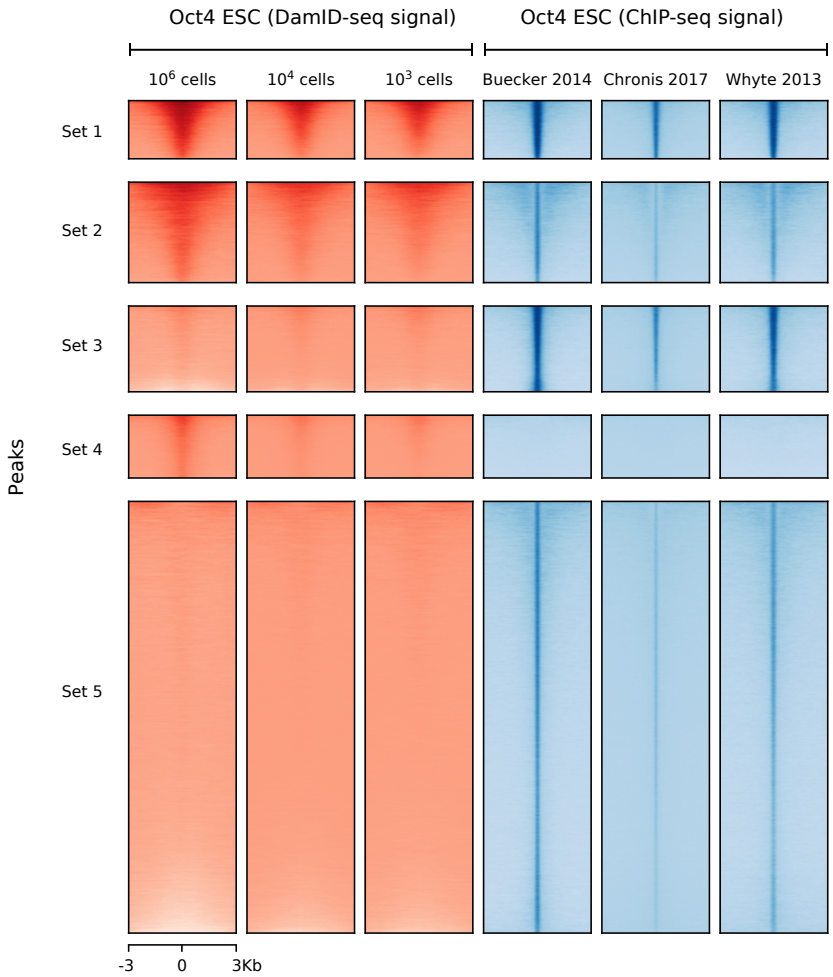

FIGURE 4.34. Read coverage at Oct4 ESC peaks from DamID-seq and ChIP-seq data.

The heatmap shows the pileup of aligned reads at overlapping and unique Oct4 ESC peaks identified from DamID-seq and ChIP-seq data. Each set of peaks relates to the sets found in the corresponding Euler diagram (see Figure 4.33).

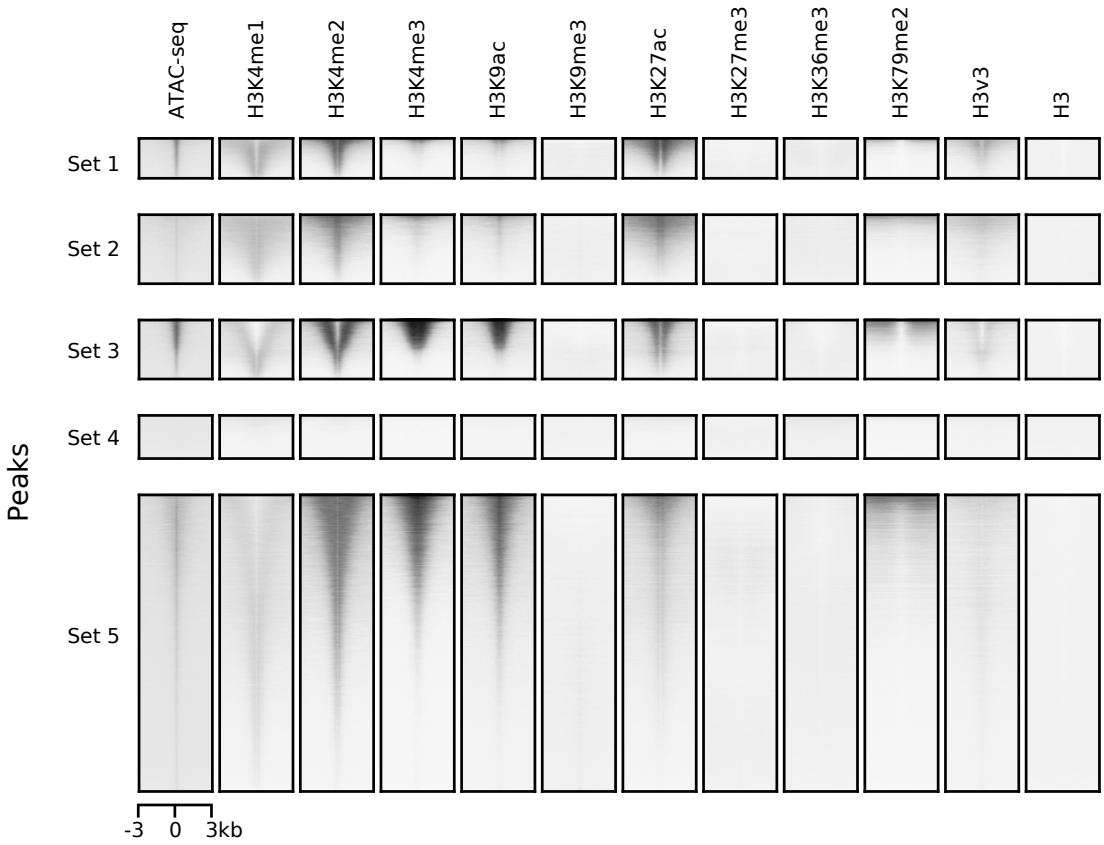

FIGURE 4.35. Chromatin accessibility and histone modification at Oct4 ESC peaks from DamID-seq and ChIP-seq data.

The heatmap shows the chromatin accessibility and histone modifications present overlapping and unique Oct4 ESC peaks from DamID-seq and ChIP-seq data. Each set of peaks relates to the sets found in the corresponding Euler diagram (see Figure 4.33).

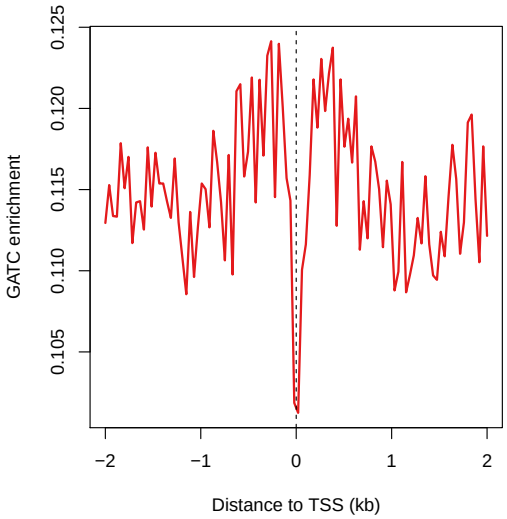

FIGURE 4.36. Fold enrichment of GATC sites at promoter regions.

Fold enrichment of GATC sites around transcriptional start sites (TSS). For each position around the TSS, fold enrichment is calculated as the number of GATC sites divided by the number of transcripts measured. Promoter regions have a depletion of GATC sequences.

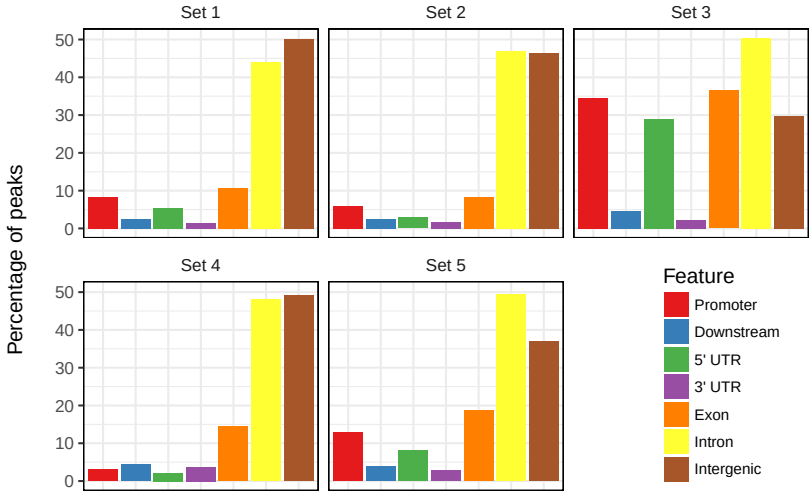

FIGURE 4.37. Annotation of overlapping and unique Oct4 ESC DamID-seq and ChIP-seq peaks.

Genomic annotation of overlapping and unique Oct4 ESC DamID-seq and ChIP-seq peaks as promoter, downstream of gene end, 5′ untranslated region, 3′ untranslated region, exon, intron, or intergenic. Each set of peaks relates to the sets found in the corresponding Euler diagram (see Figure 4.33).

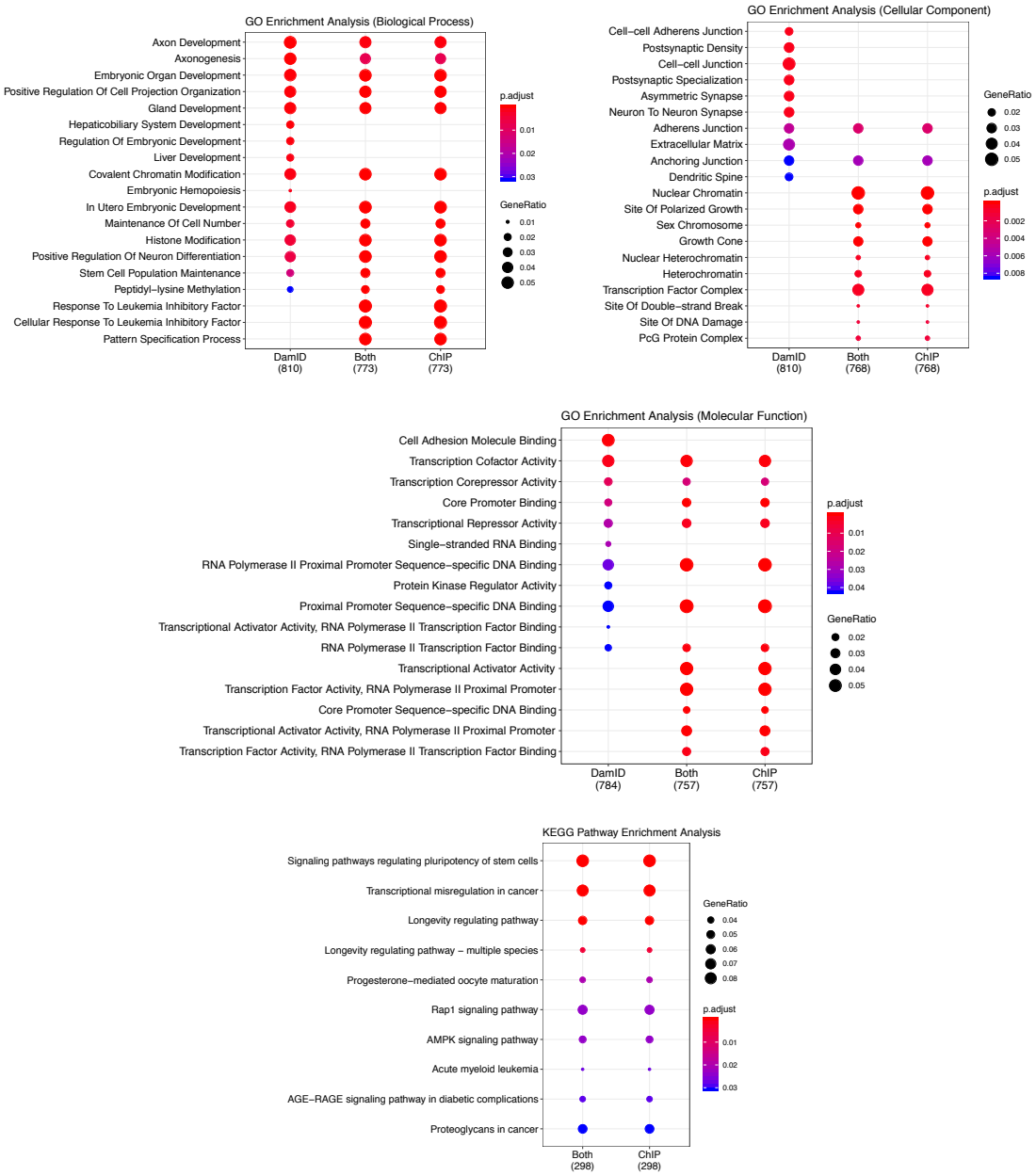

FIGURE 4.38. Ontology and pathway analysis of Oct4 ESC ChIP-seq and DamID-seq binding sites.

Gene ontology analysis of overlapping and unique Oct4 ESC ChIP-seq and DamID-seq peaks. The colour of the points in the graph reflect the enrichment significance of each category. The size of the points in the graph reflect the number of peak genes which were assigned to the relevant category.

One of the advantages of DamID-seq is the ability to generate a sufficient DNA yield from much lower cell numbers than conventional ChIP-seq experiments. Approximately 7,190 Oct4 ESC binding sites could still be detected using only 10^3 cells, which is remarkable given this number is ten thousandths that of a conventional ChIP-seq experiment (see Figure 4.39). Importantly, the peaks called using lower cell numbers formed almost an entire subset of those called using higher cell numbers and many of them contained a high number of reads which indicated that at lower cell numbers only the strongest binding sites were detected. Overall these results indicated that the DamID-seq peak calling strategy was effective and allowed the identification of binding sites from fewer cell numbers. Additionally, differences between DamID-seq and ChIP-seq peaks indicated that DamID-seq is less sensitive at promoter regions.

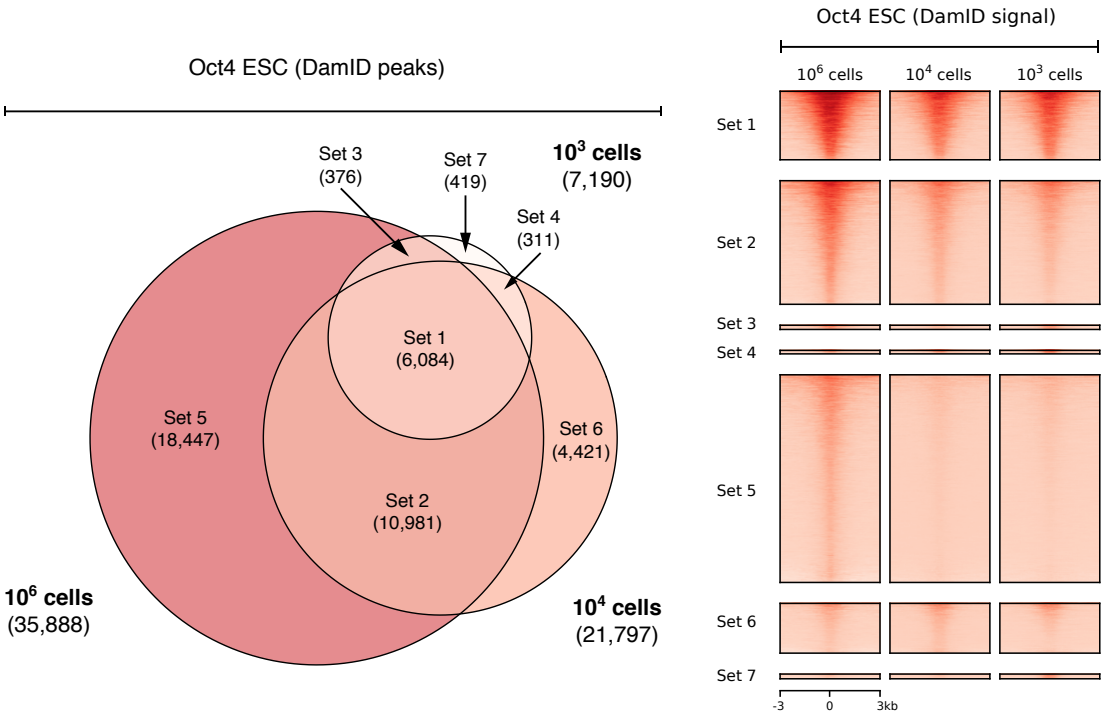

FIGURE 4.39. Comparison of Oct4 ESC peaks from low cell number DamID-seq experiments.

The Euler diagram on the left hand side represents the overlap between DamID-seq peaks from ESC experiments performed using 10^6, 10^4, and 10^3 cells. The heatmap on the right hand side displays DamID-seq read coverage at the corresponding set of peaks.

To confirm the observations about DamID-seq data and further evaluate the peak calling strategy, a comparison of Sox2 NSC binding sites identified from DamID-seq and three public ChIP-seq data sets was also performed. Importantly, this comparison used a different transcription factor and cell line so observations common to both DamID-seq analyses were reinforced. The DamID-seq data was again generated using 10^6, 10^4, and 10^3 cells whilst the public ChIP-seq data was generated using a minimum of 10^7 cells. Approximately 12,417 binding sites from a total of 46,027 sites were detected in all three ChIP-seq experiments (see Figure 4.40). Additionally, overlapping peaks were highly covered by reads which demonstrated that these sites were strongly bound and hence reproducible. Genome browser images showed a similar pileup of

DamID-seq and ChIP-seq reads at predicted binding sites, and the DamID-seq peak calls were located appropriately under regions methylated higher in the Dam-Sox2 library than the control Dam library (see Figure 4.41). Approximately 55% of DamID-seq peaks called using 10^6 cells overlapped with 28% of the ChIP-seq peaks identified from the three public experiments analysed (see Figure 4.42). This latter percentage increased to over 42% when the overlapping ChIP-seq peaks were compared, this number was very similar to the corresponding percentage from the Oct4 ESC analyses which suggested that around 60% of ChIP-seq peaks may never be detected using DamID-seq due to an unknown technical difference between the assays. One such explanation may be the existence of high-occupancy target regions in the genome which have recently been shown to generate reproducible but meaningless enrichment of proteins at highly expressed genes in ChIP-seq experiments (Wreczycka et al., 2017). Nonetheless, DamID-seq peaks which overlapped with ChIP-seq peaks exhibited high read coverage which indicated that reproducible and strong binding sites can be identified by increasing the peak calling threshold. The annotation of peaks to genomic features demonstrated that over 20% of ChIP-seq specific peaks were located at promoters, almost double the number of DamID-seq specific peaks. Additionally, both DamID-seq and ChIP-seq peaks were enriched for neuronal gene ontology categories and pathways including axon development, gliogenesis, and forebrain development (see Figure 4.45). Lastly, 2,387 binding sites could be detected using only 10^3 cells and these DamID-seq peaks also contained a high number of reads which again indicated that at lower cell numbers only the strongest bindings sites were detected. Having confirmed that the peak calling strategy was suitable and that the observations about DamID-seq and ChIP-seq peaks were reproducible across two different transcription factors and cell lines, a software package to enable the accurate and reproducible analysis of DamID-seq data was then developed.

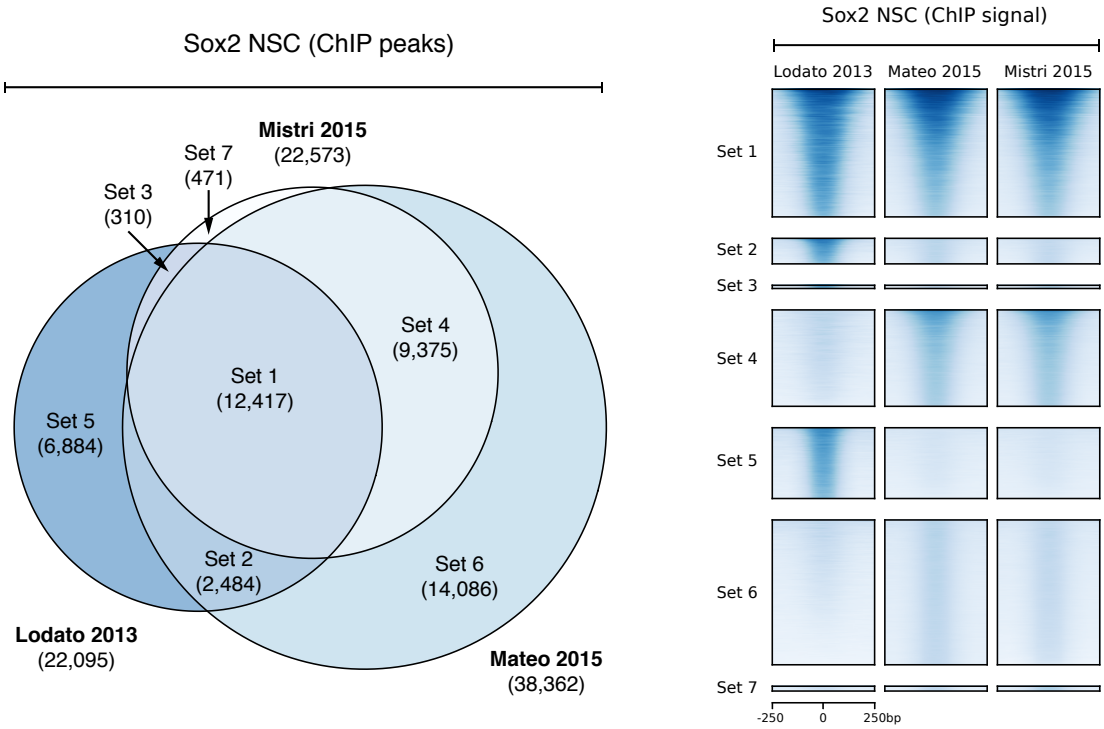

FIGURE 4.40. Comparison of Sox2 NSC peaks from published ChIP-seq experiments.

The Euler diagram on the left hand side represents the overlap between Sox2 NSC peaks from three published ChIP-seq experiments. The heatmap on the right hand side displays read coverage at the peaks from the same ChIP-seq data.

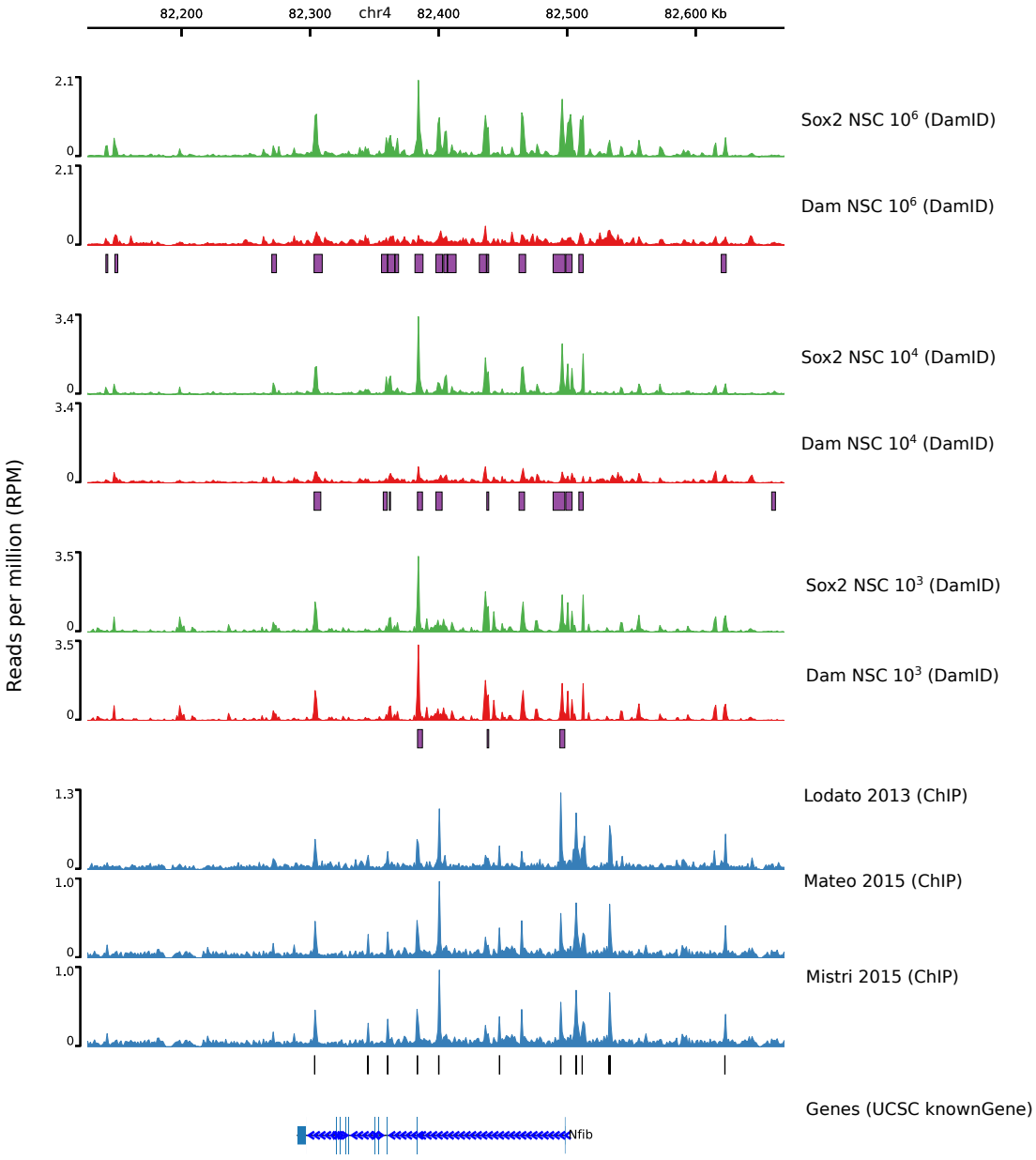

FIGURE 4.41. Genomic snapshot of Sox2 NSC DamID-seq and ChIP-seq peak calls.

Genome browser tracks show good agreement between Sox2 NSC DamID-seq and ChIP-seq read coverage and peak calls. As expected, the signal from the Dam-Oct4 libraries are higher than the Dam libraries at predicted transcription-factor binding sites.

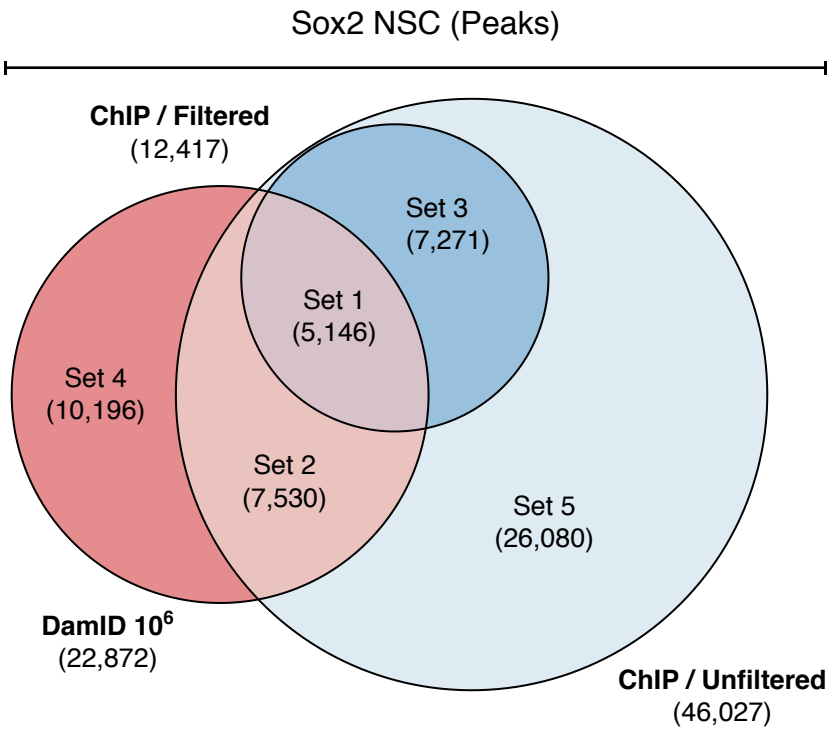

FIGURE 4.42. Comparison of Sox2 NSC peaks between DamID-seq and ChIP-seq data.

The Euler diagram represents the overlap between Sox2 NSC peaks from DamID-seq and ChIP-seq data. The filtered ChIP-seq peaks were derived from the overlap between multiple published experiments, as described previously.

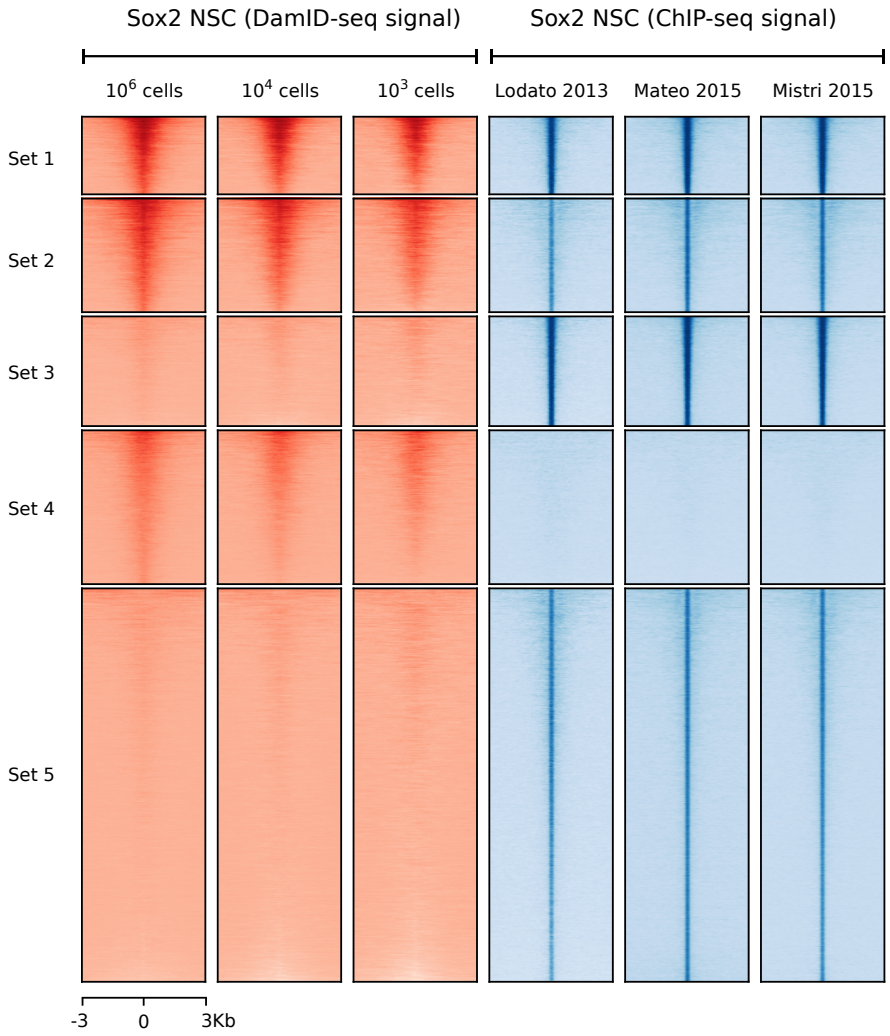

FIGURE 4.43. Read coverage at Sox2 NSC peaks from DamID-seq and ChIP-seq data.

The heatmap shows the pileup of aligned reads at overlapping and unique Sox2 NSC peaks identified from DamID-seq and ChIP-seq data. Each set of peaks relates to the sets found in the corresponding Euler diagram (see Figure 4.42)

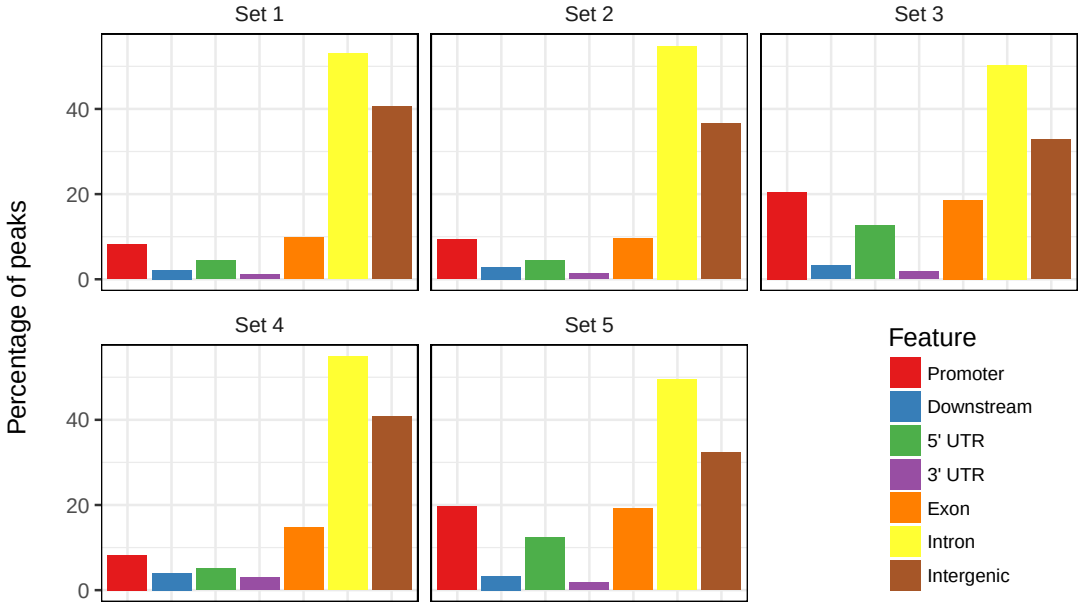

FIGURE 4.44. Annotation of overlapping and unique Sox2 NSC peaks from DamID-seq and ChIP-seq experiments.

Genomic annotation of Sox2 NSC DamID-seq and ChIP-seq peaks as promoter, downstream of gene end, 5′ untranslated region, 3′ untranslated region, exon, intron, or intergenic

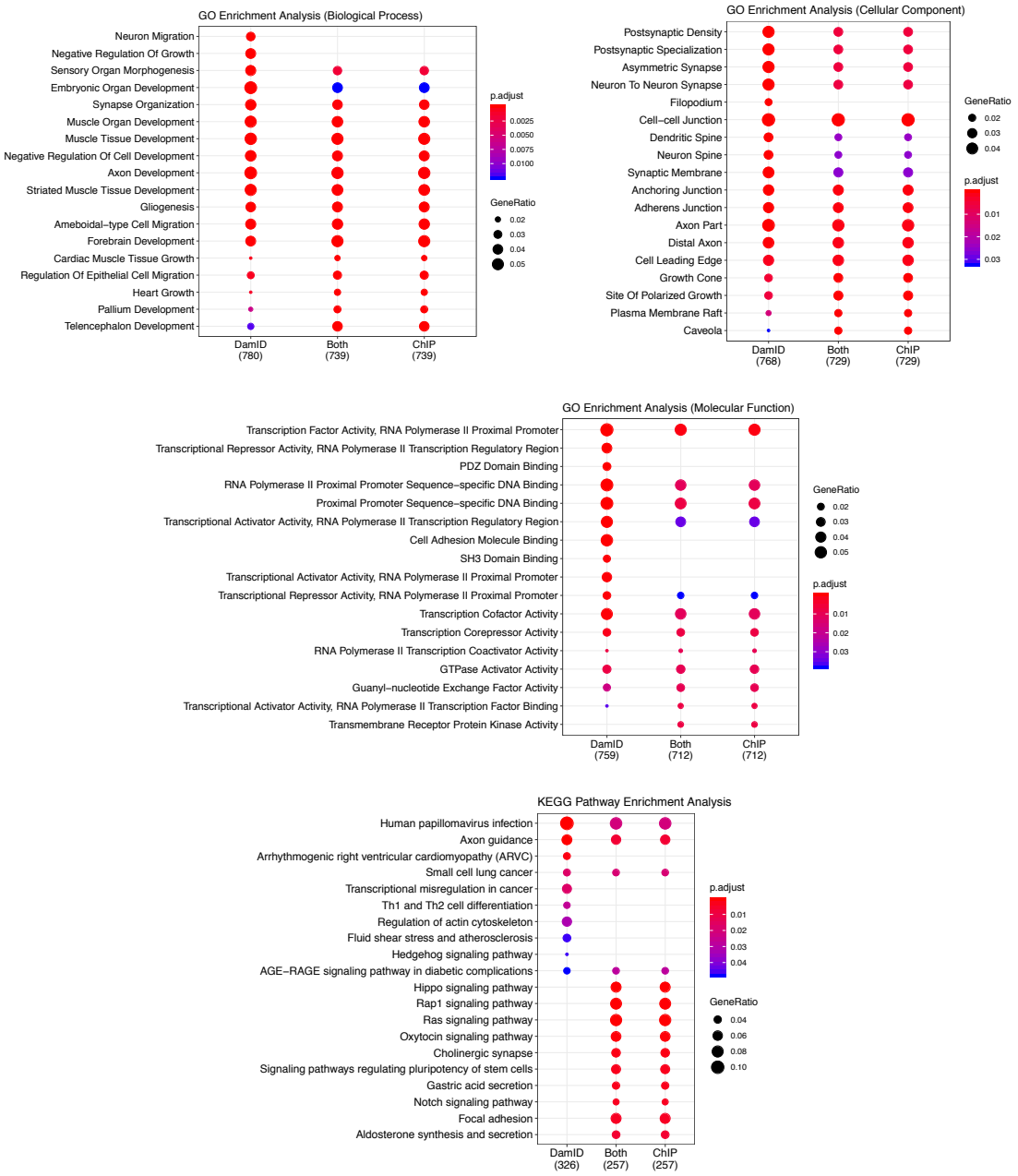

FIGURE 4.45. Ontology and pathway analysis of Sox2 NSC ChIP-seq and DamID-seq binding sites.

Gene ontology analysis of the genes assigned to the top 1,000 overlapping and unique Sox2 NSC ChIP-seq and DamID-seq peaks. The colour of the points in the graph reflect the enrichment significance of each category. The size of the points in the graph reflect the number of peak genes which were assigned to the relevant category.

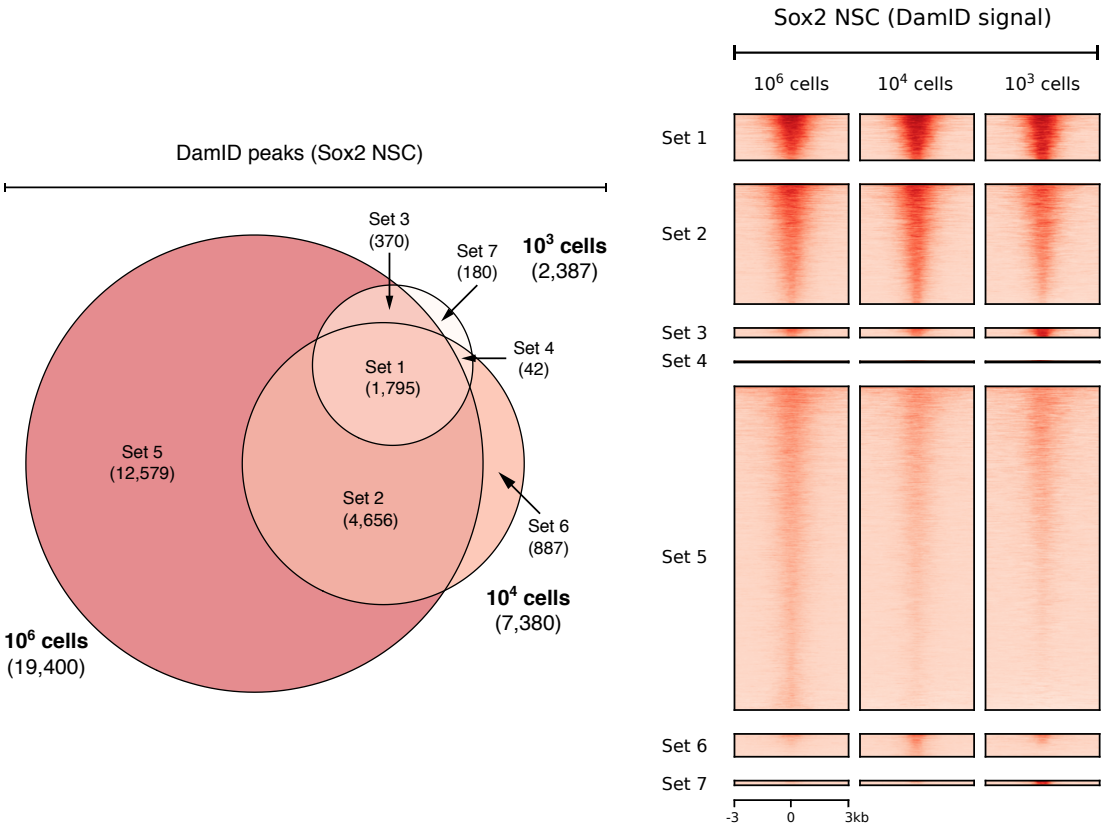

FIGURE 4.46. Comparison of Sox2 NSC peaks from low cell number DamID-seq experiments.

The Euler diagram on the left hand side represents the overlap between DamID-seq peaks from ESC experiments performed using 10^6, 10^4, and 10^3 cells. The heatmap on the right hand side displays read coverage at the peaks from the same DamID-seq data.

4.4.4 The Daim package for analysis of DamID-seq data

In order to analyse DamID-seq experiments in a reproducible and convenient manner, an R/Bioconductor package which handles quantification, pre-processing, and peak calling of the sequencing data was developed (see Figure 4.47). The Daim package offers a workflow to identify chromatin accessibility (discussed in Chapter 5) and transcription-factor binding sites from replicated DamID-seq data which has been

aligned to the genome. The package is based around the RangedSummarizedExperiment object and all functions are capable of interacting with different slots of the object, so no additional extraction of the data by the user is required. The package relies on multiple Bioconductor packages, most important of these is the limma package which provides the core functionality for differential testing using linear modelling and Bayes methods (Ritchie et al., 2015). In reference to the name of the package. Although it is spelled "Daim", it is actually pronounced "Dime", named after the chocolate brand. At the time of writing, there is no Bioconductor or CRAN package called Daim, however there is a ChIP-seq package called DIME (Taslim, Huang, and Lin, 2011). Given that packages should ideally have unique spelling names, I felt it was better to use an alternative yet unique spelling for the Daim package.

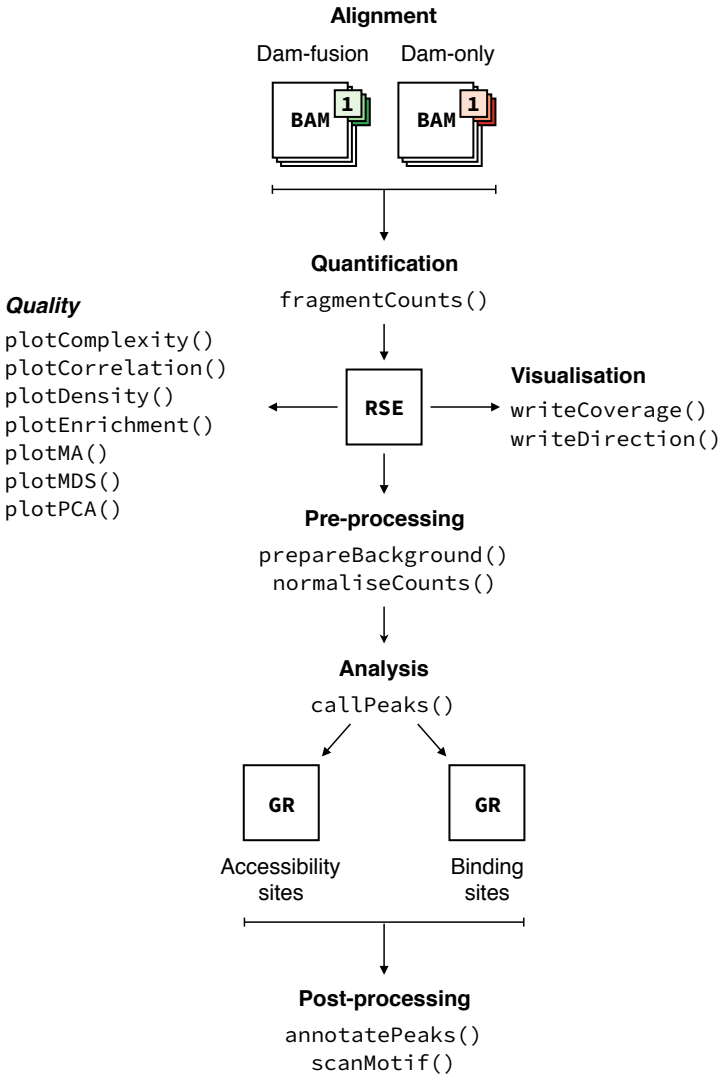

FIGURE 4.47. Overview of the Daim workflow for analysis of DamID-seq data.

The Daim package provides functions for the analysis, interpretation, and visualisation of DamID-seq data. It is primarily used to identify transcription-factor binding and chromatin accessibility sites from DamID-seq experiments.

The workflow begins with the fragmentCounts function which quantifies the level of Dam and Dam-fusion methylation by counting the number of reads aligned to DpnI restriction fragments. Notably, restriction fragments can be generated using either a

custom sequence file in FASTA format or one of the many BSgenome packages already available from the Bioconductor consortium. After the reads are counted, quality control plots and genome browser tracks can be immediately generated to decide whether the experiment was successful (see Figure 4.48). The plotComplexity function displays the number of restriction fragments which have been sequenced at different library sizes to decide whether more sequencing is required. The plotPCA and plotMDS functions can also be used to perform principal component analysis and multidimensional scaling on the read counts to measure the similarity between biological replicates. Additionally, the writeCoverage and writeDirection functions allow the user to generate genome browser files of the restriction fragment read counts and the average fold change between Dam and Dam-fusion libraries, respectively.

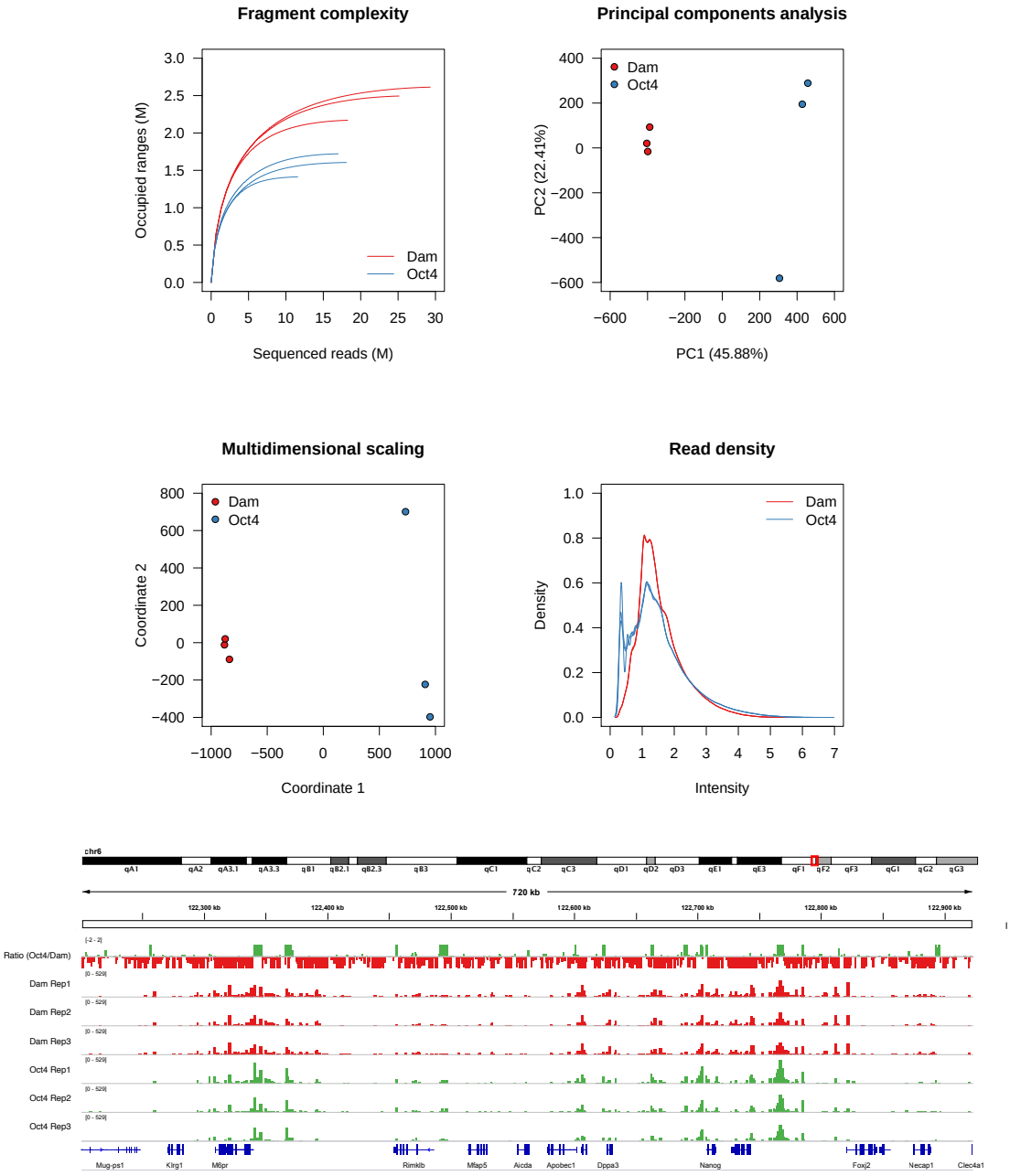

FIGURE 4.48. Quality control plots and genome browser tracks generated using the Daim package.

The Daim package can be used to generate various quality control plots and genome browser tracks to assess whether the DamID-seq experiments were successful. Specifically, Daim can generate genome browser files of the normalised restriction fragment read counts and the average fold change between Dam and Dam-fusion libraries (e.g. Oct4/Dam).

After thorough inspection of the raw data, the prepareBackground and normaliseCounts functions can be used depending on the type of analysis required. The prepareBackground function generates simulated libraries intended to measure background methylation so that a differential methylation analysis between the Dam and simulated libraries can be performed to identify chromatin accessibility sites. For each Dam sample, a simulated library representing random background methylation across the genome is generated by calculating the average number of reads aligned to neighbouring restriction fragments within a range of window sizes (e.g. 5 kb, 10 kb). After this process each restriction fragment will have a read count from each Dam sample and their associated simulated library. These counts can then be used in a conventional two group differential methylation analysis to detect restriction fragments which have been significantly methylated more in the Dam sample over the simulated library (see Subsection 5.4.3 for detailed information regarding the peak calling strategy for chromatin accessibility sites). The normaliseCounts function is more straightforward and is provided to remove technical biases via smooth quantile normalisation between the Dam and Dam-fusion libraries so that a differential methylation analysis can be performed to identify transcription-factor binding sites. The effect of normalisation can be checked using the plotDensity function which generates density plots of the normalised methylation values. The sequencing data is now ready for differential methylation analysis using the callPeaks function to identify either chromatin accessibility or transcription-factor binding sites depending upon which pre-processing function was used previously. The called peaks are returned as GRanges objects which can then be exported in broadPeak format to import into an external genome browser, or additional functional and sequences analyses can be performed. The annotatePeaks function will assign each peak to the nearest gene and genomic feature, including measuring the distance to the closest TSS to investigate the distribution of binding. Lastly, the scanMotif function can be used to search the peak sequences for the occurrence

of a given motif and its associated probability using any PWM file to filter for peaks with a high likelihood of binding or to increase the resolution of the peak regions. The package is freely available to download (`https://github.com/jma1991/daim`) and is released under an open-source MIT license to encourage adoption and further development from the research community.

The release status of Daim is currently at version 1.0 and can be treated as the first official release. The intention is to release the package within the Bioconductor framework, allowing users to download the package from the official Bioconductor repository using the BiocManager package. In order to submit Daim, all of its dependencies must be available to download on either the Bioconductor or CRAN repository. At present, the qsmooth package which Daim uses as part of the normalization strategy is only available through GitHub (Hicks et al., 2018). I have contacted the author and they have kindly submitted the qsmooth package to Bioconductor and it is currently being reviewed. In order to promote and elicit feedback on Daim I am also in the process of asking groups who have employed DamID-seq in their research to review and test the package on their experiments.

4.5 Discussion

These results show that DamID-seq data can be used to identify multiple transcription factor binding sites (Oct4 and Sox2) from different cell types (ESC and NSC) with minimal numbers of cells (10^6, 10^4, and 10^3). Approximately 40% of binding sites detected using ChIP-seq were also identified using DamID-seq – the majority of which exhibited the strongest binding and functionally relevant gene ontology. This percentage was comparable to UV ChIP-seq which ranged anywhere between 4% to 41% reproducibility with formaldehyde ChIP-seq (Steube et al., 2017). Whether this reflects an interesting aspect of the biological function of Oct4 is currently unclear. Peaks which

are identified by multiple technologies and experiments might represent the core network of a DNA-binding protein and this may reflect its core function within the majority of cells within a population. However, those peaks which exhibit variability or appear in a small number of experiments could occur for a number of reasons, both biological and technical. For example, scRNA-seq allows one to identify highly variable genes in order to profile different cell states or rare cell types within a population. Before this technology, it was very difficult to prove that variable expression of a particular gene was a genuine biological phenomenon caused by variable cell states. Whilst the development of single-cell epigenomic profiling technology (e.g. uliCUT&RUN and calling cards) will help in this endeavour, it will still be difficult to target particular cell states in future experiments because profiling will not allow the researcher to identify a marker to use for cell sorting. Data from both scRNA-seq and single-cell epigenome profiling within the same cell would have to be generated in order to find markers and determine whether the highly variable peaks correlate with highly variable genes. Furthermore, it is tempting to attribute distinct functions to binding sites which have been identified by both DamID-seq and either the filtered or unfiltered ChIP-seq peaks. However, the binding sites which are identified by DamID-seq and the filtered ChIP-seq peaks (see Set 1 in Figures 4.33 and 4.42) are those which by definition are less variable across ChIP-seq experiments given the peaks were pre-filtered based upon their presence in multiple experiments (see Figure 4.15). By comparison, the sites detected by DamID-seq and the unfiltered ChIP-seq peaks (see Set 2 in 4.33 and 4.42) are peaks which are present in less than six experiments, which is a very small number to attribute significance (see 4.15). It is also important to remember that the sites detected by DamID-seq only (see Set 4 in 4.33 and 4.42) are not seen in any of the ChIP-seq datasets (see the unfiltered category in 4.33 and 4.42). It is therefore not possible to calculate the presence of peaks, and it is not of interest to calculate how variable the read coverage of these regions is because they are not significantly

enriched above the input read coverage, so any variability would be due to technical variation rather than biological variation. Importantly, the binding sites detected using DamID-seq with 10^3 cells were also the strongest bound in ChIP-seq which indicated that only the most functionally relevant sites were maintained. This validates the DamID-seq approach and ensures that the molecular function of the DNA-binding protein will be captured using minimal numbers of cells.

In order to achieve this level of reproducibility, appropriate normalisation and modelling of the sequencing data was required. Global adjustment methods were found to be imprecise for differential methylation analysis because Dam and Dam-fusion libraries often exhibited global differences which were biologically rather than technically generated. Smooth quantile normalisation was therefore applied to remove any technical variation within groups whilst retaining biological variation between groups. Additionally, the usual assumptions of variance modelling were not suitable for differential methylation analysis because the Dam-fusion libraries were much more variable, so a group-specific variance model was instead employed. These two modifications to the analysis of the sequence count data were essential to achieving an accurate and sensitive peak calling method.

A stand-out result from this work was the apparently large variation in binding sites detected from published Oct4 ESC ChIP-seq experiments. Importantly, this could not be explained by routine quality control metrics or experimental variables. Although this data was only used to identify a subset of highly reproducible sites for comparison with DamID-seq, a deeper investigation of causative differences is recommended. The results from ChIP-seq are routinely interpreted based on the number and distribution of binding sites detected, and usually a biological explanation based on these observations is formulated. If such sites vary this drastically between experiments there is a concern that inaccurate conclusions about the molecular function of the DNA-binding

protein will be attributed.

Lastly, the development of Daim ensures that researchers without the appropriate training in computational biology will be able to comprehensively analyse their own DamID-seq data. It provides functions for quantification, pre-processing and peak calling of the sequencing data and includes downstream functions for visualising coverage, genomic annotation and sequence scanning. Together these functions provide a complete workflow for going from aligned sequencing reads to annotated peak calls. The package is implemented in R and is available under an open-source MIT license to encourage further development by the wider research community.

Chapter 5

Identification of chromatin accessibility from DamID-seq data

5.1 Introduction

Chromatin accessibility is defined as the extent to which chromatinized nuclear DNA can be physically reached or entered by a macromolecule (Klemm, Shipony, and Greenleaf, 2019). Ease of access is determined by a number of biological and structural mechanisms: the occupancy and organization of nucleosomes, the presence of histone variants, the combination of post-translational modifications on histone tails, and the action of transcription factor and chromatin remodelling proteins (Tsompana and Buck, 2014). First, the occupancy and organization of nucleosomes not only govern the higher-order structure of the chromatin but also reshape the availability of binding sites to transcription factors. Interactions between nucleosomes on the same and different chromatin fibres influence the folding of the chromatin which can obstruct macromolecules binding. Additionally, nucleosomes positioned at binding sites physically block transcription factors from interacting with regulatory elements required

for gene expression. Second, the presence of histone variants changes the global structure of the nucleosome allowing the DNA to "breath" or "constrict" around the histone proteins. Multiple variants exist for each core histone protein exhibiting distinct functions within the genome. For example, the variants H2A.Z and H3.3 are enriched within actively transcribed genes and help maintain an accessible state for transcriptional activity during development (Henikoff and Smith, 2015). By comparison, the variant macroH2A is enriched within transcriptionally silent domains such as the inactivated X chromosome (Chadwick and Willard, 2001). Third, the combination of post-translational modifications on histone tails change the affinity of DNA wrapped around the nucleosome (Bannister and Kouzarides, 2011). For example, the acetylation of particular lysines on multiple histones (notably H3K9, H3K14, H3K18, H4K5, H4K8, and H4K12) by histone acetyltransferases (HATs) neutralise its positive charge weakening the interaction between histones and the DNA molecule. This modification can also be removed by histone deacetylases (HDACs) to strengthen the interaction and decrease chromatin accessibility (Görisch et al., 2005). Similarly, the phosphorylation of a variety of amino acids (serines, threonines, and tyrosines) on different histones (H3 and H4) by kinases introduce a highly negative charge to the histone which loosens the chromatin. Phosphatases though carry out the opposite modification, removing phosphate groups which reduces the negative charge including chromatin accessibility (Schick et al., 2015). Finally, the action of transcription factors in response to external stimuli provide sequence-specific changes in accessibility to the chromatin. Given the wide variety of transcription factors, a number of models have been discovered which facilitate chromatin remodelling: the factor can passively compete for binding sites with dynamic nucleosomes which turnover regularly, the factor and active chromatin remodellers bind to non-nucleosomal DNA to displace nucleosomes in *cis* stabilized by a secondary transcription factor, the factor binds to accessible regulatory elements which recruit active chromatin remodellers to

displaced nucleosomes in *trans* stabilised by a secondary transcription factor, or the factor binds directly to the DNA and displaces the nucleosome with or without help from chromatin remodelling proteins (Klemm, Shipony, and Greenleaf, 2019). Overall, the interplay between these biological and structural mechanisms generates a dynamic chromatin accessibility landscape which is often cell-type specific and reflects their requirement for particular genes to be expressed through regulatory elements such as promoters and enhancers.

Chromatin accessibility assays (including ATAC-seq, DNase-seq, and FAIRE-seq) allow researchers to determine which regions of the genome are open and therefore implicated in genomic regulation (see Section 1.2 for a description of each method). Genomic features such as promoters and enhancers are typically exposed and bound by multiple DNA-binding proteins in order to drive transcription and other epigenetic modifications. The wide application of these assays has already provided insights into the mechanisms of iPSC reprogramming and embryo development, highlighting the importance of studying chromatin biology (Li et al., 2017; Gao et al., 2018). Whilst these assays continue to be used with great success, they are limited in a number of ways: DNase-seq and FAIRE-seq require at least 1 x 10^6 cells in order to generate a sufficient DNA yield, consequently they are impractical for use on rare cell types (Song and Crawford, 2010; Simon et al., 2012). FAIRE-seq further suffers from a low signal-to-noise ratio, therefore capturing only the most accessible regions and making biological interpretation challenging (Tsompana and Buck, 2014). DNase-seq additionally exhibits DNase I cleavage bias, resulting in favouritism between sites which must be normalised (Yardımcı et al., 2014). The recent development of ATAC-seq has largely usurped these two other assays because the experimental protocol is relatively quick and simple, but impressively can be performed using single cells (Buenrostro et al., 2013). There is however still one drawback, none of these assays can measure

chromatin accessibility from individual cell types without isolation. A laborious process which not only lends itself to human error, but also to undesirable alteration of the chromatin landscape through chemical handling (Tsompana and Buck, 2014).

Having determined in previous chapters that Dam preferentially binds to regions of the genome generally considered open (enriched for H3K4me1, H3K4me2, and H3K27ac histone modifications) it is conceivable that Dam may be used independently to measure chromatin accessibility. This re-purposing of the sequencing data would double the application of a single DamID-seq experiment and give more context to the DNA-binding sites which have already been identified. Encouragingly, one or two research groups have already begun to exploit this phenomenon and have used transgene-driven Dam expression to profile *in vivo* chromatin accessibility from individual cell types (Sha et al., 2010; Aughey et al., 2018). This solves the aforementioned complications associated with isolation and culture of rare cell types.

Whilst these studies provide a first proof of concept, many of these experiments were performed with large amounts of starting material from organisms with comparatively small genomes (e.g. *C. elegans* and *D. melanogaster*). Consequently, there is no direct evidence that this approach is usable for large mammalian genomes in situations where cell number is limited. Comparative analyses between DamID-seq and other chromatin accessibility assays are also currently limited, leading one to worry about its efficacy and potential wide adoption. For example, whilst Aughey and colleagues do show that DamID-seq yields comparable results to ATAC-seq and FAIRE-seq, their study does not include a comparison with DNase-seq and there is no attempt at reproducing the results using data from different studies (Aughey et al., 2018). By using data from more than one study, the accuracy and reproducibility of DamID-seq could be better evaluated given the expected levels of variability between and within

replicates from different assays. These arguments therefore warrant the need for further investigation of the chromatin accessibility sites identified from DamID-seq data in order to establish whether or not this is a convincingly forthcoming and perhaps superior method.

5.2 Aims

The aims of this chapter are to determine whether Dam methylation can be used to measure chromatin accessibility, how accurate this measurement is in comparison to previously established assays, and whether accessibility can be measured using lower cell numbers. To answer these questions, DamID-seq data from mouse embryonic stem cells (10^6, 10^4, 10^3, and 10^2 cells) and embryonic fibroblast cells (10^6 cells) will be compared to publicly available ATAC-seq, DNase-seq and FAIRE-seq data. A computational method to identify chromatin accessibility sites from DamID-seq data will also be described and is implemented in the Daim software package.

5.3 Attribution

The ATAC-seq, ChIP-seq, DNase-seq, FAIRE-seq, and RNA-seq libraries presented in this chapter were generated by the relevant research groups and the raw sequencing data was analysed by the author, James Ashmore. The DamID-seq libraries were generated by Dr Luca Tosti, a previous PhD student in Prof Keisuke Kaji's research group.

5.4 Results

5.4.1 Evaluation of Dam methylation and chromatin accessibility

In order to determine whether Dam methylation can be used to measure chromatin accessibility, the agreement between DamID-seq and other chromatin accessibility assays from ESCs and MEFs was investigated. Firstly, read coverage at representative accessible regions in the genome showed that Dam methylation corresponded with the signal from multiple independent ATAC-seq, DNase-seq, and FAIRE-seq experiments (see Figures 5.1 and 5.2). These regions also contained CTCF binding and histone modifications (including H3K4me1, H3K4me3, and H3K27ac) associated with promoters and enhancers. Interestingly, the resolution of accessibility regions measured by Dam also appeared to be higher than the resolution of transcription factor binding measured by Dam-fusion. This observation may be explained by the fact that the Dam-fusion protein methylates accessible GATC sites around the transcription factor binding site, where as the Dam protein directly methylates the accessible region. The signal of Dam methylation also appeared to be much higher than all other assays displayed. This was apparent given the scaling required for the genome browser tracks for each assay, whereby the maximum value for the DamID-seq samples was nearly three times higher than the maximum value for the other assays. Next, read coverage at promoters showed that Dam methylation was localised around the TSS and decreased with RNA-seq expression and RNA PolII occupancy (see Figure 5.3). Read coverage at enhancers also showed Dam methylation was present and decreased with p300, H3K4me1, and H3K27ac occupancy (see Figure 5.4). Importantly, these patterns were consistent with those from other chromatin accessibility assays.

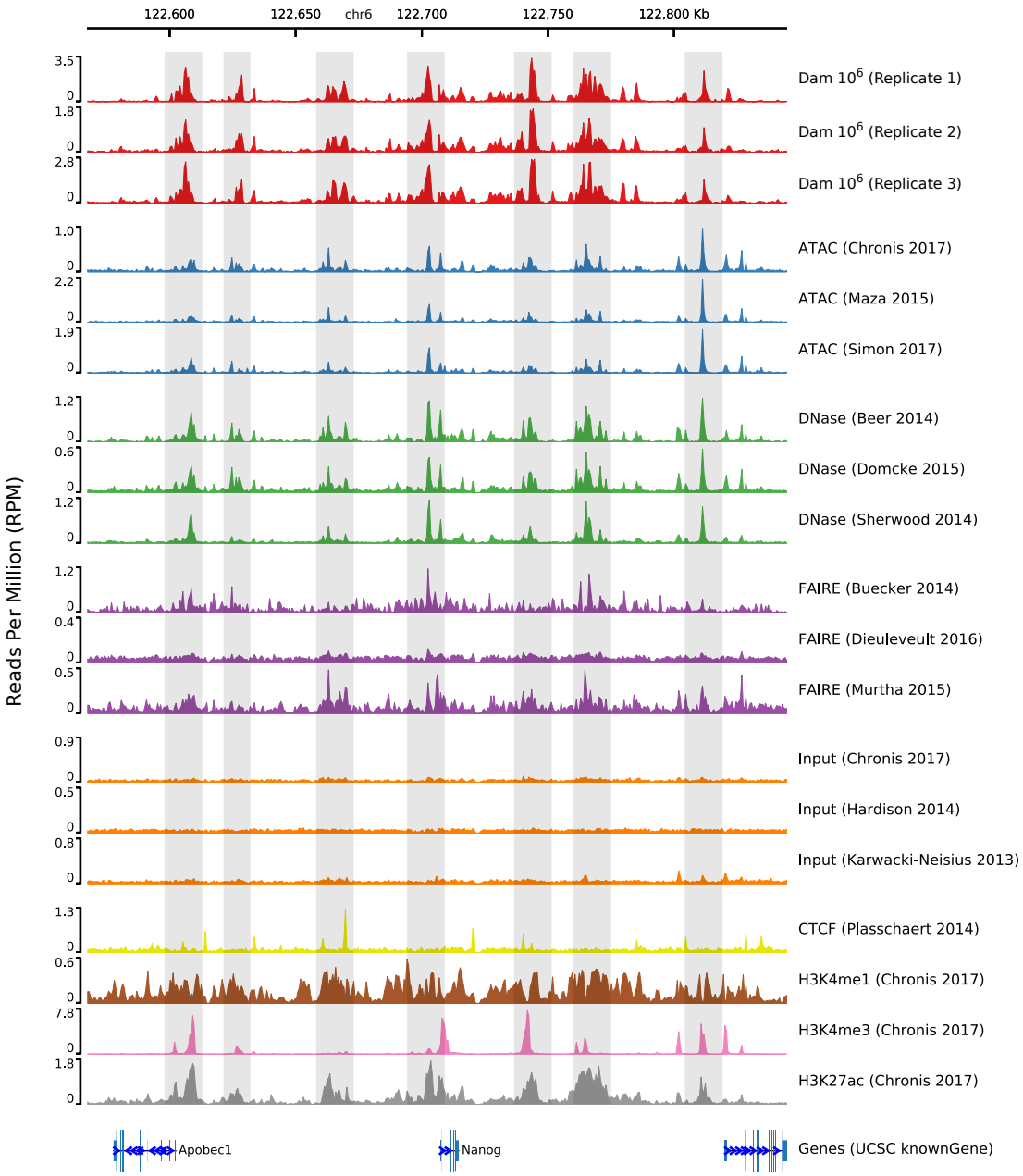

FIGURE 5.1. Genomic snapshot of Dam methylation in ESCs at the *Nanog* locus.

Comparison of DamID-seq with ATAC-seq, DNase-seq and FAIRE-seq data. Active promoters and enhancers are highlighted using CTCF, H3K4me1, H3K4me3, and H3K27ac ChIP-seq data. The vertical grey bars highlight regions of agreement between assays.

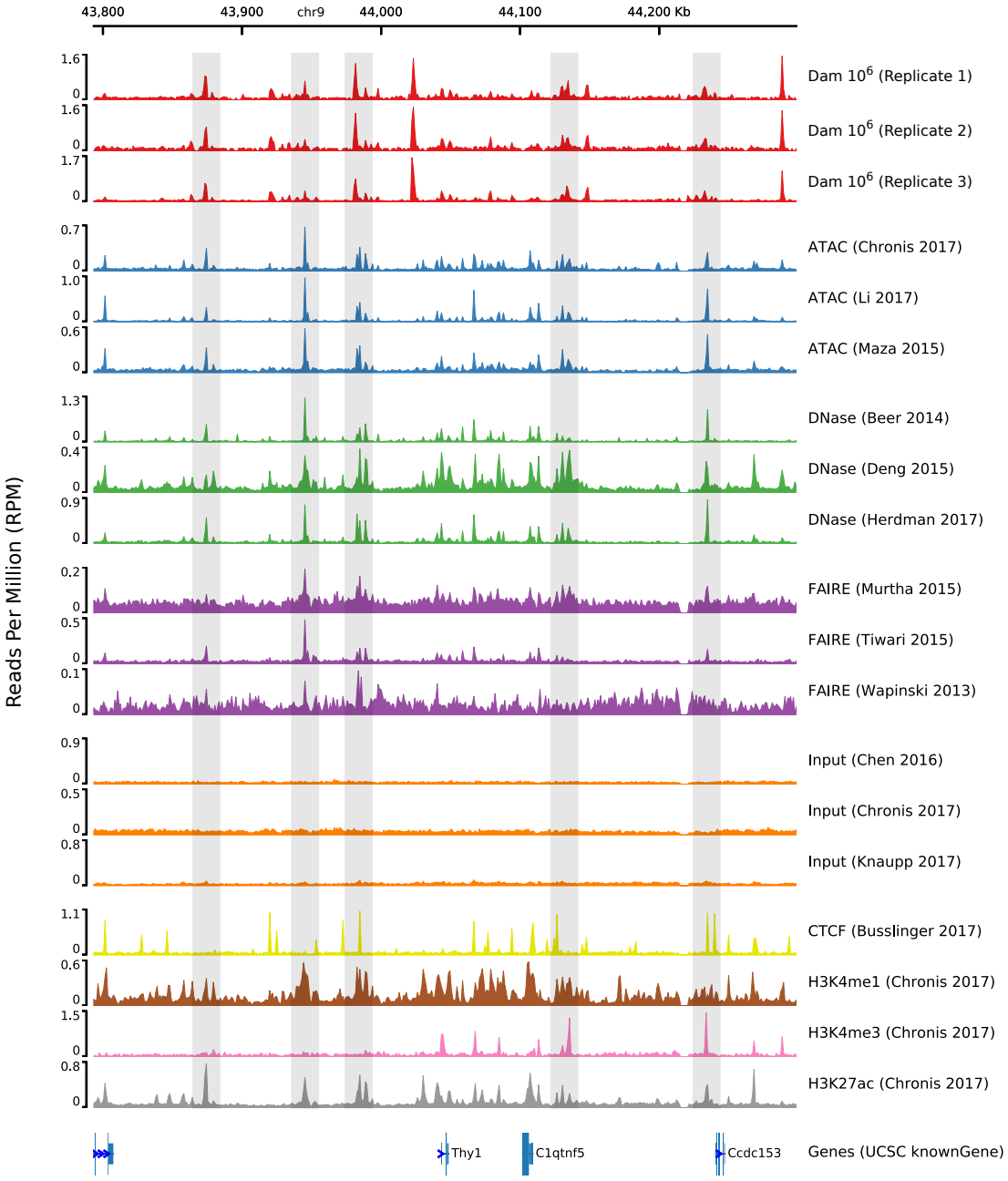

FIGURE 5.2. Genomic snapshot of Dam methylation in MEFs at the *Thy1* locus.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. Active promoters and enhancers are highlighted using CTCF, H3K4me1, H3K4me3, and H3K27ac ChIP-seq data. The vertical grey bars highlight regions of agreement between assays.

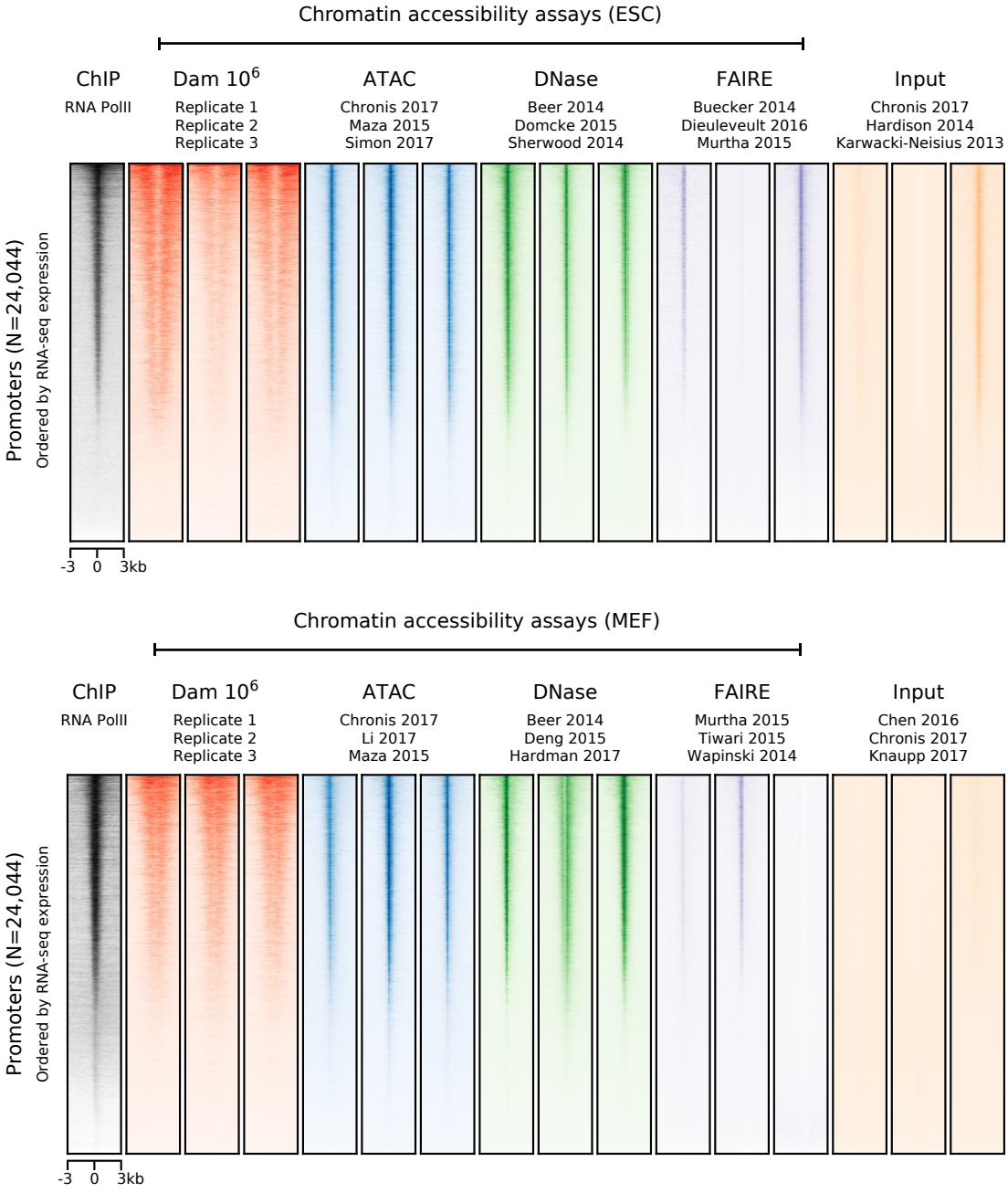

FIGURE 5.3. Heatmap of Dam methylation in ESCs and MEFs at promoter regions.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. The promoter regions are ranked by decreasing RNA-seq expression. Active promoters are highlighted using RNA PolII ChIP-seq data.

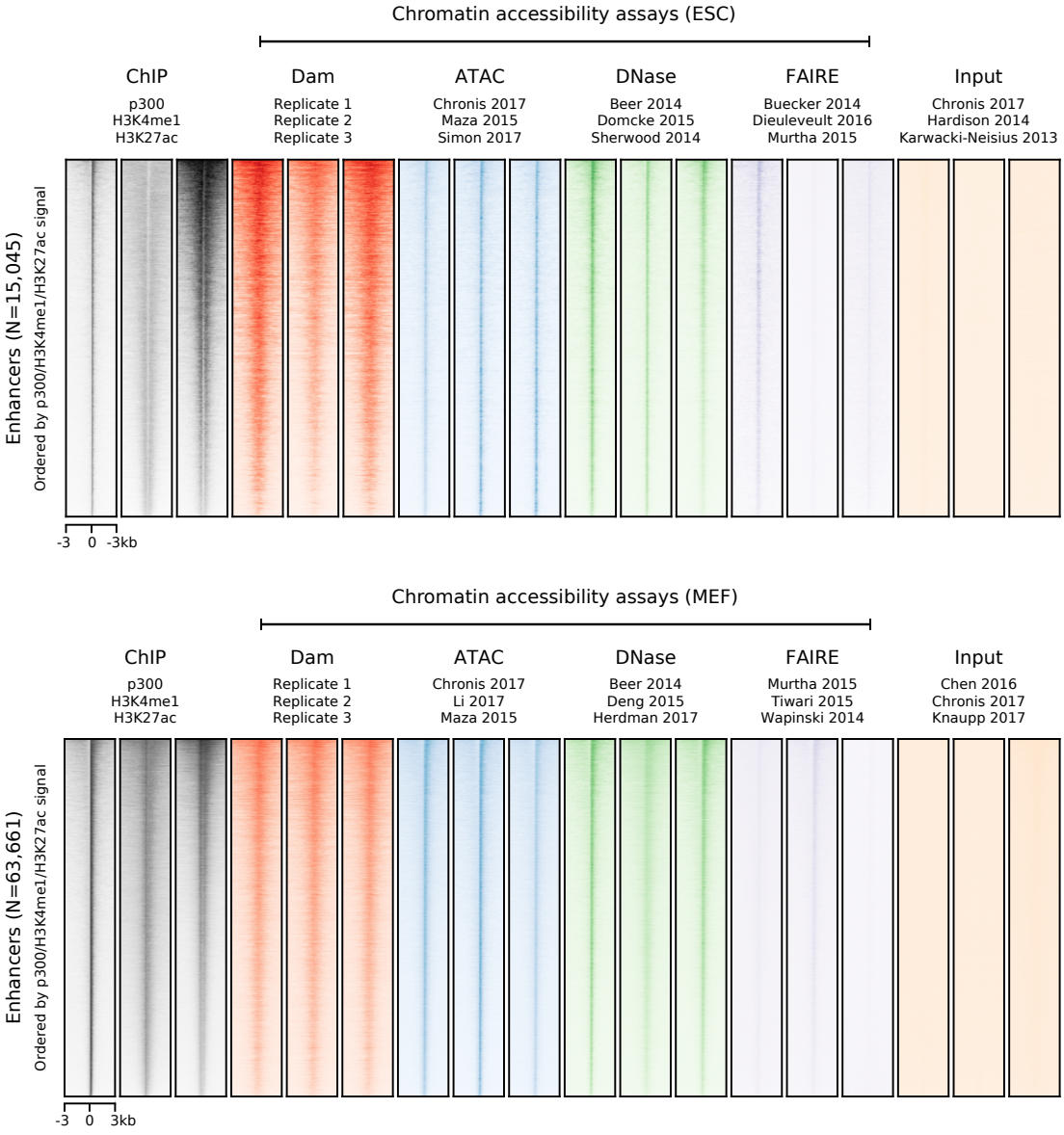

FIGURE 5.4. Heatmap of Dam methylation in ESCs and MEFs at enhancer regions.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. The enhancer regions are ranked by decreasing RNA-seq expression. Active enhancers are highlighted using p300, H3K4me1, and H3K27ac ChIP-seq data.

To ensure Dam methylation was not generically deposited, differentially accessible regions between ESCs and MEFs were examined. To begin with, read coverage at

representative promoters and enhancers showed that Dam methylation corresponded with cell-type specific gene expression and enhancer activity (see Figures 5.5, 5.6, 5.7, and 5.8). The promoters and enhancers of pluripotency and fibroblast related genes were only accessible in the relevant cell-types, as seen from the multiple independent ATAC-seq, DNase-seq, and FAIRE-seq experiments. Genome-wide expression profiling was then used to identify differentially expressed genes from public RNA-seq data (Milagre et al., 2017). Principal component analysis demonstrated that the ESC and MEF replicates were more similar within groups than between groups, and that there were substantial differences between their gene expression profiles (see Figure 5.9). A volcano plot of gene expression changes demonstrated that a large number of genes were significantly differentially expressed (see Figure 5.10). Moreover, the differentially expressed genes were involved in processes indicative of each cell-type's function and origin (see Figures 5.11 and 5.12). Surprisingly, read coverage at the promoters of these genes showed very little difference in accessibility and methylation in the ESC experiments (see Figure 5.13). A possible explanation for this is that many promoters in ESCs reside in a bivalent state (contain both active and repressive chromatin modifications) meaning that although the promoter is accessible the gene may not be transcribed (Harikumar and Meshorer, 2015). By comparison, read coverage in the MEF experiments showed a substantial difference in accessibility and methylation as one would have expected. In order to identify differentially active enhancers, overlapping and unique enhancer regions were located by intersecting peak calls from ChIP-seq data of p300, H3K4me1, and H3K27ac enhancer-associated chromatin modifications (see Figure 5.14). Reassuringly, read coverage at these enhancers showed substantial differences in accessibility and methylation corresponding to cell-type specific enhancer activity (see Figure 5.4). Together, these results strongly indicated that Dam methylation was correlated with accessibility, and that sequencing data from Dam libraries could be used to profile the chromatin landscape.

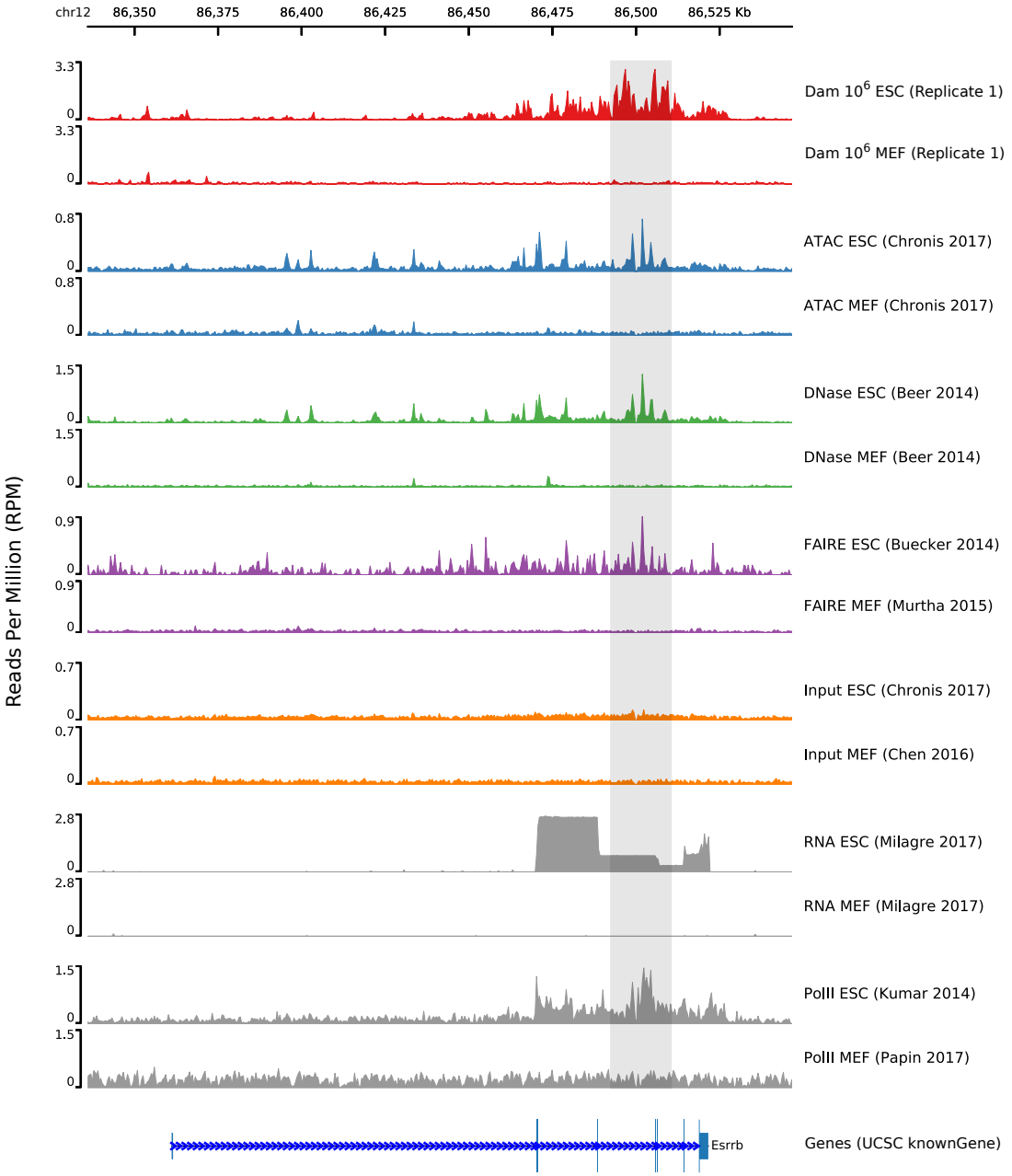

FIGURE 5.5. Genomic snapshot of Dam methylation in ESCs and MEFs at the *Esrrb* locus.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. Differentially expressed genes are highlighted using RNA-seq and RNA PolII ChIP-seq data. The vertical grey bars highlight regions of agreement and disagreement between assays and cell types, respectively.

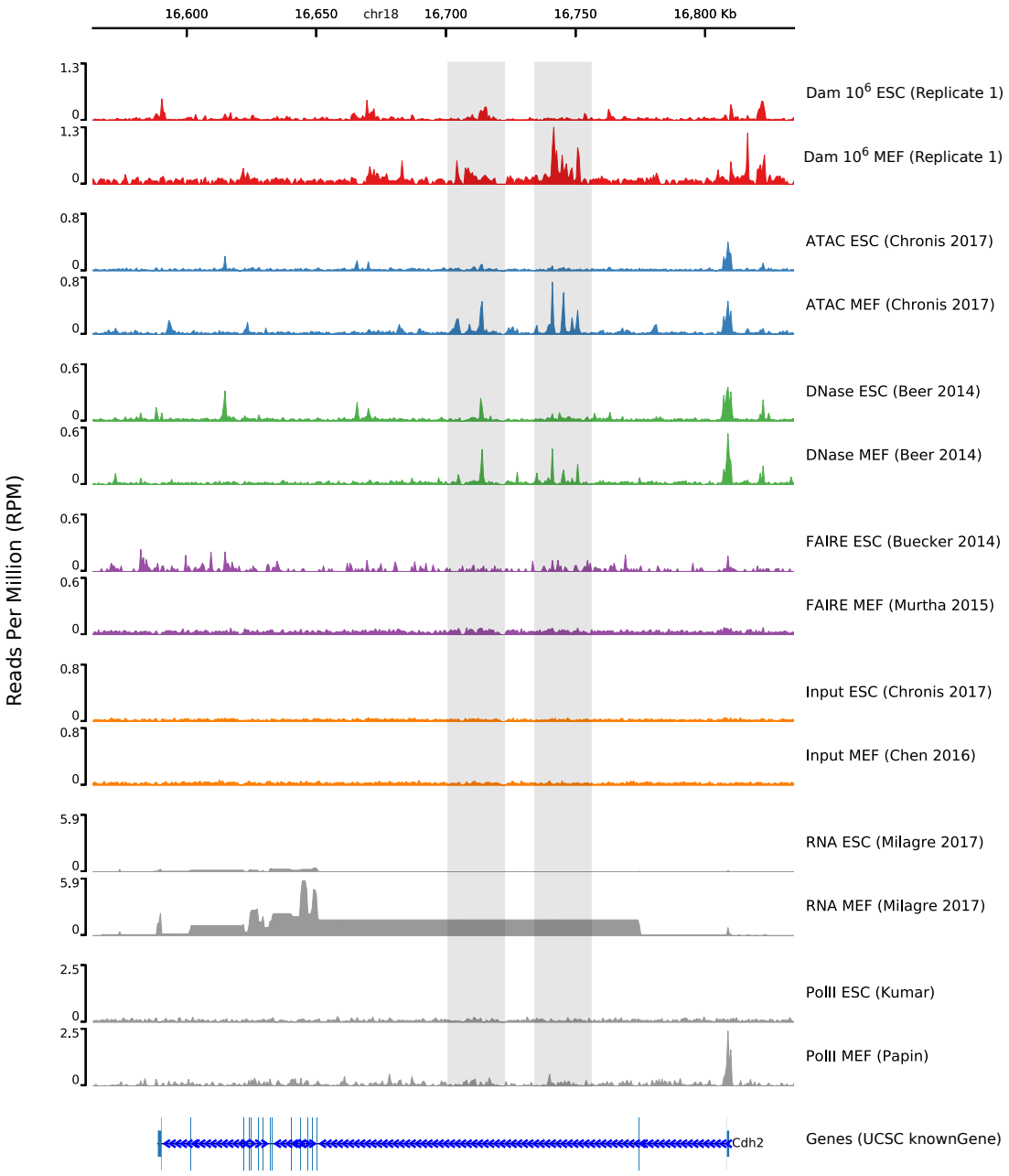

FIGURE 5.6. Genomic snapshot of Dam methylation in ESCs and MEFs at the *Cdh2* locus.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. Differentially expressed genes are highlighted using RNA-seq and RNA PolII ChIP-seq data. The vertical grey bars highlight regions of agreement and disagreement between assays and cell types, respectively.

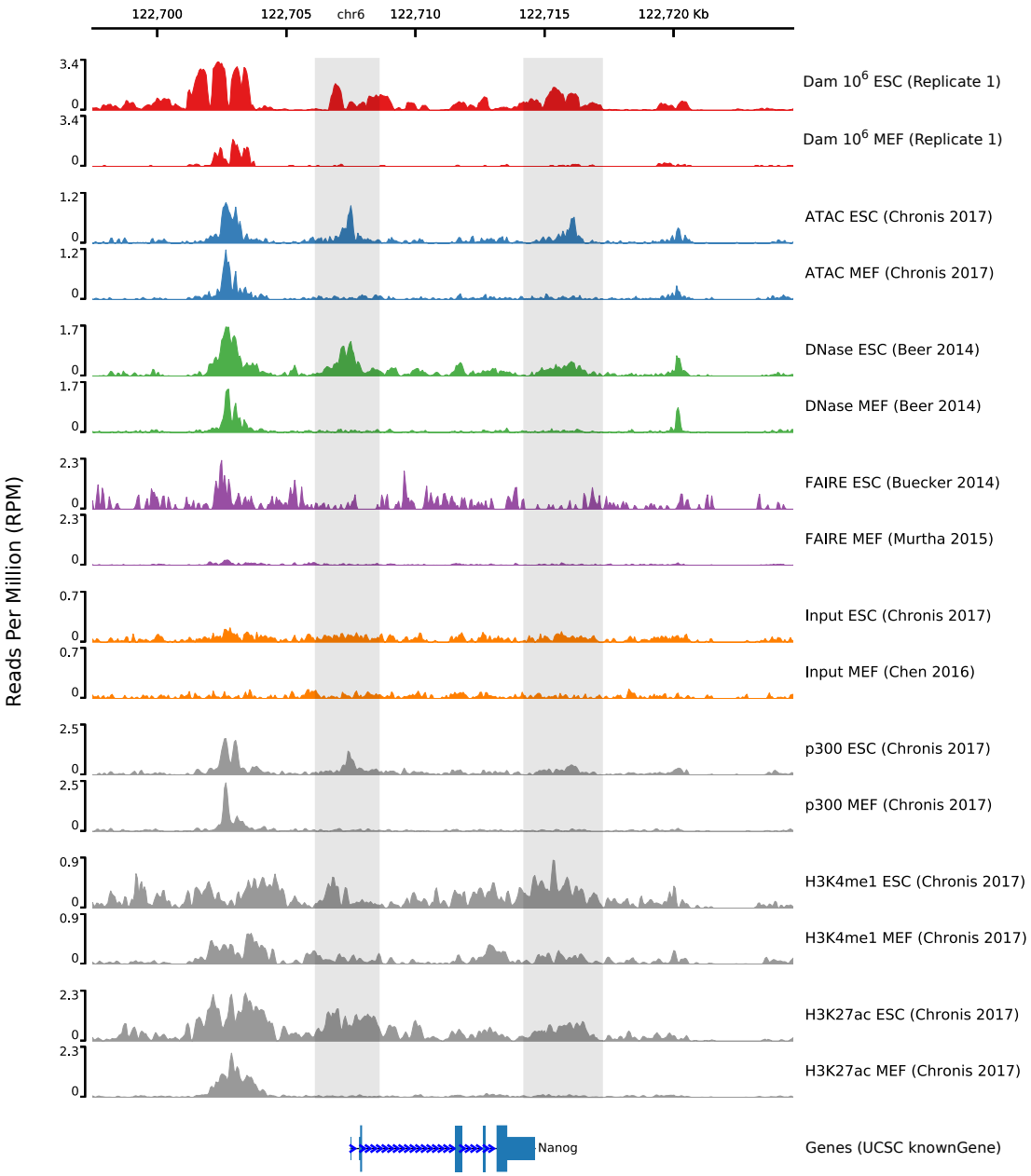

FIGURE 5.7. Genomic snapshot of Dam methylation in ESCs and MEFs at the *Nanog* locus.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. Differentially expressed enhancers are highlighted using p300, H3K4me1, and H3K27ac ChIP-seq data. The vertical grey bars highlight regions of agreement and disagreement between assays and cell types, respectively.

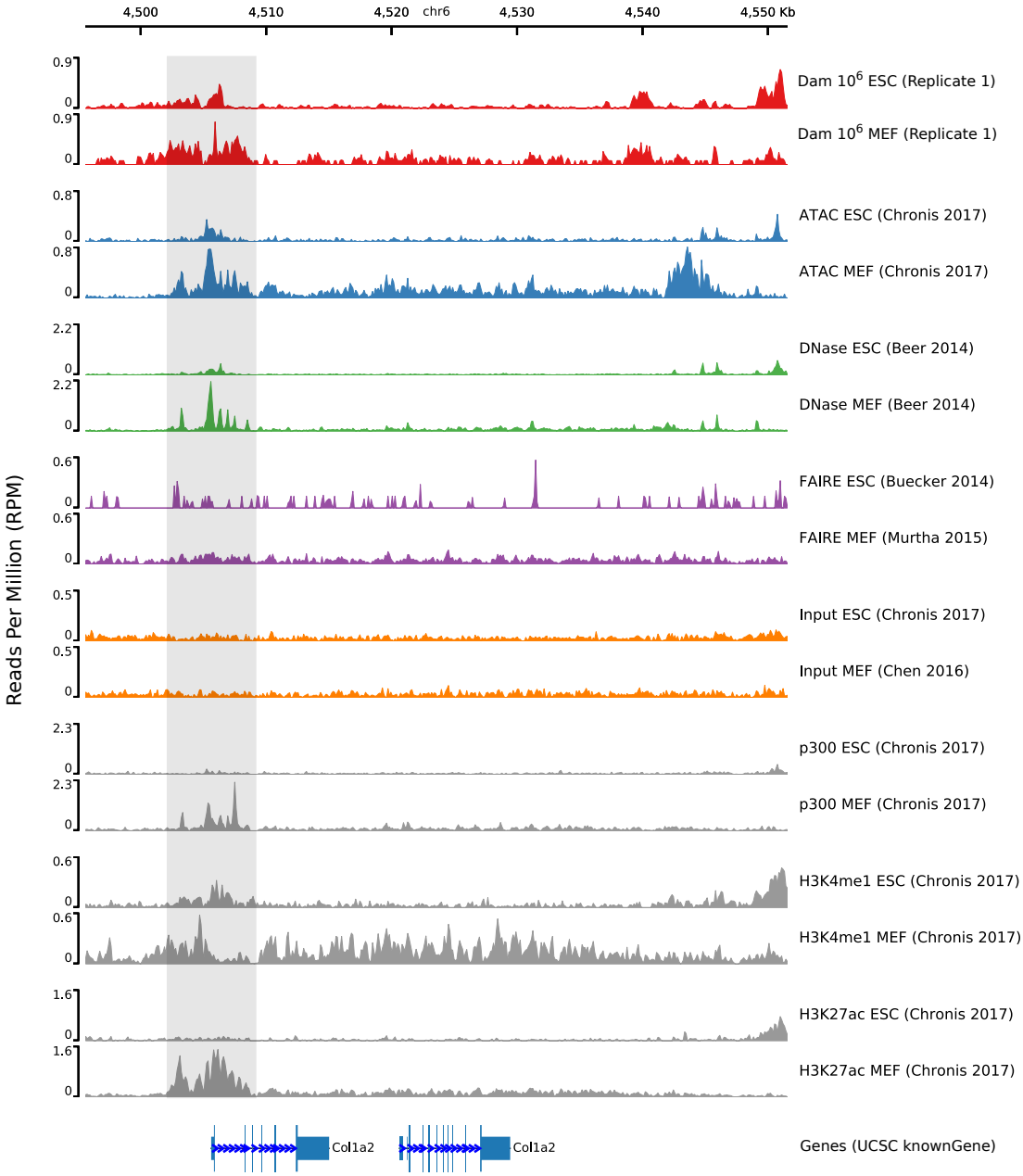

FIGURE 5.8. Genomic snapshot of Dam methylation in ESCs and MEFs at the *Cola12* locus.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. Differentially expressed enhancers are highlighted using p300, H3K4me1, and H3K27ac ChIP-seq data. The vertical grey bars highlight regions of agreement and disagreement between assays and cell types, respectively.

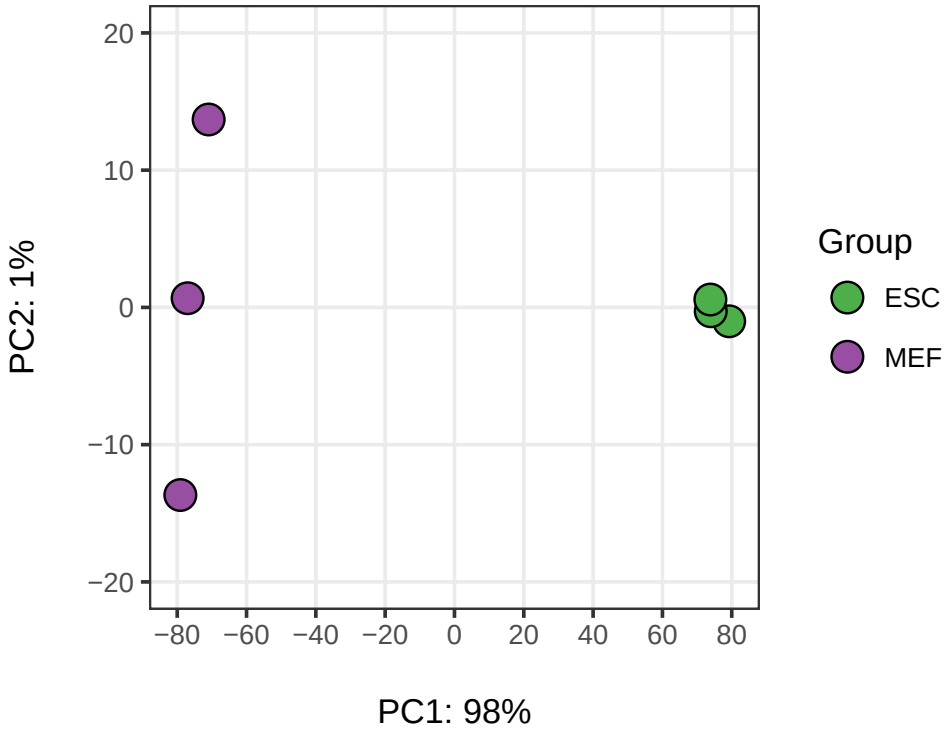

FIGURE 5.9. Principal component analysis of ESC and MEF gene expression.

Principal component analysis of ESC and MEF RNA-seq data. The X axis measures the first principal component (PC1) which accounts for 98% of the variance. The Y axis measures the second principal component (PC2) which accounts for 1% of the variance.

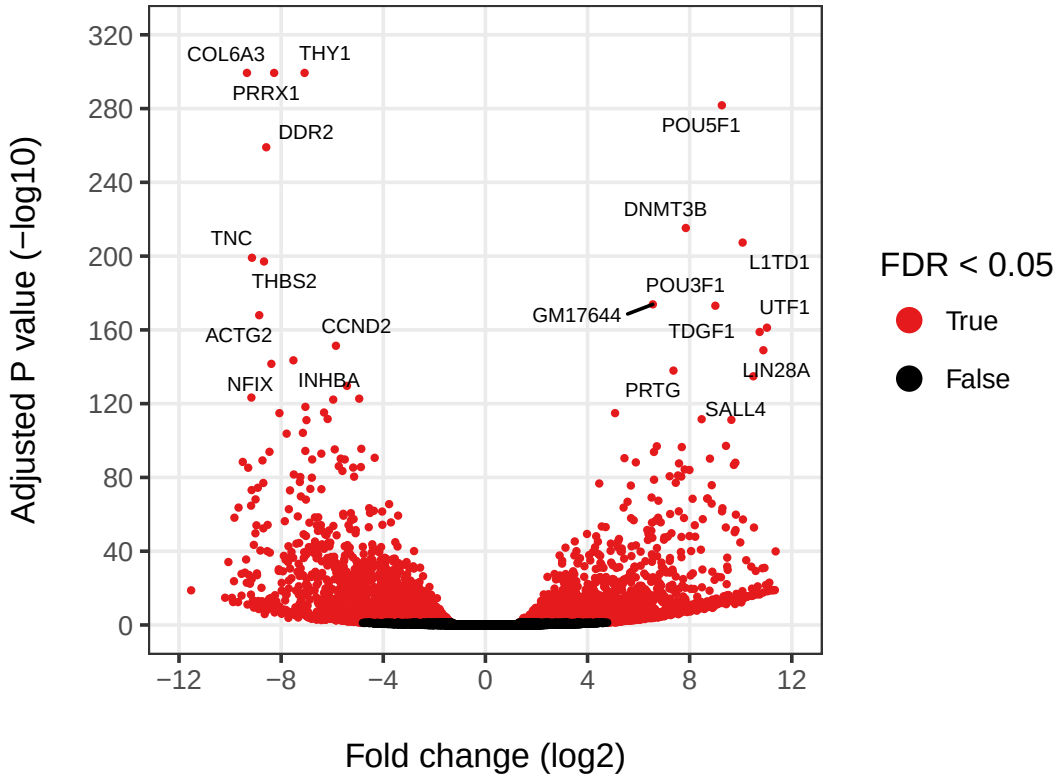

FIGURE 5.10. Volcano plot of ESC and MEF gene expression changes.

Volcano plot of ESC and MEF RNA-seq data. There are 2,164 and 2,194 significantly differentially expressed genes (FDR < 0.05) between ESCs and MEFs, respectively. The X axis measures the gene expression fold change and the Y axis measures the adjusted P value from differential expression analysis. Significantly differentially expressed genes are coloured red and the top 10 differentially expressed genes are labelled.

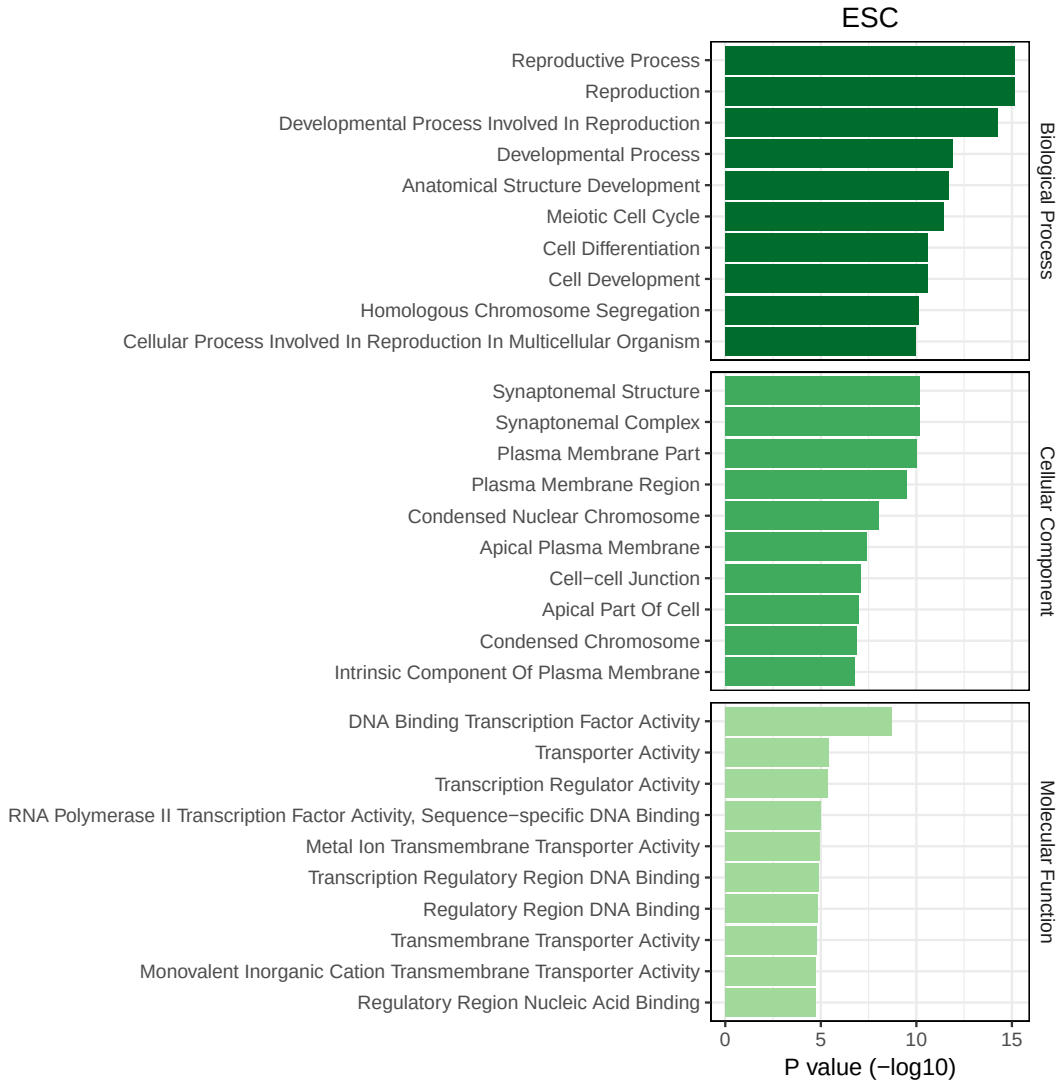

FIGURE 5.11. Gene ontology analysis of differentially expressed ESC genes.

Gene ontology analysis of differentially expressed ESC genes for biological process, cellular component, and molecular function categories.

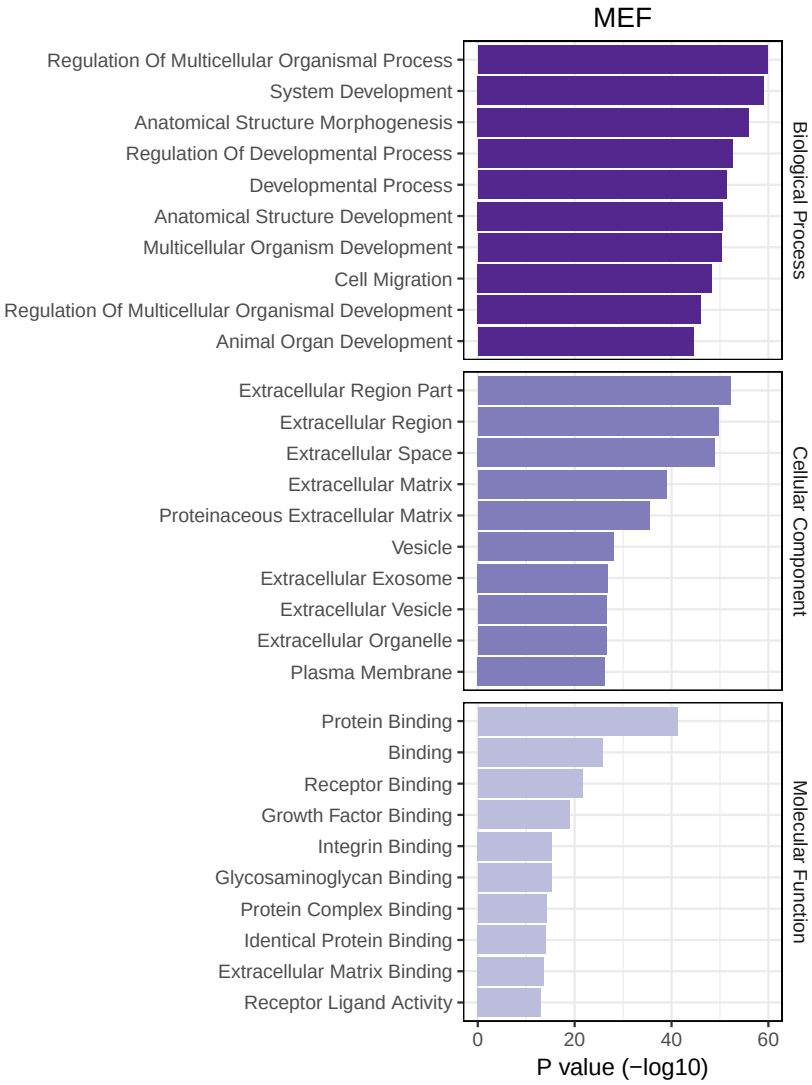

FIGURE 5.12. Gene ontology analysis of differentially expressed MEF genes.

Gene ontology analysis of differentially expressed MEF genes for biological process, cellular component, and molecular function categories.

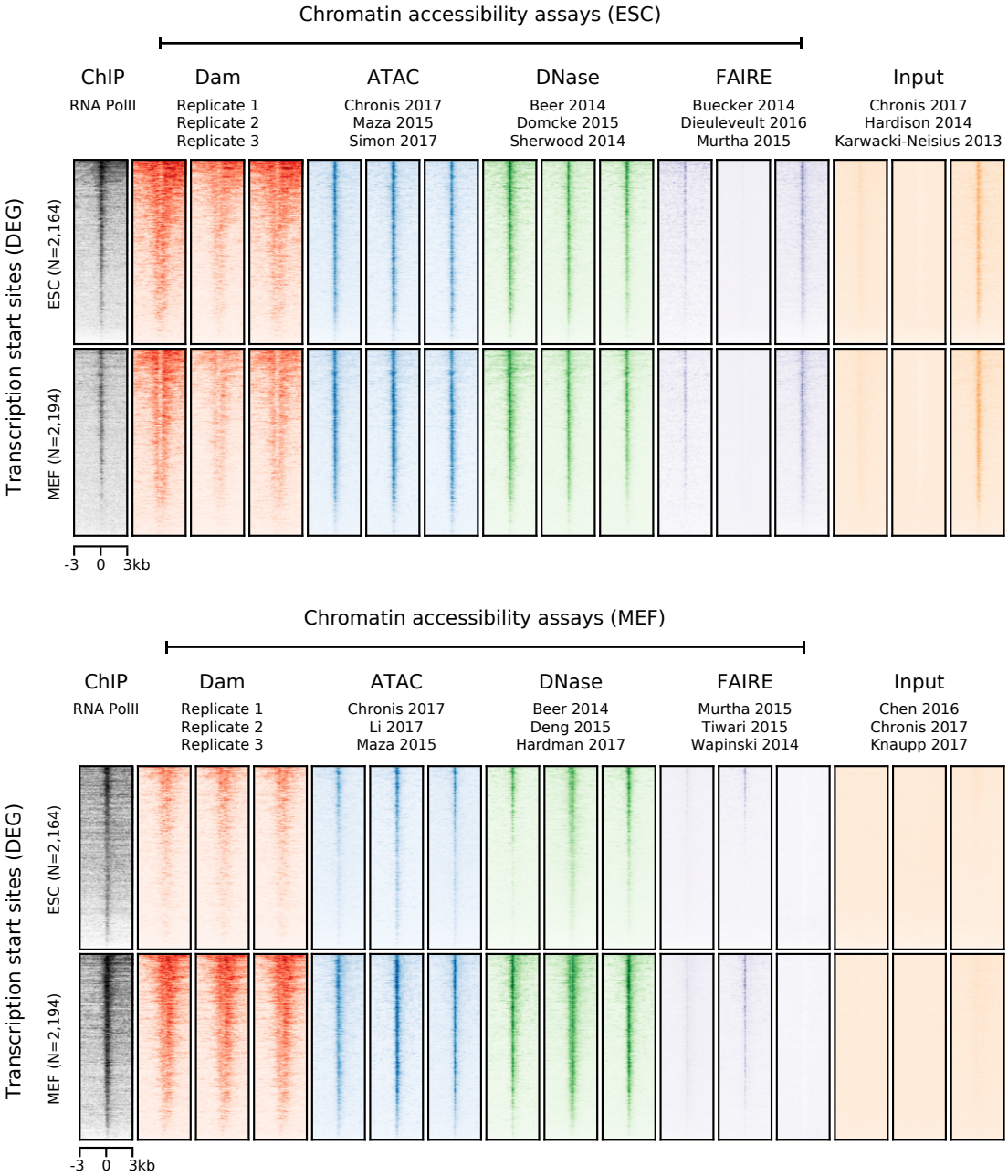

FIGURE 5.13. Heatmap of Dam methylation in ESCs and MEFs at DEG promoter regions.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. The DEG promoter regions are ranked by decreasing RNA-seq expression. Active promoters are highlighted using RNA PolII ChIP-seq data.

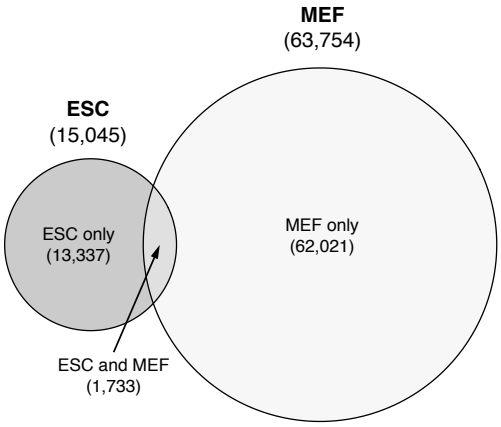

FIGURE 5.14. Euler diagram of ESC and MEF enhancer regions.

Diagram representing the overlap between ESC and MEF enhancer regions.

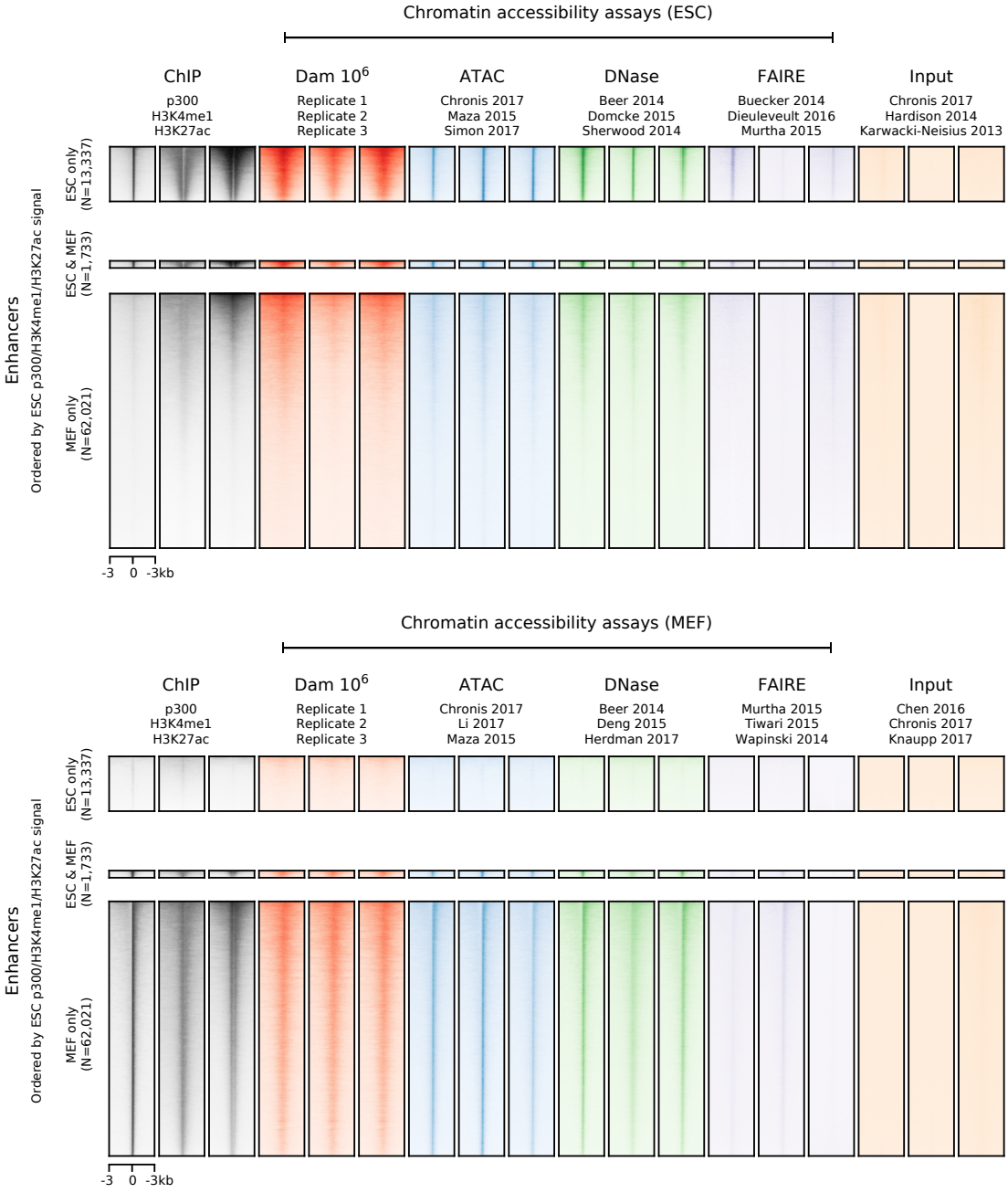

FIGURE 5.15. Heatmap of Dam methylation in ESCs and MEFs at DB enhancer regions.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq data. The DB enhancer regions are ranked by enhancer expression. Active enhancers are highlighted using p300, H3K4me1, and H3K27ac ChIP-seq data

5.4.2 Comparison of accessibility sites between and within assays

In order to evaluate the accuracy and sensitivity of DamID-seq data, a reliable set of chromatin accessibility sites from already established assays was first required. Peaks were called using publicly available ATAC-seq, DNase-seq, and FAIRE-seq data from three independent ESC and MEF experiments. The level of variability between and within assays was measured by comparing the number of overlapping and unique peaks identified from the different experiments. For easier comprehension, different groups of overlapping and unique peaks were enumerated: Set 1 refers to peaks in all three experiments; Sets 2 to 4 refer to peaks in two out of three experiments; and Sets 5 to 7 refer to peaks in only one experiment. First, there were 82,357 and 77,340 ATAC-seq peaks which overlapped all three ESC and MEF experiments, respectively (see Figure 5.16). Read coverage at peaks in Set 1 was higher than in all other sets, indicating that these sites were highly accessible and reproducible between experiments. By comparison, read coverage at peaks in Sets 2 to 7 was visibly reduced indicating that these sites were only partially accessible and in some experiments completely closed. There were also slightly more peaks in Sets 5 to 7 than in Set1 which demonstrated that there were more unique than overlapping peaks between experiments. In general, the ATAC-seq data generated a large number of both highly reproducible and distinct chromatin accessibility sites. Next, there were 6,223 and 50,629 DNase-seq peaks which overlapped all three ESC and MEF experiments, respectively (see Figure 5.17). One possible reason for the disparity in numbers, is that peak calling in the ESC experiments was not optimal, evidenced by the fact that peaks in Sets 5 to 7 display read coverage in the experiments where a peak had for some reason not been called. Unexpectedly, peaks from two of the ESC experiments formed nearly a complete subset of the third experiment. Apart from unknown laboratory variation, the only variable which may account for this result is library size. However, the total

number of reads was comparable (~30 M per experiment) and small differences were unlikely to have had such a profound effect on peak calling. Similar to the ATAC-seq data, peaks in Set 1 were highly accessible and reproducible compared to peaks in Sets 2 to 7 which were more closed. Given such discrepancy between the ESC and MEF experiments it was more difficult to conclude anything from the DNase-seq data, apart from that the total number of peaks was slightly fewer than the ATAC-seq data. Lastly, there were 1,965 and 481 FAIRE-seq peaks which overlapped all three ESC and MEF experiments, respectively (see Figure 5.18). Worryingly, the total number of peaks for each experiment was drastically different and hardly any peaks overlapped between experiments. This result could not be attributed to imprecise peak calling as read coverage over peaks in Sets 5 to 7 was only visible from the experiment where the peak had been called. Additionally, read coverage over peaks in Sets 5 to 7 was remarkably high which indicated that these sites were apparently highly accessible but irreproducible. The large discrepancy between experiments could also not be attributed to library size because experiments with the largest number of reads sometimes had the fewest number of peaks. Obviously, the FAIRE-seq data was highly irreproducible and chromatin accessibility sites identified from such experiments should be carefully interpreted.

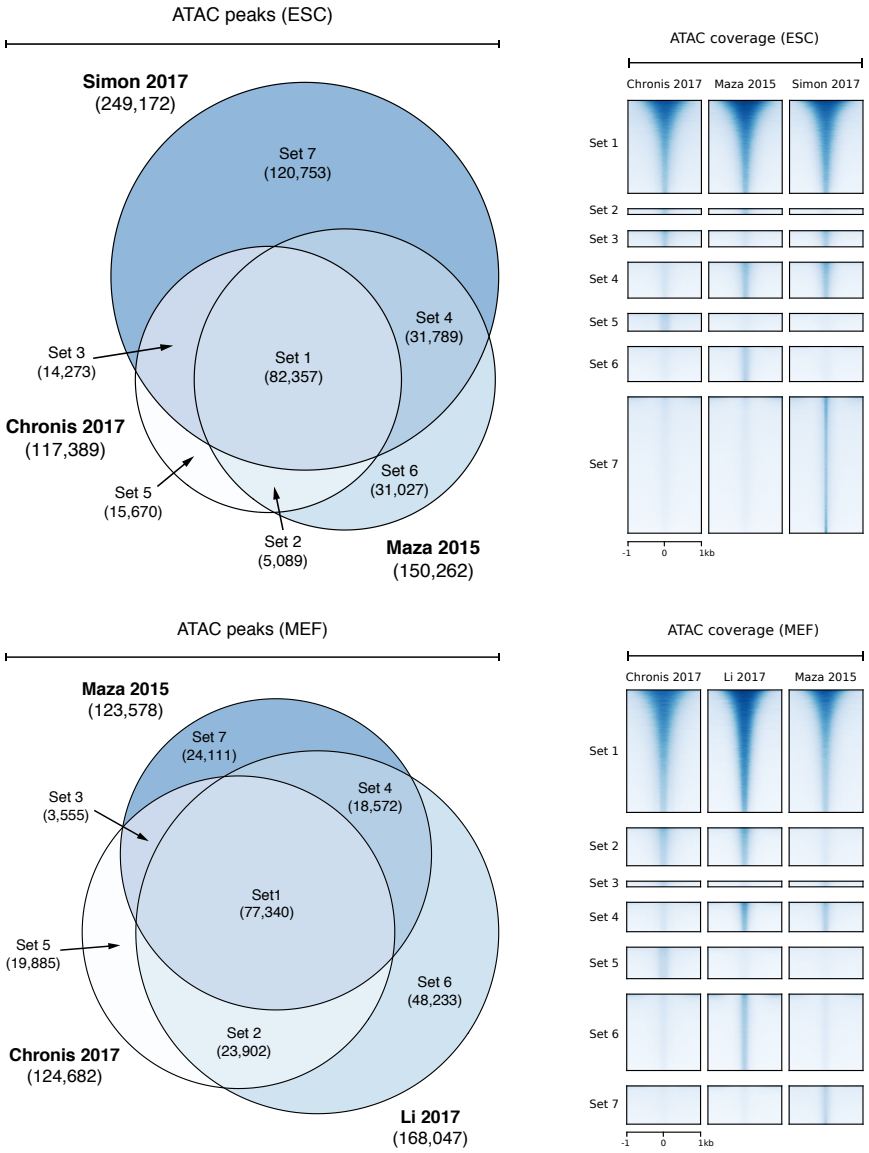

FIGURE 5.16. Comparison of ATAC-seq peaks from ESC and MEF experiments.

Euler diagrams on the left hand side represent the overlap between ATAC-seq peaks from ESC and MEF experiments. Heatmaps on the right hand side display chromatin accessibility from ATAC-seq data.

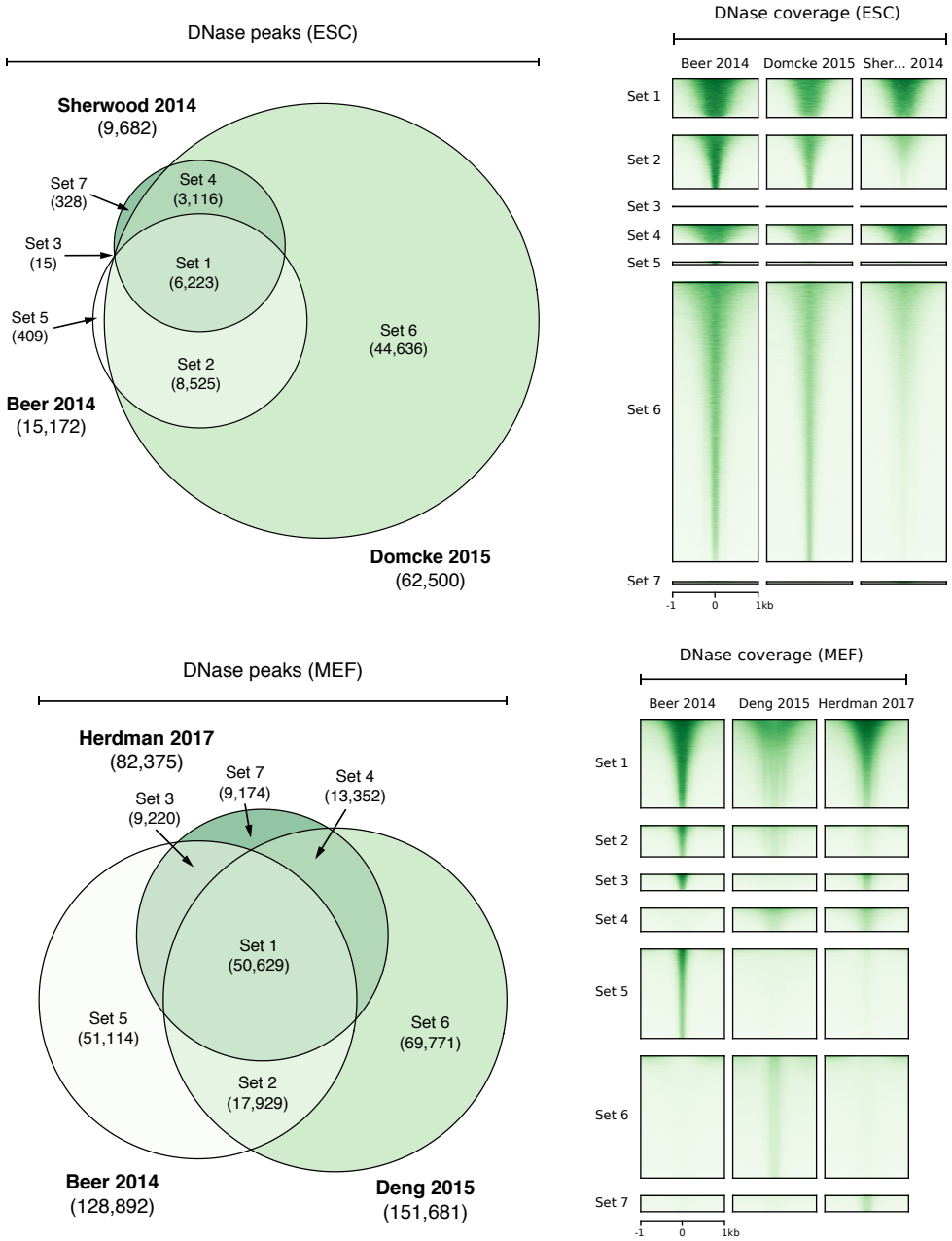

FIGURE 5.17. Comparison of DNase-seq peaks from ESC and MEF experiments.

Euler diagrams on the left hand side represent the overlap between DNase-seq peaks from ESC and MEF experiments. Heatmaps on the right hand side display chromatin accessibility from DNase-seq data.

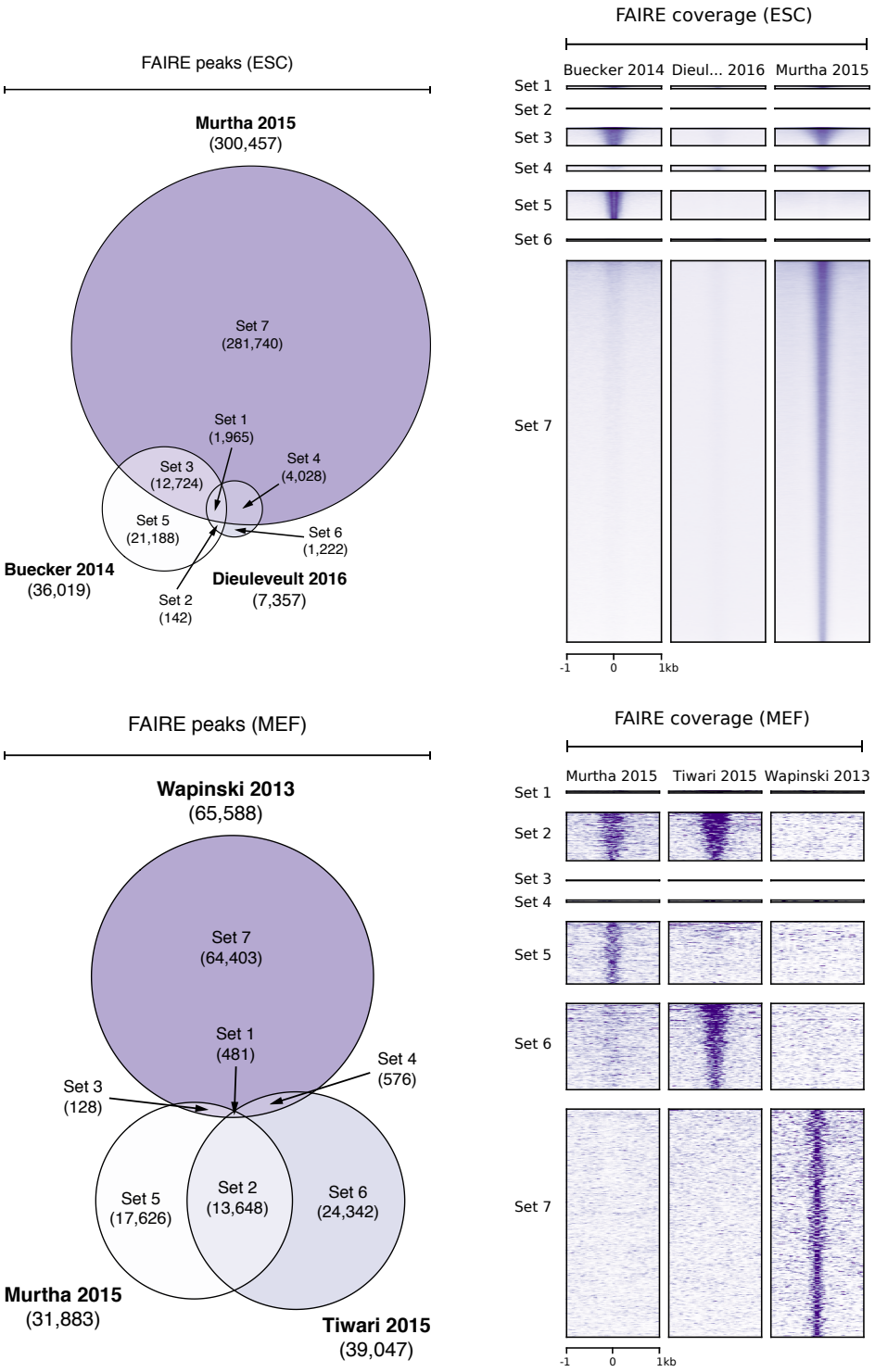

FIGURE 5.18. Comparison of FAIRE-seq peaks from ESC and MEF experiments.

Euler diagrams on the left hand side represent the overlap between FAIRE-seq peaks from ESC and MEF experiments. Heatmaps on the right hand side display chromatin accessibility from DNase-seq data.

Having investigated the level of variability within assays, the number of overlapping and unique peaks between assays was then considered. Peaks for each assay were generated from Sets 1 to 4 (peaks in more than two experiments) because in some cases the third outlier experiment still exhibited moderate read coverage even if a peak had for some reason not been called. As described previously, groups of overlapping and unique peaks were enumerated: Set 1 refers to peaks in all three assays; Sets 2 to 4 refer to peaks in two out of three assays; and Sets 5 to 7 refer to peaks in only one assay. Approximately 7,543 and 14,001 peaks overlapped all three assays from ESC and MEF experiments, respectively (see Figure 5.19). This relatively small number was due to very few overlapping peaks between FAIRE-seq experiments, and consequently a much larger number of peaks overlapped between the ATAC-seq and DNase-seq assays. The majority of DNase-seq and FAIRE-seq peaks also formed a subset of the ATAC-seq peaks, which suggested that ATAC-seq is either more sensitive or less accurate than the other assays. To determine whether the relatively large number of unique ATAC-seq peaks was caused by the peak generation strategy, the number of overlapping and unique peaks between assays was calculated using peaks generated from Sets 1 to 7 (peaks in at least one experiment) (see Figure 5.20). The large number of unique ATAC-seq peaks did not decrease dramatically, but rather the number of unique peaks for all assays increased. To identify hallmark differences between the peaks from each assay, read coverage and histone modification of overlapping and unique peaks was examined (see Figures 5.21 and 5.22). As expected, the read coverage of peaks from Set 1 was higher than all others which indicated that these sites were highly accessible and reproducible between assays. This set of peaks also contained H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H3K27ac modifications which together mark accessible regulatory regions such as transcription factor binding sites, transcribed genes, active promoters, and active enhancers (Wang, Li, and

Hu, 2014; Liu et al., 2016; Karmodiya et al., 2012; Creyghton et al., 2010). Genomic annotation of the peaks likewise showed that reproducible peaks between assays were mostly located at promoters and within gene bodies, where as peaks from other sets were predominantly found within introns and intergenic regions (see Figure 5.23). Encouragingly, there was substantial read coverage from DamID-seq at peaks from Sets 1 to 5 which provided further evidence that Dam can measure chromatin accessibility. Unlike FAIRE-seq, both ATAC-seq and DNase-seq specific peaks exhibited partial read coverage and H3K4me1, H3K4me2, and H3K27ac modifications which suggests that these two assays are more sensitive than FAIRE-seq and on average capture a larger number of partially open chromatin regions. Together, these results demonstrated that there are substantial differences in the number of peaks called between assays, and that truly reproducible chromatin accessibility sites are a minority which exhibit high read coverage and relevant histone modifications. Importantly, it is up to the researcher to determine whether or not they are interested in partially open sites which would be irreproducible between multiple experiments. Whilst it is difficult to fully determine causative differences between assays from public sequencing data, it is clear that FAIRE-seq is not as accurate or reproducible as ATAC-seq or DNase-seq and will no longer considered.

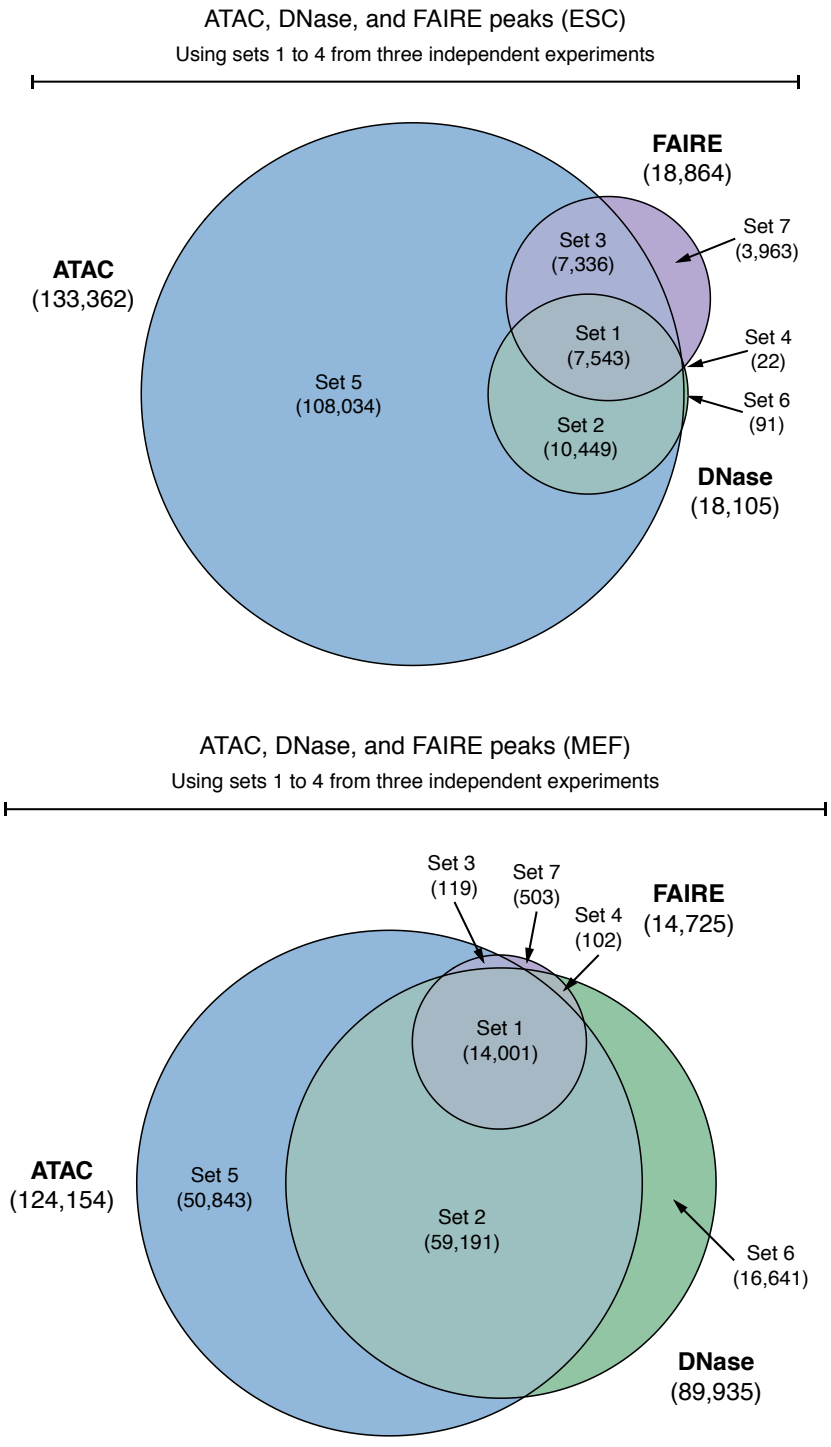

FIGURE 5.19. Overlap between ATAC-seq, DNase-seq and FAIRE-seq peaks in ESCs and MEFs using peaks in more than one experiment.

Euler diagrams represent the overlap between ATAC-seq, DNase-seq, and FAIRE-seq peaks from ESC and MEF experiments. Assay-specific peaks were generated using peaks in more than one experiment (Sets 1 to 4).

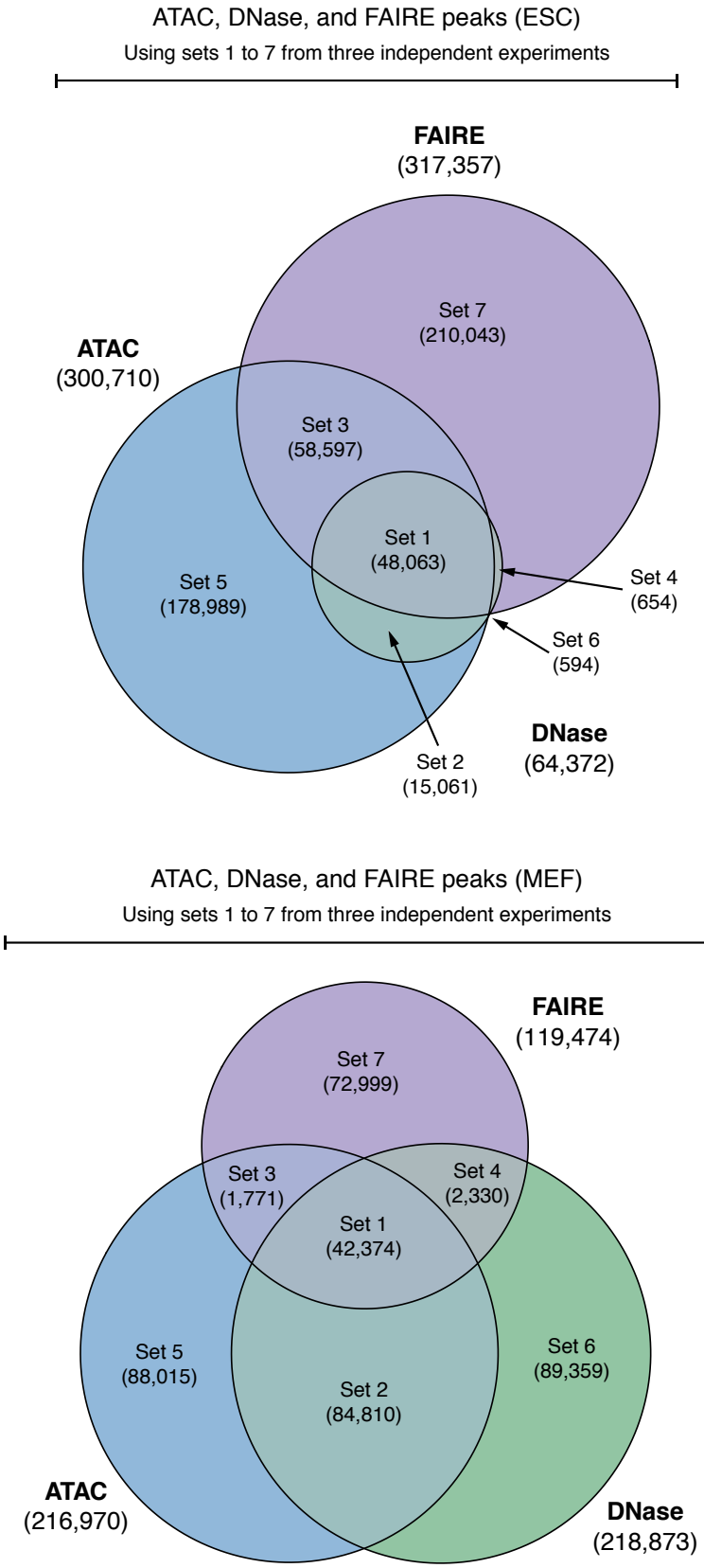

FIGURE 5.20. Overlap between ATAC-seq, DNase-seq and FAIRE-seq peaks in ESCs and MEFs using peaks in one or more experiments.

Euler diagrams represent the overlap between ATAC-seq, DNase-seq, and FAIRE-seq peaks from ESC and MEF assays. Assay-specific peaks were generated using peaks present in one or more experiments (Sets 1 to 7).

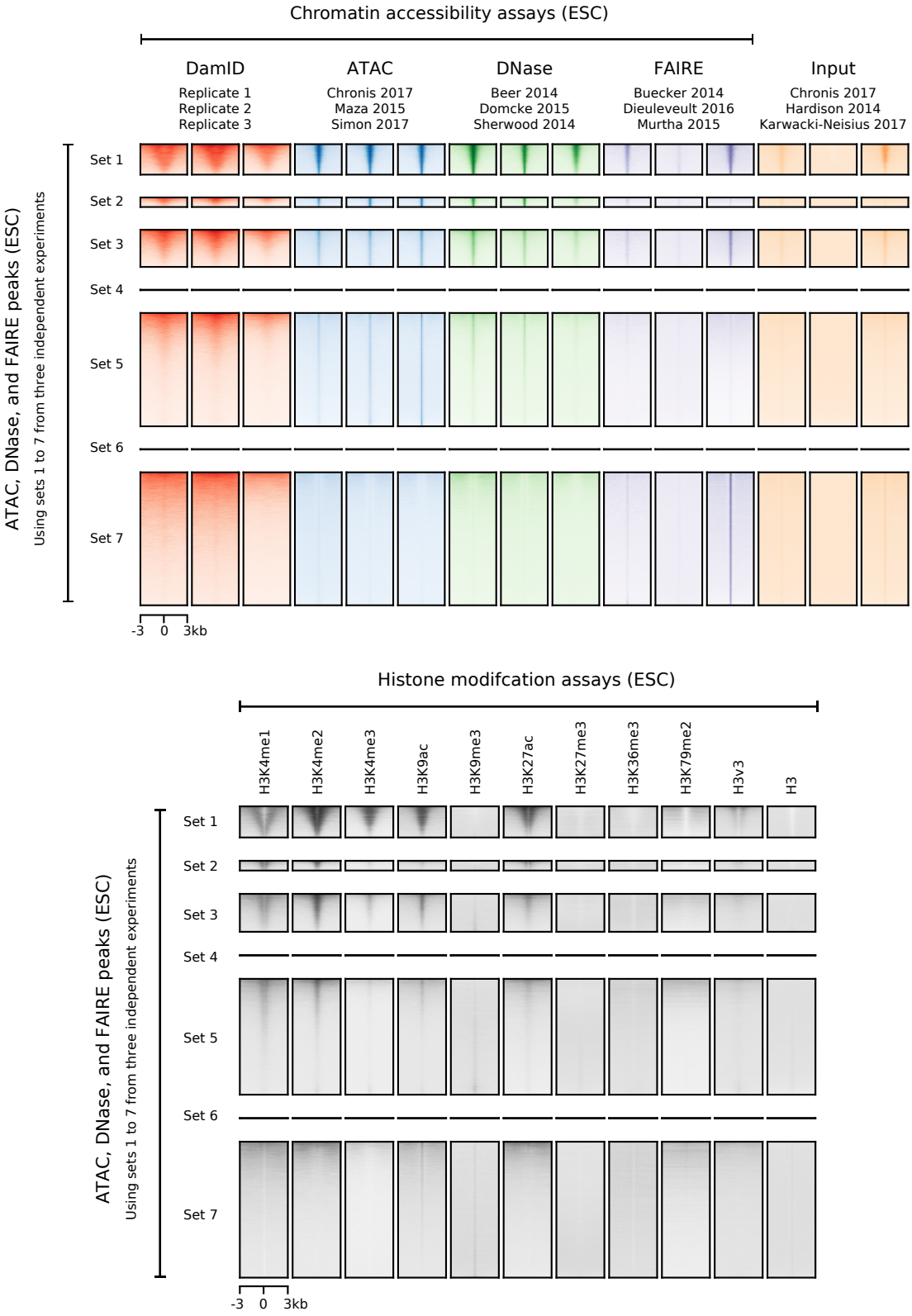

FIGURE 5.21. Chromatin accessibility and histone modifications in ESCs at overlapping ATAC-seq, DNase-seq, and FAIRE-seq peaks.

The top panel contains a heatmap illustrating DamID-seq, ATAC-seq, DNase-seq, and FAIRE-seq signal at overlapping and unique ATAC-seq, DNase-seq, and FAIRE-seq peaks. The bottom panel contains a heatmap illustrating histone modification ChIP-seq signal at the same peak regions.

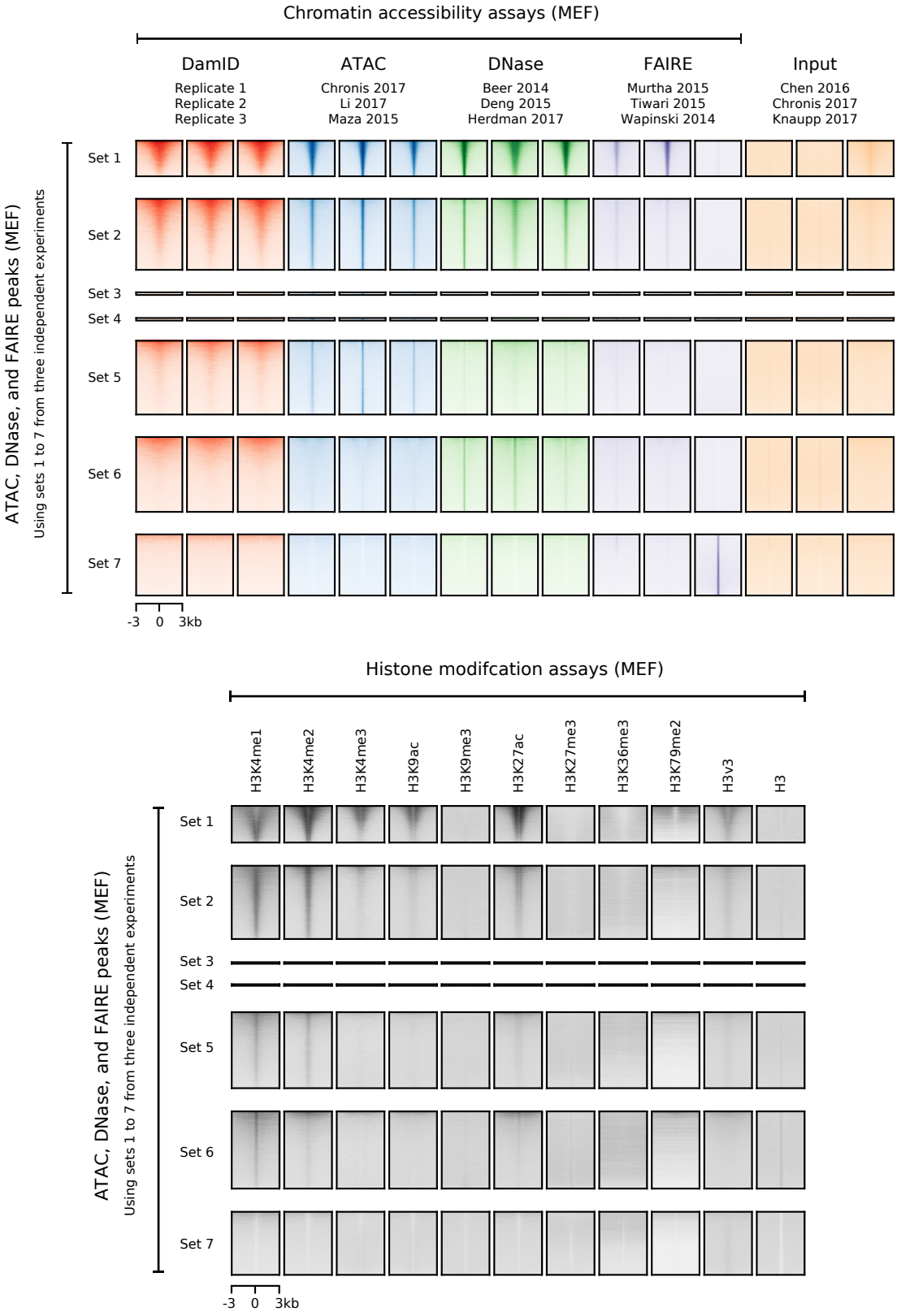

FIGURE 5.22. Chromatin accessibility and histone modifications in MEFs at overlapping ATAC-seq, DNase-seq, and FAIRE-seq peaks.

The top panel contains a heatmap illustrating DamID-seq, ATAC-seq, DNase-seq, and FAIRE-seq signal at overlapping and unique ATAC-seq, DNase-seq, and FAIRE-seq peaks. The bottom panel contains a heatmap illustrating histone modification ChIP-seq signal at the same peak regions.

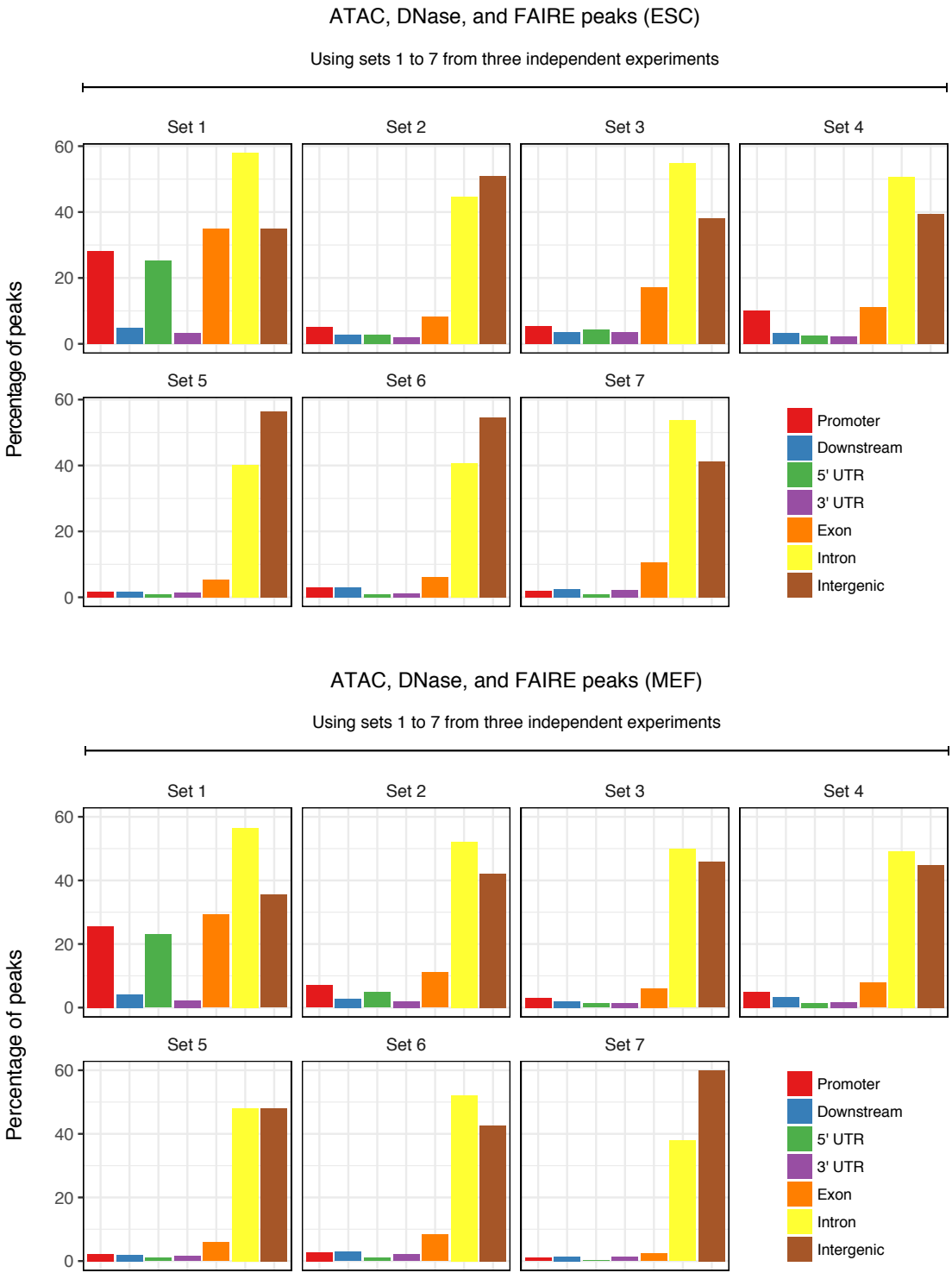

FIGURE 5.23. Annotation of overlapping ATAC-seq, DNase-seq, and FAIRE-seq peaks from ESC and MEF experiments.

Genomic annotation of chromatin accessibility peaks as promoter, downstream of gene end, 5′ untranslated region, 3′ untranslated region, exon, intron, or intergenic.

5.4.3 Identification of accessibility sites from DamID-seq data

Having generated a reliable set of chromatin accessibility sites and investigated the expected level of variability between already established assays, a method to call peaks from DamID-seq data was now required. First, a differential methylation analysis between the Dam libraries and simulated libraries representing the background level of methylation was performed (see Figure 5.24). To generate the simulated libraries, restriction fragment read counts were calculated by measuring the average number of reads aligned to neighbouring restriction fragments within a range of window sizes (e.g. 5 kb, 10 kb) from the Dam libraries. This technique is commonly used in single-sample peak calling algorithms (e.g. MACS2) because it protects against local fluctuations in read coverage due to chromatin structure, PCR amplification, and genome copy number variation (Zhang et al., 2008). To identify significantly methylated restriction fragments, differential methylation analysis using linear modelling and empirical Bayes methods was performed (Ritchie et al., 2015). Importantly, the relationship between the Dam libraries and the simulated libraries was explicitly included in the linear modelling strategy (see Table 5.1). This approach is conceptually similar to Reduced Representation Bisulfite Sequencing (RRBS) data where methylated and unmethylated counts generated from the same sample are modelled with the same dispersion but different means (Chen et al., 2017). Restriction fragments within 1 kb of each other were then stitched together and a Fisher's combined P value and multiple testing corrected P value were calculated (Lun and Smyth, 2016). The method described here was implemented in the Daim software package and can be tested by following the online manual and using the example data provided.

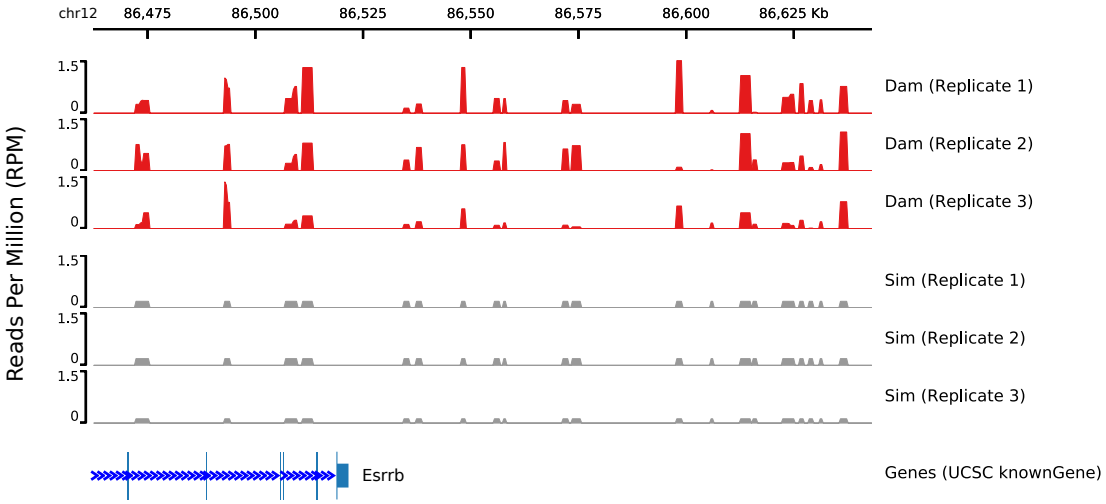

FIGURE 5.24. Snapshot of Dam and simulated libraries in ESCs at the *Esrrb* locus.

For each Dam library, a simulated library representing the background level of methylation was generated. Differential methylation analysis between the Dam and simulated libraries was performed.

Rep1	Rep2	Rep3	Dam
1	0	0	1
1	0	0	0
0	1	0	1
0	1	0	0
0	0	1	1
0	0	1	0

TABLE 5.1. Design matrix for assessing differential methylation between Dam and simulated libraries.

The table represents a design matrix used with linear modelling to assess differential methylation between the Dam and simulated libraries. The first three coefficients are used to model the total number of reads (Dam or simulated) for three replicates. The fourth coefficient is used to to estimate the log ratio of Dam methylation to simulated background methylation.

In order to evaluate the peak calling strategy, a comparison of sites identified from

DamID-seq and other chromatin accessibility assays from ESC and MEF experiments was performed. Approximately 120,597 and 91,306 significant peaks (FDR < 0.1) were called using Daim for the ESC and MEF DamID-seq data generated with 10^6 cells, respectively (see Table 5.2). These numbers were comparable to the number of significant peaks (FDR < 0.1) called using MACS2 for the ATAC-seq, DNase-seq, and FAIRE-seq experiments (see Table 5.3). Genome browser tracks showed that the DamID-seq peak calls were appropriately located under the areas of Dam methylation and were in similar positions to the ATAC-seq, DNase-seq, and FAIRE-seq peak calls (see Figures 5.25 and 5.26). One noticeable difference was that the DamID-seq peak calls were much larger (100 bp to 10,000 bp) than those from the other chromatin accessibility assays (100 bp to 1,000 bp) (see Figure 5.27). The most likely explanation for this is that DamID-seq peaks were called from restriction fragments, where as peaks from the other assays were called using MACS2, which divides the genome into short overlapping windows (Zhang et al., 2008). The approach by MACS2 would not be suitable for DamID-seq data because the assay is performed by digesting the DNA library with methylation-sensitive enzymes which cut at GATC restriction sites, it does not methylate each individual nucleotide along the genome. Although the genome browser figures appear to show very similar size peaks between the different technologies, it is clear from the distributions that DamID-seq peaks are generally much larger. These larger peaks also tend to be evenly distributed across the genome, and not in particular regions (see Figure 5.28).

Cell	Number	Peaks
ESC	10^6	120,597
ESC	10^4	138,919
ESC	10^3	65,401
ESC	10^2	0
MEF	10^6	91,306

TABLE 5.2. Number of peaks called from DamID-seq data.

This table contains the number of significant peaks (FDR < 0.1) called from DamID-seq data using the Daim software package.

Reference	Assay	Cell	Peaks
Chronis 2017	ATAC-seq	ESC	117,389
Maza 2015	ATAC-seq	ESC	150,262
Simon 2017	ATAC-seq	ESC	249,172
Chronis 2017	ATAC-seq	MEF	124,682
Li 2017	ATAC-seq	MEF	168,047
Maza 2015	ATAC-seq	MEF	123,578
Beer 2014	DNase-seq	ESC	15,172
Domcke 2015	DNase-seq	ESC	62,500
Sherwood 2014	DNase-seq	ESC	9,682
Beer 2014	DNase-seq	MEF	128,892
Deng 2015	DNase-seq	MEF	151,681
Herdman 2017	DNase-seq	MEF	82,375
Buecker 2014	FAIRE-seq	ESC	36,019
Dieuleveult 2016	FAIRE-seq	ESC	7,357
Murtha 2015	FAIRE-seq	ESC	300,457
Murtha 2015	FAIRE-seq	MEF	31,883
Tiwari 2015	FAIRE-seq	MEF	39,047
Wapinski 2013	FAIRE-seq	MEF	65,588

TABLE 5.3. Number of peaks called from ATAC-seq, DNase-seq, and FAIRE-seq data.

This table contains the number of significant peaks (FDR < 0.1) called from ATAC-seq, DNase-seq and FAIRE-seq data using MACS2. The reference column contains the author and year of the sequencing experiment (see Chapter 2 for a full list of accession numbers).

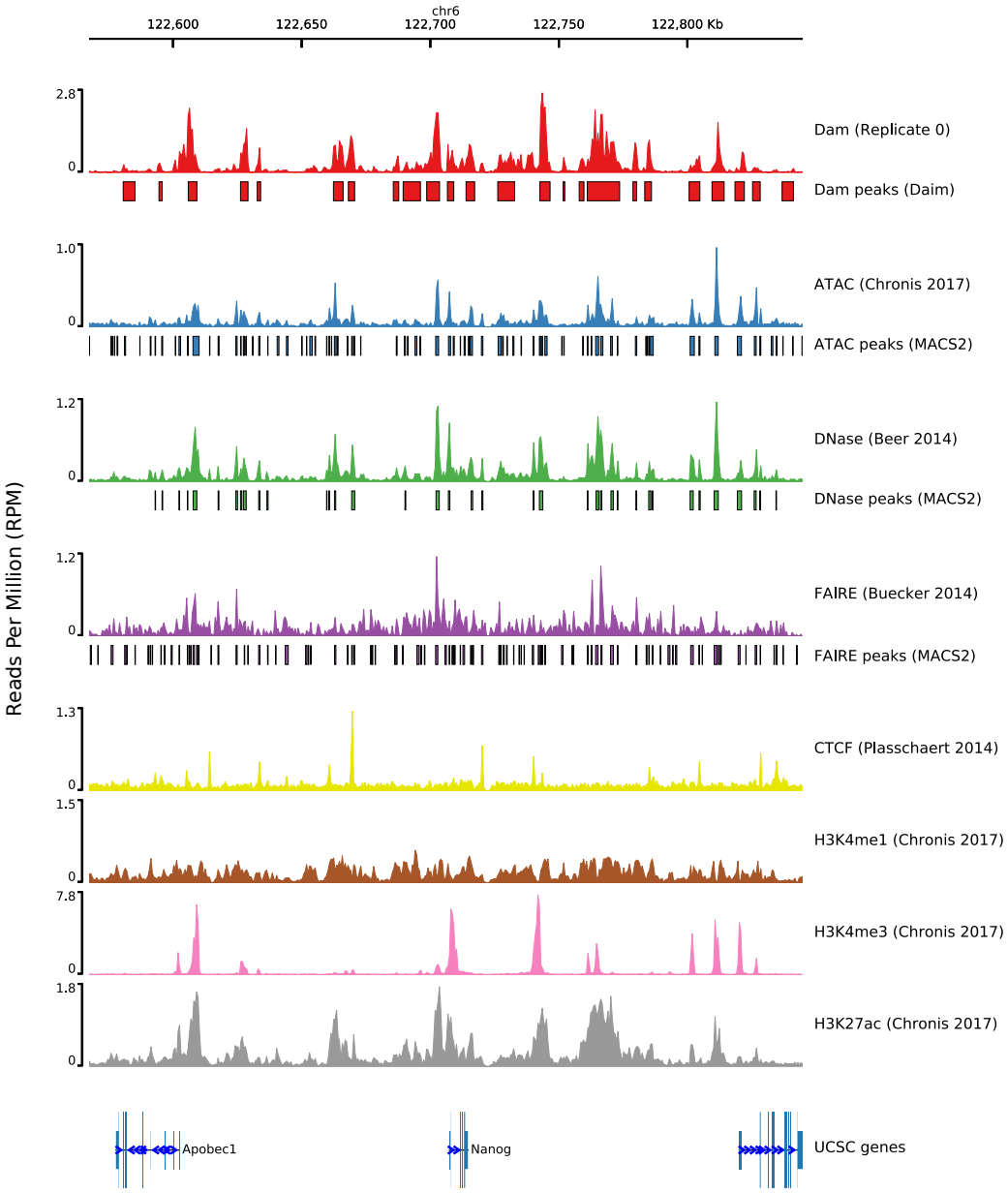

FIGURE 5.25. Genomic snapshot of peaks from ESC chromatin accessibility data.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq peaks. The DamID-seq peaks were called using Daim (FDR < 0.1) and the ATAC-seq, DNase-seq, and FAIRE-seq peaks were called using MACS2 (FDR < 0.1).

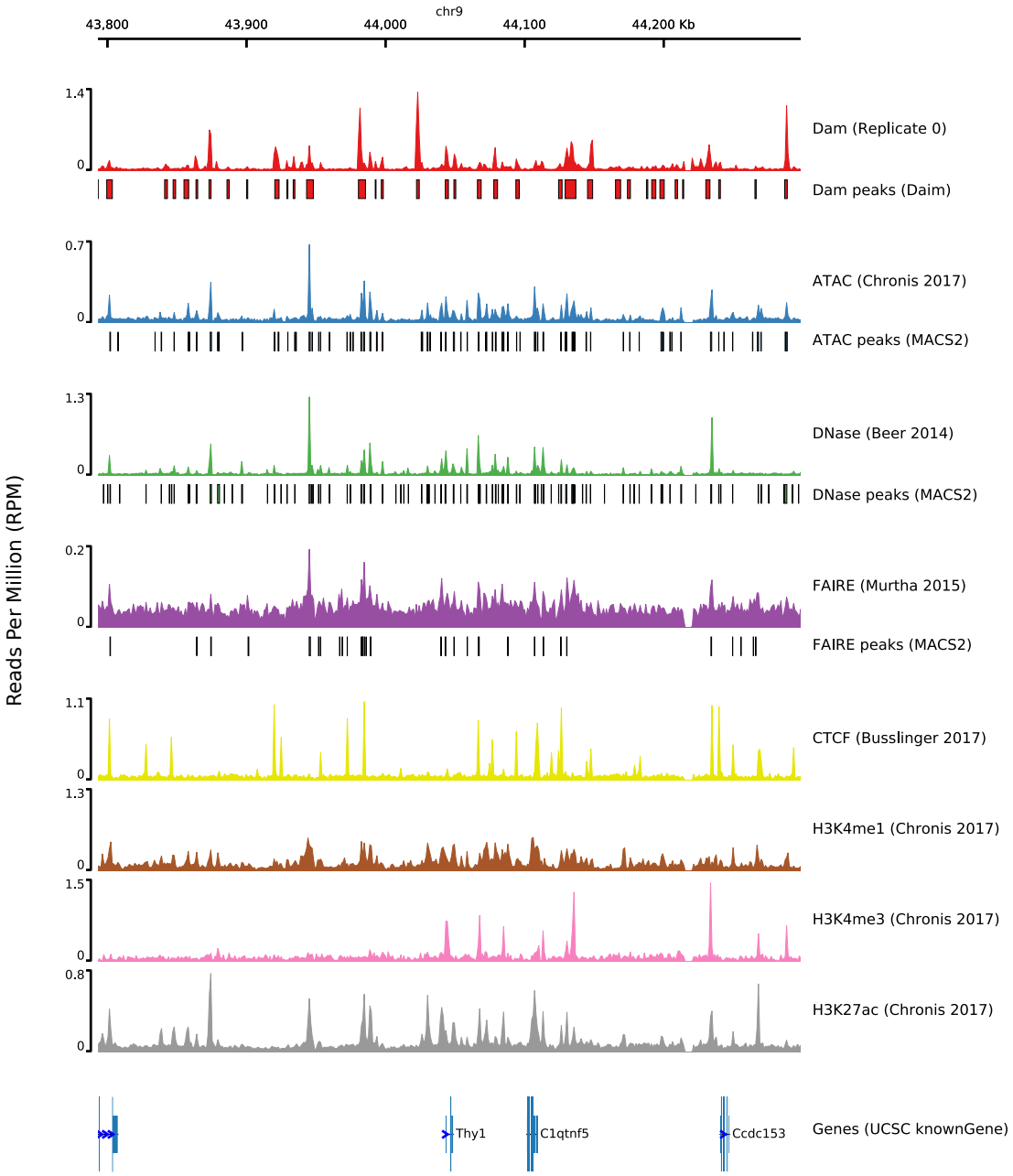

FIGURE 5.26. Genomic snapshot of peaks from MEF chromatin accessibility data.

Comparison of DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq peaks. The DamID-seq peaks were called using Daim (FDR < 0.1) and the ATAC-seq, DNase-seq, and FAIRE-seq peaks were called using MACS2 (FDR < 0.1).

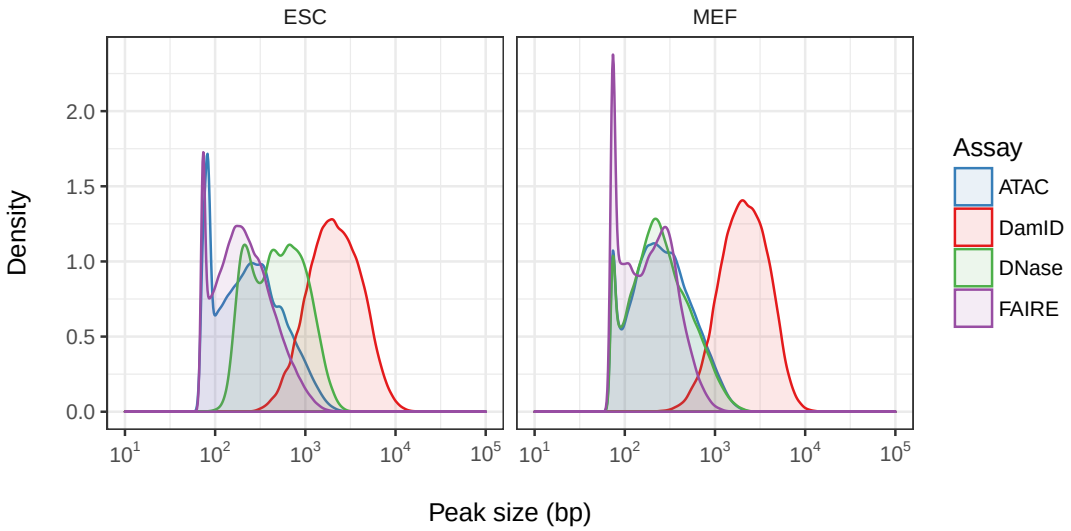

FIGURE 5.27. Distribution of peak sizes from chromatin accessibility assays.

Graphs display the distribution of peak sizes from ESC and MEF ATAC-seq, DamID-seq, DNase-seq, and FAIRE-seq data.

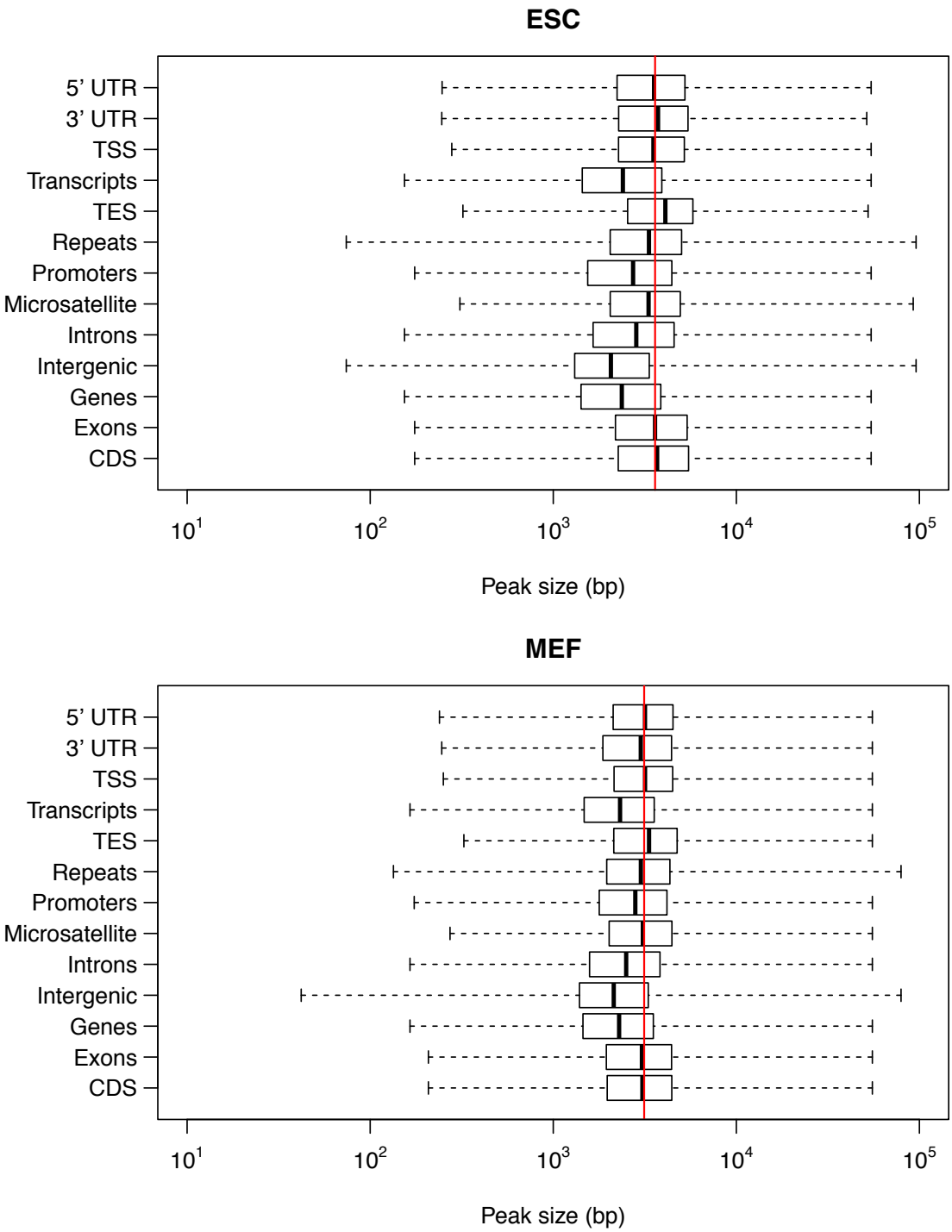

FIGURE 5.28. Distribution of DamID-seq peak sizes by genomic feature.

Boxplots display the distribution of peak sizes by genomic feature from ESC and MEF DamID-seq data. The vertical red line indicates the average peak size.

To assess the similarity between DamID-seq and other chromatin accessibility assays, the number of overlapping and unique peaks was calculated using peaks from Sets 1 to 4 between experiments (see Figure 5.29). Approximately 68,191 (50.31%) and 56,054 (54.23%) DamID-seq peaks overlapped with at least one other assay in the ESC and MEF experiments, respectively. These proportions were similar to the ATAC-seq peaks (56.03% and 71.94%) but were much lower than DNase-seq peaks (99.59% and 86.46%). Interestingly, very few peaks uniquely overlapped between DamID-seq and DNase-seq in both the ESC and MEF experiments which suggests that chromatin accessibility sites identified by ATAC-seq and DNase-seq are more related. In addition, the number of unique ATAC-seq and DamID-seq peaks were similar in both the ESC (58,705 and 67,356) and MEF (34,612 and 47,313) experiments. To verify that the rather large number of unique DamID-seq peaks were genuine, the proportion of overlapping and unique peaks was calculated using peaks from Sets 1 to 7 between experiments (see Figure 5.30). However, the number of unique DamID-seq peaks decreased only slightly in both the ESC (67,356 to 49,141) and MEF (47,313 to 34,687) experiments. In order to identify hallmark differences in the overlapping and unique DamID-seq peaks, read coverage and histone modification data was examined (see Figure 5.31). The read coverage data showed that peaks found in all assays were highly accessible, which indicated that the strongest peaks were also the most reproducible. Interestingly, peaks which were found in either DamID-seq and ATAC-seq only (Set 2) or DamID-seq and DNase-seq only (Set 3) were measured as highly accessible in the DamID-seq assay but only marginally accessible in the ATAC-seq and DNase-seq assays. Likewise, peaks found in ATAC-seq and DNase-seq only were as accessible as the peaks found in all three assays. The histone modification data showed that peaks found in all three assays were highly enriched for H3K4me1, H3K4me2, H3K4me3, H3K9ac, and H3K27ac which are commonly associated with transcribed genes, active promoters, and active enhancers (see Figure 5.32). Interestingly, peaks which were

unique to DamID-seq lacked any histone modifications and peaks which were shared between DamID-seq and just one other assay had a reduced H3K4me3 and H3K9ac level. Lastly, the distribution of overlapping and unique DamID-seq peaks within genomic features was examined (see Figure 5.33). Generally, the genome annotations showed that all sets of peaks were predominantly found in intergenic and intron regions, but that peaks common to all assays were also to a large extent located within promoter regions. Worryingly, peaks which were unique to ATAC-seq and DNase-seq were also found at promoter regions which suggests that DamID-seq is unable to detect chromatin accessibility sites at certain promoters. However, as discussed previous (see Subsection 4.4.3) this difference could be explained by the depletion of GATC sites in promoter regions in the mouse genome (see Figure 4.36) which may explain why fewer accessibility sites were detected at promoters using DamID-seq data. Together these results indicate that high coverage DamID-seq peaks predominantly overlap with high coverage peaks from other chromatin accessibility assays, however unlike ATAC-seq and DNase-seq, unique DamID-seq peaks are not enriched for histone modifications associated with accessible genomic features. Whether these are true accessibility sites or are simply a technical artefact remains to be understood.

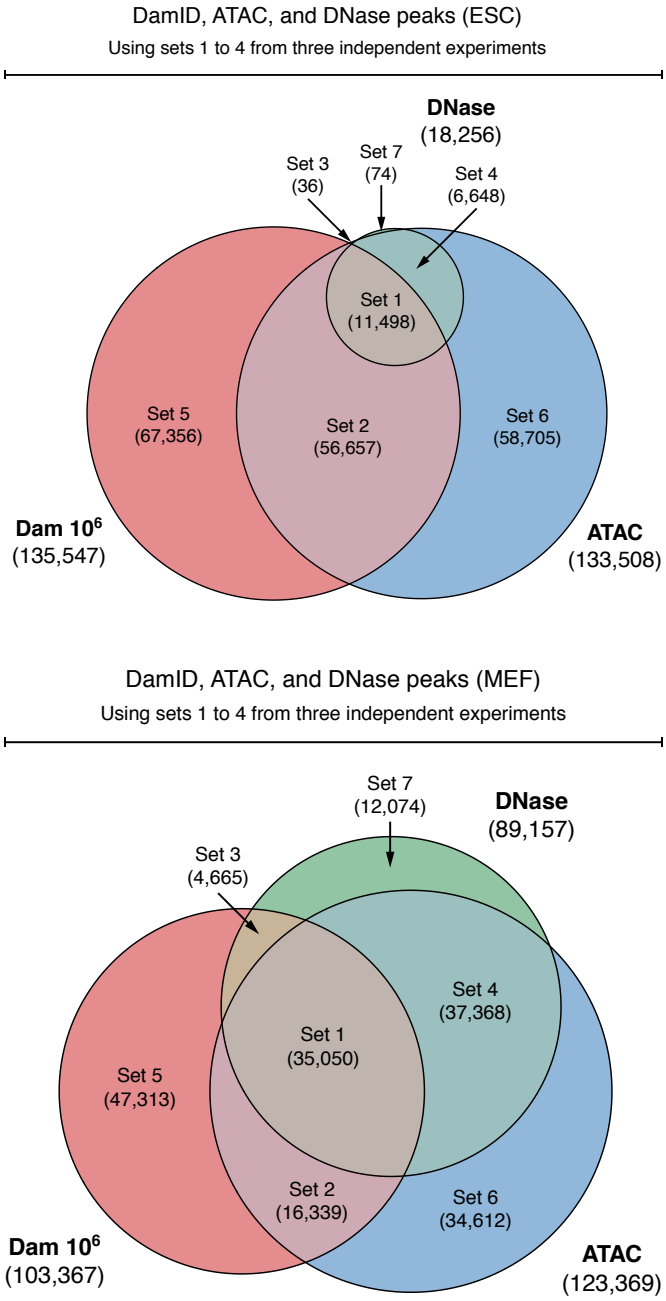

FIGURE 5.29. Overlap between DamID-seq, ATAC-seq, and DNase-seq peaks in ESCs and MEFs using peaks in more than one experiment.

Euler diagrams represent the overlap between DamID-seq, ATAC-seq, and DNase-seq peaks from ESC and MEF experiments. The ATAC-seq and DNase-seq peaks were generated using peaks in more than one experiment (Sets 1 to 4).

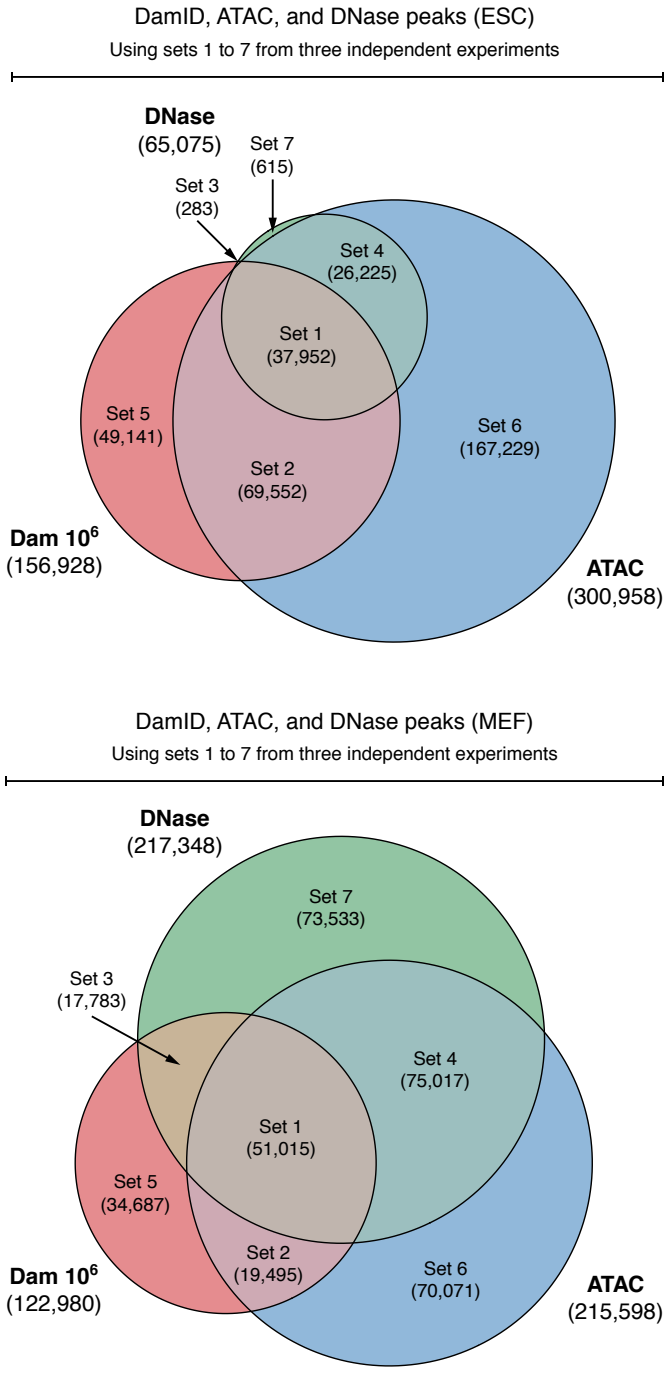

FIGURE 5.30. Overlap between DamID-seq, ATAC-seq, and DNase-seq peaks in ESCs and MEFs using peaks in one or more experiments.

Euler diagrams represent the overlap between DamID-seq, ATAC-seq, and DNase-seq peaks from ESC and MEF experiments. The ATAC-seq and DNase-seq peaks were generated using peaks in one or more experiments (Sets 1 to 7).

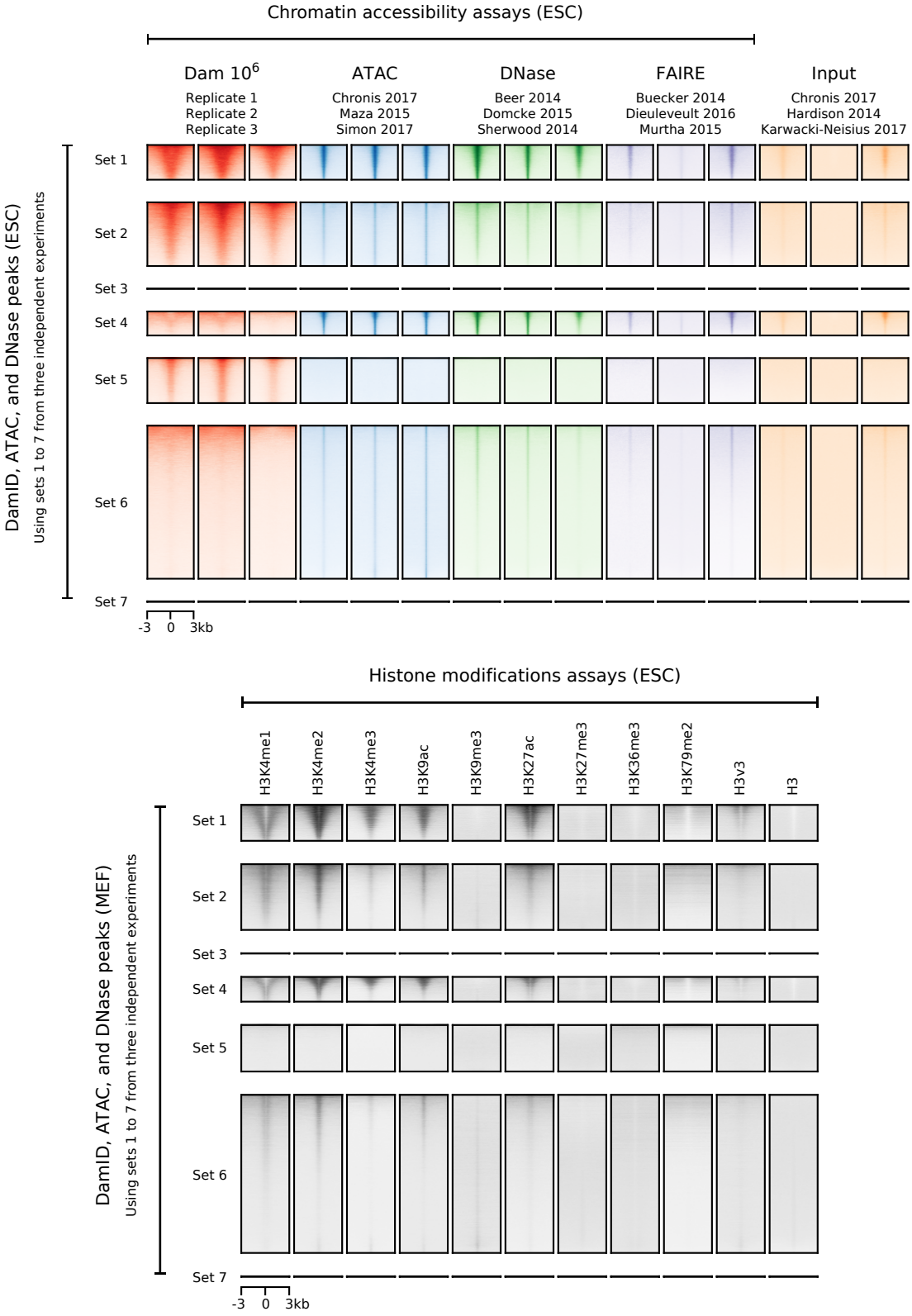

FIGURE 5.31. Chromatin accessibility and histone modifications in ESCs at overlapping DamID-seq, ATAC-seq, and DNase-seq peaks.

The top panel contains a heatmap illustrating DamID-seq, ATAC-seq, DNase-seq, and FAIRE-seq signal at overlapping and unique DamID-seq, ATAC-seq, and FAIRE-seq peaks. The bottom panel contains a heatmap illustrating histone modification ChIP-seq signal at the same peak regions.

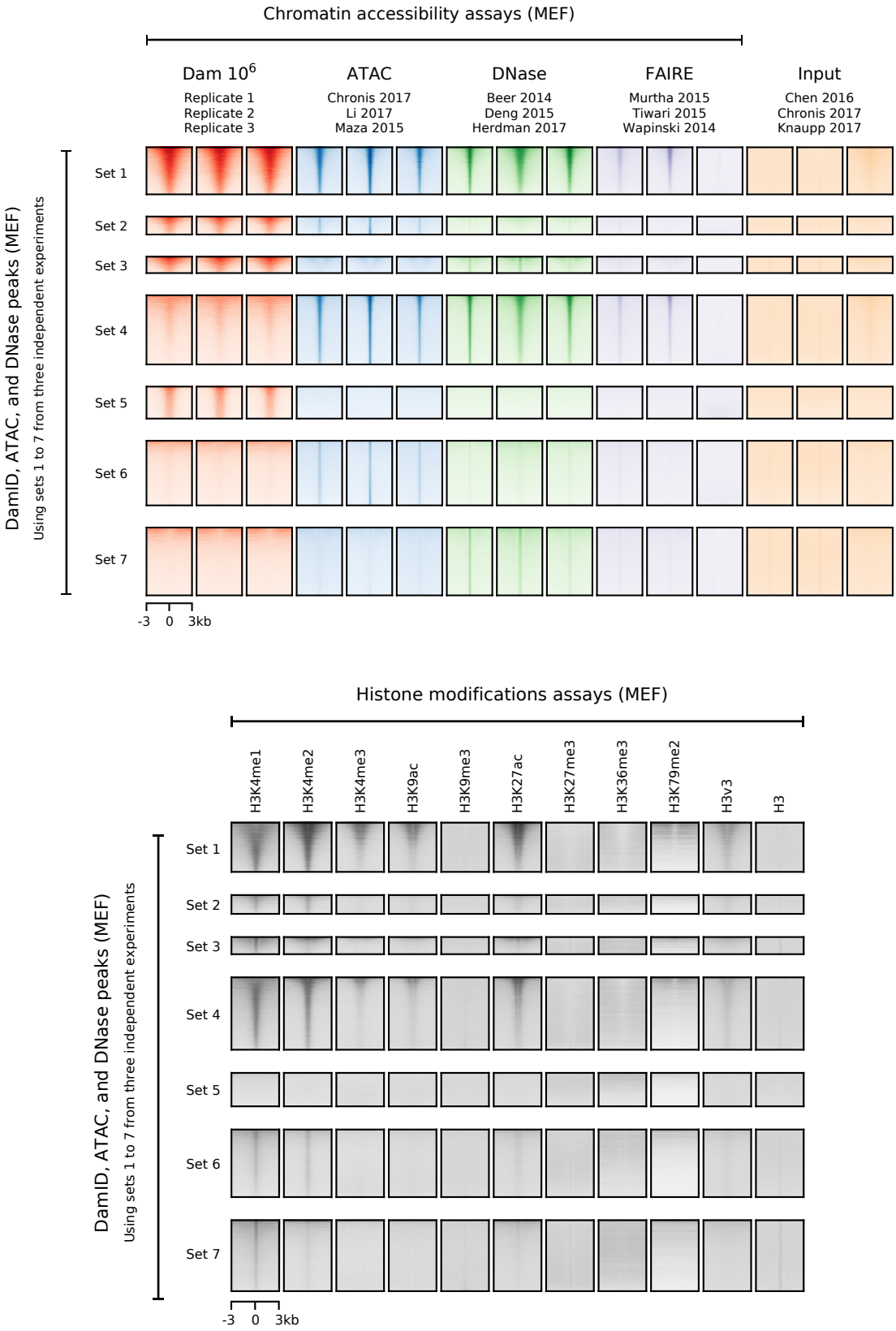

FIGURE 5.32. Chromatin accessibility and histone modifications in MEFs at overlapping DamID-seq, ATAC-seq, and DNase-seq peaks.

The top panel contains a heatmap illustrating DamID-seq, ATAC-seq, DNase-seq, and FAIRE-seq signal at overlapping and unique DamID-seq, ATAC-seq, and FAIRE-seq peaks. The bottom panel contains a heatmap illustrating histone modification ChIP-seq signal at the same peak regions.

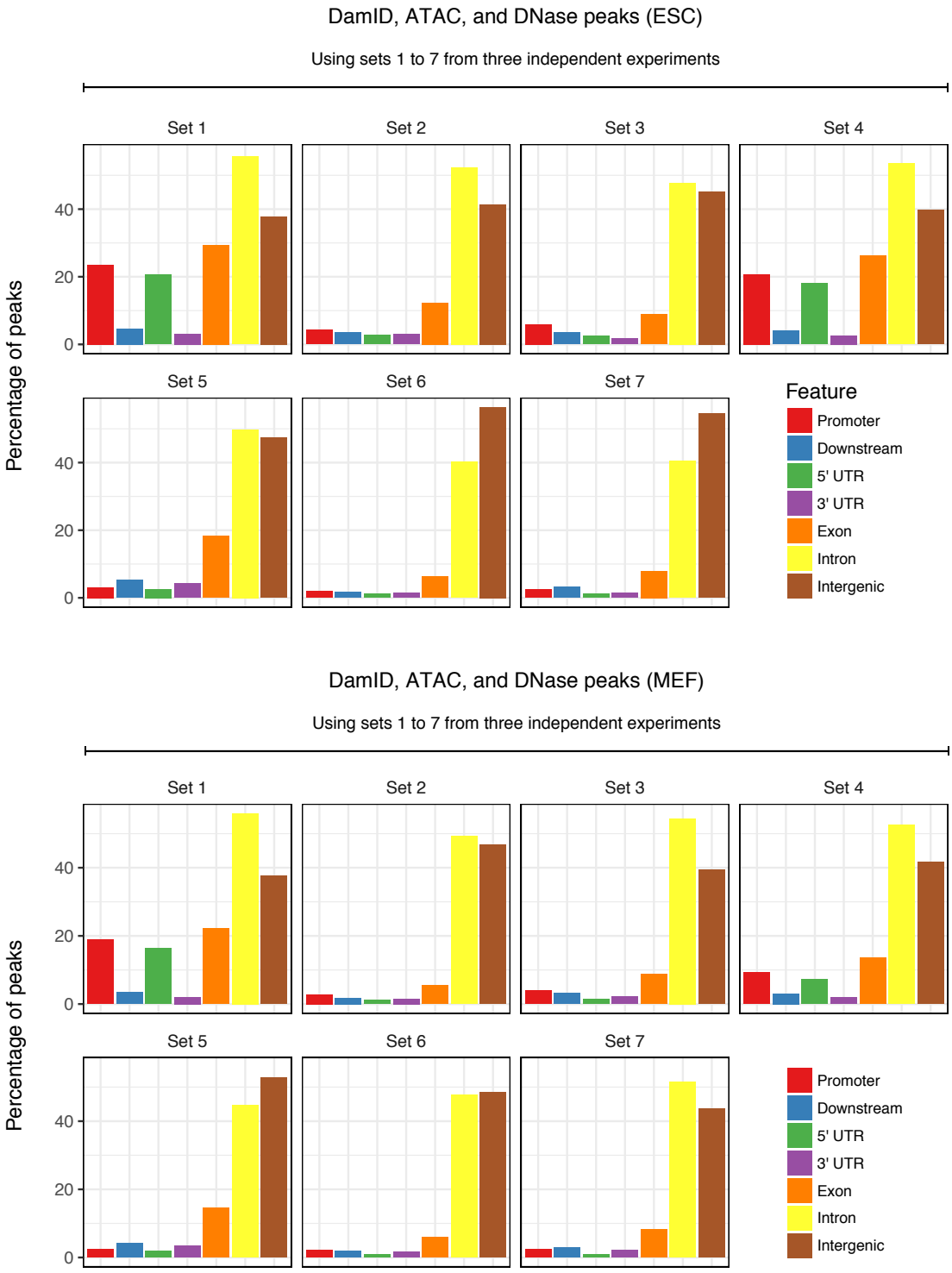

FIGURE 5.33. Annotation of overlapping DamID-seq, ATAC-seq, and DNase-seq peaks from ESC and MEF experiments.

Genomic annotation of chromatin accessibility peaks as promoter, downstream of gene end, 5′ untranslated region, 3′ untranslated region, exon, intron, or intergenic.

Having demonstrated that chromatin accessibility sites can be detected from DamID-seq data generated with 10^6 cells, the sensitivity of DamID-seq at lower cell numbers was then examined. Peaks were called from ESC DamID-seq data generated with 10^4, 10^3, and 10^2 cells using the method implemented in the Daim software package (see Figure 5.34). A similar number of peaks were called using 10^6 (N=120,597) and 10^4 (N=138,919) cells, but this dropped by over half using 10^3 (N=65,401) cells and no significant peaks were detected using 10^2 cells (see Table 5.2). To determine why no significant peaks were detected using 10^2 cells, correlation plots of the restriction fragment read counts between biological replicates were generated (see Figure 5.36). The correlation plots showed that Spearman's rank correlation coefficient decreased as the number of cells decreased (ranging from 0.80 to 0.29) and that past 10^3 cells the correlation between biological replicates was extremely low, hence the variation was too high to achieve a significant probability level (FDR < 0.1). Another possible explanation for the drop in the number of peaks was that as the number of cells decreased the complexity of the libraries also decreased (see Figure 5.35). Reassuringly, the majority of peaks called at lower cell numbers formed a subset of those called from higher cell numbers (see Figure 5.37). Read coverage of the overlapping and unique peaks also demonstrated that the peaks called using lower cell numbers were the strongest accessible peaks called using higher cell numbers, which are also reproducible across assays. Lastly, the distribution of peaks within genomic features was similar between different cell numbers which indicated that peak calling at specific genomic features was not biased by cell number (see Figure 5.38). These results indicated that a large number of chromatin accessibility sites can be detected using a minimum of 10^3 cells, and that the method implemented in the Daim software package is accurate at lower cell numbers.

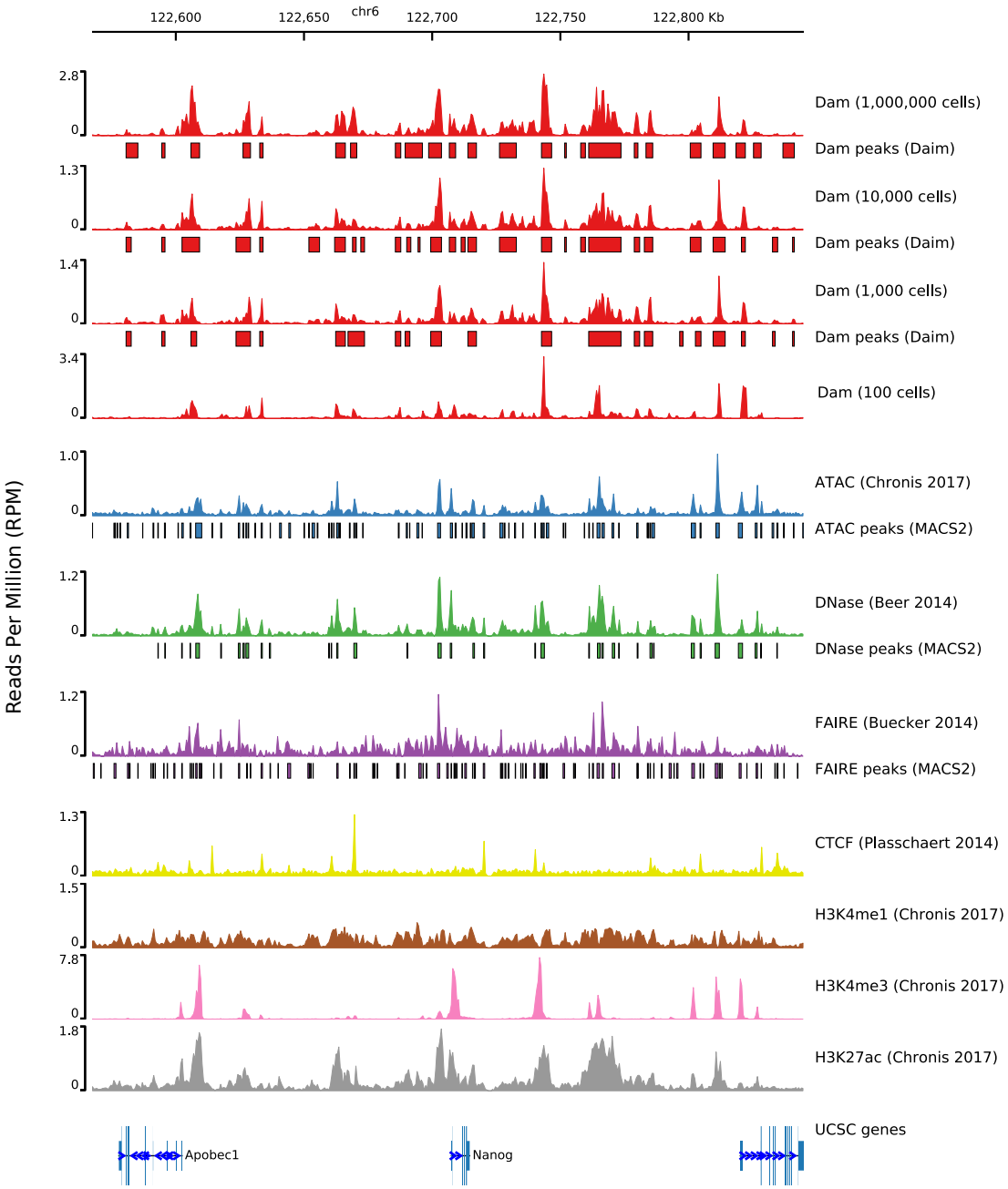

FIGURE 5.34. Genomic snapshot of low cell number DamID-seq peak calls in ESCs at the *Nanog* locus.

Comparison of low cell number DamID-seq with ATAC-seq, DNase-seq, and FAIRE-seq peaks. The DamID-seq peaks were called using Daim (FDR < 0.1). The ATAC-seq, DNase-seq, and FAIRE-seq peaks were called using MACS2 (FDR < 0.1).

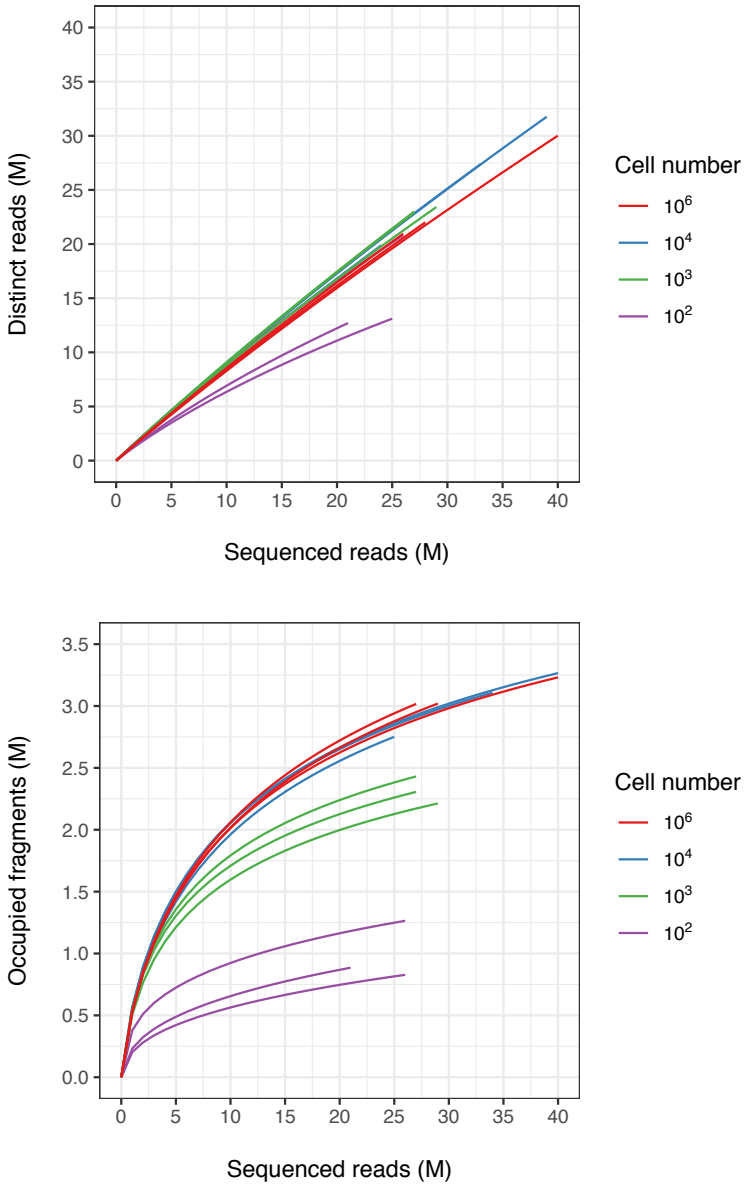

FIGURE 5.35. Library complexity of low cell number ESC DamID-seq data.

The top graph measures the sequencing read complexity at different library sizes. The bottom graph measures the number of restriction fragments occupied at different library sizes.

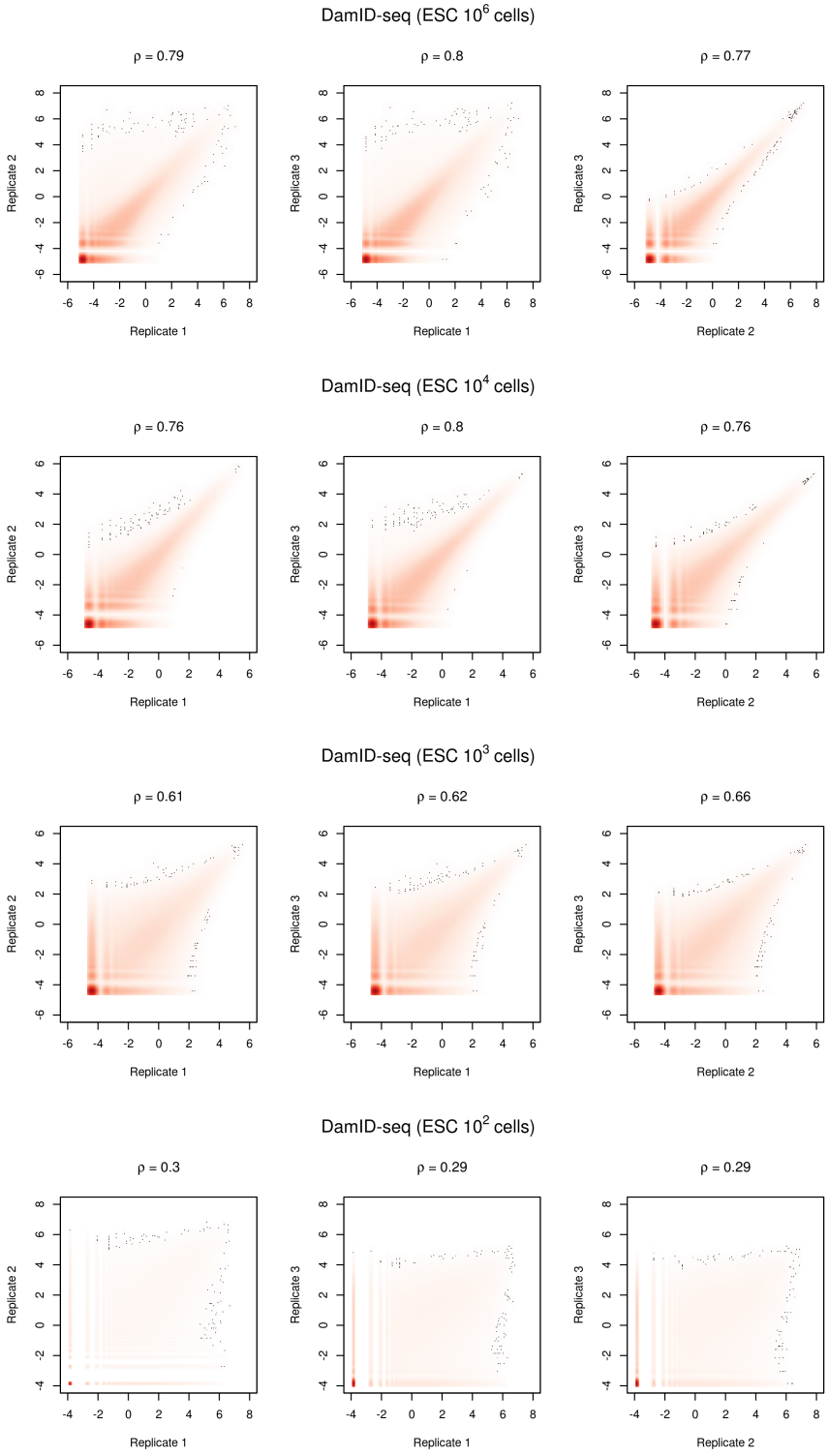

FIGURE 5.36. Spearman correlation between low cell number ESC DamID-seq replicates.

Graphs of Spearman correlations between low cell number DamID-seq biological replicates calculated using the normalised number of counts for each restriction fragment. The correlation coefficient's show that the reproducibility between Dam libraries decreases as the number of cells decreases.

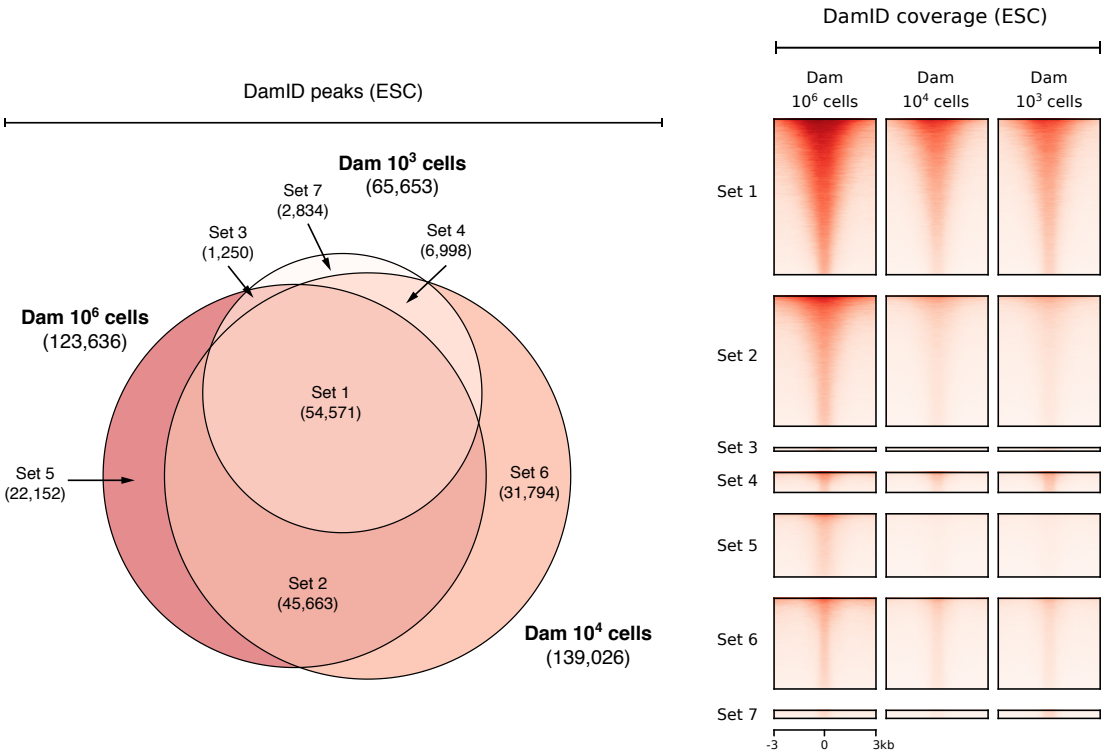

FIGURE 5.37. Comparison of low cell number DamID-seq peaks from ESC experiments.

Euler diagrams on the left hand side represent the overlap between low cell number DamID-seq peaks from ESC experiments. Heatmaps on the right hand side display chromatin accessibility from DamID-seq data.

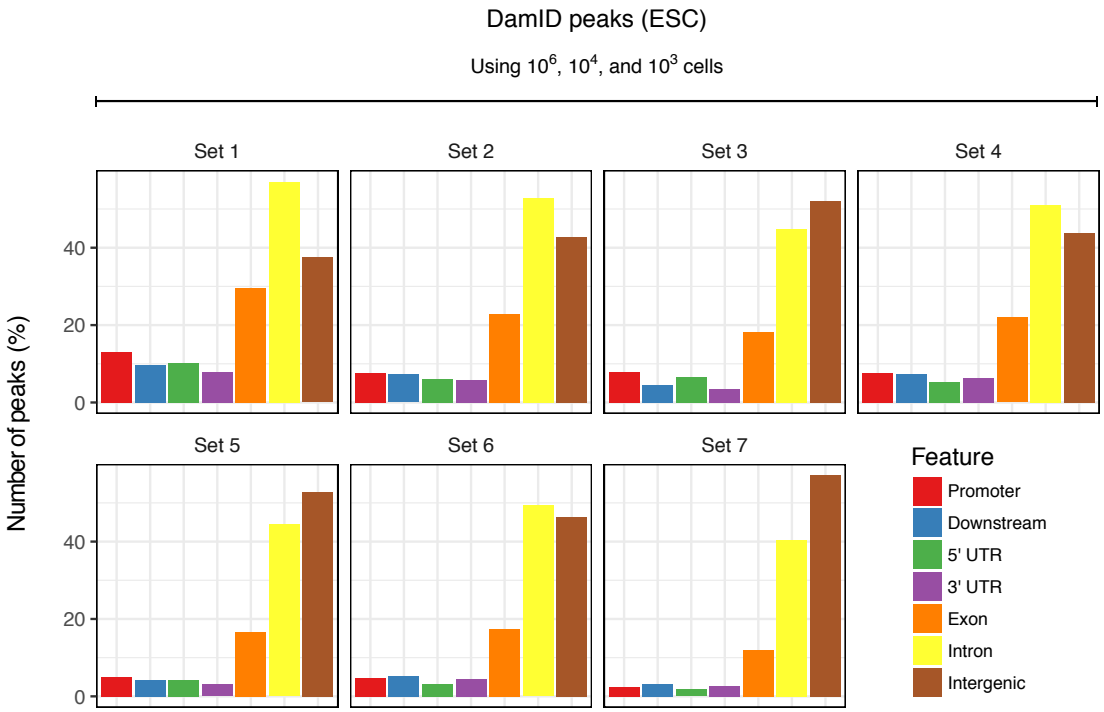

FIGURE 5.38. Annotation of overlapping low cell number DamID-seq peaks ESC experiments.

Genomic annotation of chromatin accessibility peaks as promoter, downstream of gene end,5'untranslated region, 3' untranslated region, exon, intron, or intergenic.

5.5 Discussion

These results show that DamID-seq can be used to measure chromatin accessibility with minimal numbers of cells (10^6, 10^4, 10^3) from a large mammalian genome. This finding complements previous research by providing further evidence that Dam methylates accessible chromatin in a manner similar to already established assays. It also provides the first proof that a large number of reproducible chromatin accessibility sites can be identified from DamID-seq data prepared with 10^3 cells at a minimum. This is especially significant when combined with transgene-drive Dam expression, because it provides researchers limited by starting material with an accurate

and sensitive *in vivo* profiling method. This work also demonstrates that truly reproducible chromatin accessibility sites, based upon comparative analyses across experiments and assays, are in a minority and it is not clear whether they are relevant or important to the biology under investigation. Most of the reproducible sites are located near regulatory genomic features such as promoters and enhancers, whilst the rather large number of unique sites appear to be randomly distributed within intergenic regions. This suggests that many chromatin accessibility sites identified using these assays are not functionally important and instead represent a heterogenous landscape caused either by tissue heterogeneity or chemical handling. In reality, multiple biological and technical factors collude in this regard, and that comparative analysis of public sequencing data is insufficient to thoroughly understand this phenomenon. Despite these complications, it is clear that ATAC-seq outperformed all of the other assays and only in the particular situation where isolation of rare cell types is too challenging should DamID-seq be specifically applied. The resolution of DamID-seq peaks is also very low compared to ATAC-seq, and whilst this is problematic, it is difficult to envisage how an analysis of chromatin accessibility would be particularly biased. Most analyses of this type of data do not rely on nucleotide-level resolution, primarily because chromatin accessibility usually spans multiple kilobases along the genome. Perhaps more concerning, is that FAIRE-seq data is extremely irreproducible and past research relying on this technology should be either carefully interpreted or performed again with a different assay. Lastly, the development of a peak calling method for DamID-seq and its implementation in an open-source software package will allow researchers to readily analyse their own sequencing data, which hopefully will create an impetus for the wide adoption of the technology and allow for future large-scale comparative analyses of chromatin accessibility sites from DamID-seq experiments.

Chapter 6

Discussion

6.1 Summary of research

The results presented in this thesis show that DamID-seq is capable of identifying transcription factor binding sites and chromatin accessibility sites using a minimum of 10^3 cells. In order to reach this conclusion a comprehensive investigation of biases in the experimental procedure, differential methylation analysis of the sequence count data, and similarity between DamID-seq and already established assays was conducted. First, it was demonstrated that technical variables such as polymerase usage and DpnII digestion were important in generating sufficient enrichment of methylated chromatin and that identification of differentially methylated restriction fragments was influenced by sequence composition. Second, the statistical properties of the sequence count data revealed that global adjustment methods were not appropriate for normalisation and that the variance between replicate experiments should be modelled within groups independently. Third, comparative analyses with already established genome-wide methods verified that the target sites identified using DamID-seq were genuine and that with fewer cells only the strongest sites were detected. These findings culminated in the development of Daim – an R/Bioconductor software package for the

analysis, interpretation, and visualisation of DamID-seq data. The availability of this software should help streamline reproducible analyses and alleviate the concerns of researchers looking to perform DamID-seq experiments who lack the relevant training in computational biology.

6.2 Contributions of research

This work contributes significantly to the understanding of DamID-seq data and lays the appropriate theoretical groundwork for development of new tools and interpretation of results:

First, a systematic evaluation of potential biases in the sequencing data revealed that the lower resolution of DamID-seq was limited to multiple kilobases. By comparison the strand-specific nature of ChIP-seq can be used to achieve near base pair resolution. The removal of PCR duplicates was not detrimental to the signal at transcription factor binding sites, and in some cases increased relative to the background. The binding of Dam was not biased by the local nucleotide composition of GATC sites, unlike the Tn5 transposase used in ATAC-seq experiments. The GC content and length of restriction fragments influenced the ability to detect differential methylation, which indicated that downstream corrective measures were required. Lastly, a chromatin state model of Dam binding demonstrated that the protein binds principally at genomic regions associated with chromatin accessibility, such as enhancers and promoters.

Second, the evaluation of distributional properties of the restriction fragment read counts showed that global adjustment methods were not appropriate and instead smooth quantile normalisation was suitable to retain genuine differences between the

Dam and Dam-fusion libraries. Additionally, it was shown that the Dam-fusion libraries were much more variable than the Dam libraries which indicated that the variance of the restriction fragment read counts should be modelled independently for each group when performing differential methylation analysis. Together, these observations informed the development of a peak-calling strategy which was able to identify binding sites from multiple transcription factors and cell lines using a minimum of 10^3 cells.

Third, having observed that Dam preferentially binds genomic regions generally considered to be accessible the opportunity to repurpose the sequencing data to identify chromatin accessibility sites was examined. Through comparison with ATAC-seq, DNase-seq, and FAIRE-seq data it was shown that Dam can be used to measure chromatin accessibility. Given this result, a method to identify accessibility sites from DamID-seq data was described and evaluated through comparative analyses with multiple independent ATAC-seq, DNase-seq, and FAIRE-seq experiments.

6.3 Comparison with previous research

Whilst the results presented in this thesis largely complement and extend research by previous groups, a number of small points of disagreement were discovered:

First, the removal of the DpnII digestion step as advocated by recently published DamID-seq protocols caused a significant reduction in the methylation signal from Dam and Dam-fusion binding. The original role of DpnII was to remove restriction fragments caused by non-specific methylation to help decrease the background in the sequencing data and make peak-calling more accurate and sensitive. However, the conjecture from newer publications is that the occurrence of non-specific methylation is so low that DpnII digestion is redundant and that only DpnI digestion is required.

While this observation may have been true for these publications, it is interesting to note that many of them were investigating nuclear lamin structure and assembly. The interaction of lamin with DNA spans hundreds of kilobases and the analysis of such data typically involves aggregating read counts from large windows so individual reductions in binding may not have been compelling or even noticed. The footprint of transcription factors on the other hand is localised to less than a hundred base pairs and hence DpnII appears to be essential for proper enrichment and removal of the non-specific background.

Second, the use of the Clontech Advantage polymerase used in the original and a number of subsequent DamID-seq protocols was inefficient at amplifying certain restriction fragments and reduced the library complexity. This was particularly important because restriction fragments containing DNA-binding sites were no longer represented in the sequencing library making the assay overall less sensitive. The extent of this problem however is probably dependent on the experimental organism as smaller genomes will on average lose fewer restriction fragments than larger genomes, especially if the average GC content and size of the restriction fragments is suited to the polymerase. In the case of transcription factors and other small footprint DNA-binding proteins, it is essential that as many methylated restriction fragments as possible are amplified in the sequencing library. Each of these restriction fragments may contain a single DNA-binding site, and for each one which has failed to amplify the number of peak calls is reduced. This issue is less concerning however for large scale interactions such as nuclear lamin, where the absence of a few restriction fragments within kilobase pair regions can be tolerated. Future DamID-seq experiments of small footprint DNA-binding proteins should therefore probably substitute the Clontech Advantage polymerase for a more processive one, such as the Kapa HiFi polymerase which was examined.

Third, routine quality control metrics proposed by the ENCODE consortia and the CISTROME project were surprisingly ineffective at predicting the success of a ChIP-seq experiment. Large-scale comparative analysis of published Oct4 ESC ChIP-seq data demonstrated a high level of variability in the number of peaks called between experiments which could not be explained by sequence quality, mapping quality, library complexity, or ChIP enrichment. Whilst many of these metrics in theory are reasonable, they do not consider variation at the experimental level which can be argued is more important to consider. However, cross-referencing of the peaks with known experimental variables such as cell culture media and antibody selection also did not provide any explanation. Instead, it might be suggested that the combination of technical and biological factors is irretrievably mixed such that any single metric is hardly predicative. Given the inherent variation in experimental procedures it would be quite difficult to design an experiment to measure the power of each quality metric in relation to the number of peaks called. Perhaps a bigger question would be whether or not the variability between replicate experiments reflects the genuine binding which occurred in the biological sample, in which case post-hoc filtering which is commonly performed on ChIP-seq peaks to retain only the reproducible sites may not be justified. This could be tested using a number of different experiments: Highly multiplexed single-cell ChIP-qPCR could be used to target highly variable peaks (HVP) to check whether Oct4 binding can be measured in any cells from a population (VanInsberghe et al., 2018). This approach could be extended using single-cell CUT&RUN and more recently single-cell calling cards (ssCC) to increase the chances of detecting HPVs by assaying a larger sample of the population (Hainer et al., 2018b; Moudgil et al., 2019). Such HPVs could also be distinguished by applying the CRISPR/Cas9 system to mutate or completely remove a subset of highly variable ChIP-seq peak regions in Oct4 knock-out cells (King and Klose, 2017). After performing a combination of ChIP-seq experiments in mutated versus wildtype and knock-out versus wildtype cells, certain

HVPs could first be identified as genuine Oct4 binding sites (i.e. peaks present only in the wildtype versus knock-out cells) and then this could be cross-referenced against HVPs which are still present even in the absence of the binding site (i.e. peaks originating from non-specific DNA sliding versus specific DNA binding).

The fact that most published studies analysed in this thesis did not perform replicate experiments and are therefore difficult to reproduce is not entirely surprising. Ioannidis and colleagues argue that the reward system in science, described as the purchasing of academic "goods" (e.g. promotion and other powers) with "currency" (e.g. publications and grants) can unintentionally select for "prolifically mediocre and/or irreproducible research" (Ioannidis, 2014). They exemplify the problem by measuring the repeatability of 18 published microarray gene expression analyses, concluding that only two in principle and six partially were reproducible compared to ten that could not be reproduced. The reasons for this were mainly unavailability of data, and discrepancies in data annotation and analysis protocols (Ioannidis et al., 2009). When considering why few experiments are replicated, a number of reasons have been suggested: Large-scale studies are particularly expensive and performing a minimum of three replicates is prohibitively expensive. Some biological samples are difficult to obtain, such as in clinical research, and there is simply not enough material to spread across multiple experiments. Resources can be better spent on assaying samples once via different technologies (e.g. ChIP-seq, RNA-seq, and ATAC-seq) than multiple times with only one technology. Replication is not incorporated up front in designing the research agenda in a given field, this is particularly prominent in ChIP-seq experiments where the most popular peak calling algorithms (e.g. GEM, MACS, and MUSIC) do not model variability between replicates and simply average or sum the biological signal across datasets (Thomas et al., 2017). This is in sharp contrast to RNA-seq experiments where the most popular tools model the data in such a way

that a minimum of two replicates is required to even run the analysis software (e.g. DESeq2, edgeR, and limma) (Conesa et al., 2016). Overall, Ioannidis and colleagues suggest a number of ways to mediate the problem including "the adoption of large-scale collaborative research; replication culture; registration; sharing; reproducibility practices; better statistical methods; standardization of definitions and analyses; more appropriate (usually more stringent) statistical thresholds; and improvement in study design standards, peer review, reporting and dissemination of research, and training of the scientific workforce." (Ioannidis, 2014). Although not all of these interventions are practical, it is clear that the current system does not reward replication and that the adoption of some of these practices is necessary.

6.4 Scientific and engineering implications

The development of the Daim software package for the analysis of DamID-seq data will help facilitate research in conditions where the number of cells are limited or an effective antibody for the protein of interest has not been produced. Using the peak calling strategy described, transcription factor binding and chromatin accessibility sites were detected using a minimum of 10^3 cells. This represents a significant decrease from the 10^7 cells required for conventional ChIP-seq experiments, which should allow unprecedented interrogation of DNA-binding and chromatin accessibility within rare cell types and *in vivo* organisms. In fact, Daim has already been used successfully within our research group to measure for the first time Oct4 binding in the gastrulating mouse embryo (Tosti et al., 2018). The implementation of a single method to detect binding and accessibility from the same sequencing data also allows for a deeper understanding of the epigenetic environment in which the transcription factor is binding. In iPSC cells and reprogramming a number of so-called pioneer factors bind specifically to closed chromatin to recruit nucleosome remodelling proteins – this

characteristic could readily be captured in a single DamID-seq experiment. Whilst the development of new functional genomic technologies is nearly always welcome, it is important to be conscious of the fact that DamID-seq fills a particular niche and that it is not competing in same space as already established assays. For example, chromatin accessibility can already be measured from 50 to 50,000 cells using ATAC-seq and has recently been upgraded for use in single cell experiments. This technology has largely usurped all previous chromatin accessibility assays including DNase-seq and FAIRE-seq because of its simple and quick experimental protocol. However, DamID-seq is the only one of these assays which can be combined with tissue-specific promoter expression to measure *in vivo* chromatin accessibility across different cell types (Southall et al., 2013). For detection of DNA binding sites, the advantage of DamID-seq over conventional ChIP-seq is the ability to use far fewer cells and no reliance on having an effective antibody for the protein of interest. However, recently a technique called CUT&RUN which is based on antibody-targeted cleavage of the protein bound DNA has been shown to map transcription factor binding sites with high resolution from 10^3 cells (Skene and Henikoff, 2017). Whilst this is a significant improvement, it still relies on having an efficient antibody and cannot be used to profile *in vivo* binding across different cell types.

6.5 Limitations of the research

Despite showing that DamID-seq can be used to identify transcription factor binding and chromatin accessibility sites from a minimum of 10^3 cells, there were a number of factors that limited the scope of this work:

First, all of the DamID-seq data analysed in this work was generated within our research group so the level of variability between independent experiments could not

be measured. As seen from the comparative ChIP-seq analyses, a multitude of unknown technical and biological factors contribute to the accuracy and sensitivity of a sequencing experiment. This could therefore change the interpretation of the results described in this work depending on how drastically the number of peaks changed. The lack of comparable sequencing data is mainly due to the fact that only in the last year did our research group describe an optimised DamID-seq protocol which would allow the identification of transcription factor binding sites from minimal numbers of cells (Tosti et al., 2018). It is hoped that in the coming years the adoption of DamID-seq by different research groups will generate a large supply of sequencing data so that technical artefacts may be identified and methods to accommodate or remove their influence on the results can be developed. Two notable examples include the identification of so-called blacklist regions and high-occupancy target (HOT) regions in a number of functional genomic assays (Landt et al., 2012; Li et al., 2016). The blacklist regions were discovered by ENCODE through comparative analysis of hundreds of sequencing experiments, which allowed them to identify what they describe as regions of the genome that show artificially high read mapping independent of the experimental condition. Sequencing data is now routinely filtered to remove these anomalously mapped reads in order not to impact the calculation of quality control metrics and call peaks within these regions (Carroll et al., 2014). The HOT regions were originally defined based on an unusually high number of transcription factor binding sites located at specific genes, and at first appeared to be relevant to the developmental processes of the cell. However, through comparative analyses of knockout transcription factor ChIP-seq data these regions were later found to be artefactual and instead were caused by specific sequence characteristics and enrichment of DNA tertiary structures (Wreczycka et al., 2017). Given the infancy of mammalian DamID-seq it is not inconceivable that technical artefacts in the data are waiting to be discovered and may currently be misinterpreted as DamID-seq specific peaks. In the

absence of a large amount of published sequencing data, it may be prudent to design functional experiments which target DamID-seq specific peaks to check their binding. One possible approach would be to design a qPCR experiment to measure the ratio of Dam and Dam-fusion methylation which would independently confirm that binding had occurred. Another would be to compare peaks called using Dam-fusion proteins containing either a functioning or non-functioning transcription factor. Those peaks which are present in the non-functioning assay would allow one to identify characteristics of DamID-seq artefactual regions which could then be used in future experiments to remove peaks within these regions.

Second, the peak calling strategy implemented in the Daim software package relies upon replicated DamID-seq data in order to identify transcription factor binding and chromatin accessibility sites. This is a departure from many conventional peak-calling algorithms which do not model the variability between replicate experiments (Pepke, Wold, and Mortazavi, 2009). Instead, peaks are called using merged data from multiple experiments or separately which then requires post-hoc analyses to identify reproducible peaks across experiments. The sophistication of these post-hoc analyses varies from simply looking at overlapping peaks to modelling the irreproducible discovery rate (IDR) to identify at what significance threshold peaks from two experiments stop being reproduced (Landt et al., 2012). The former method has clear limitations because a user-defined definition for what amount of overlap and in how many replicates constitutes a reproducible peak is required. The latter method is much more independent in that only a pre-defined value for the IDR significance threshold - usually 0.05 in the biological sciences - is required. The method implemented in the Daim software package however is to directly model the significance of each peak by looking at the variation in read counts between experiments. If a read count is highly variable it is unlikely to be assigned a P value small enough to cross the pre-defined value for the

false discovery rate (FDR) significance threshold. This disparity in measuring reproducible peaks therefore creates an artificial difference between assays which may be misinterpreted or attributed to factors unrelated. For example, a peak which exhibits a high average read count but also high variance between experiments is less likely to be called by the Daim method because the level of variability reduces the likelihood of achieving a P value smaller than the pre-defined FDR significance threshold. This variability however may be an intractable feature of the sequencing data and enrichment should instead be modelled using simpler non-parametric assumptions. This is perhaps the main advantage of conventional peak-calling algorithms because any difference in the signal-to-noise ratio between replicate experiments is not explicitly measured. With particularly noisy sequencing data, this allows the researcher to be more flexible in their definition of a genuine peak based on prior biological understanding of the system, at the inevitable cost of increasing the false-positive rate. In the specific context of DamID-seq, it may be advantageous to investigate potential non-parametric peak-calling alternatives especially at lower cell numbers where the variance between experiments is increased. There is currently only one published non-parametric algorithm, however it has not been made readily available and has only been demonstrated using sequencing data from the much smaller *Drosophila* genome (Li, Hempel, and Jiang, 2015). Peaks are defined based upon re-sampling of the data to generate a distribution of average fold changes which is then used to identify restriction fragments which achieve a fold change greater than the 95th percentile. Whilst this approach reportedly generated a comparable number of peaks to ChIP-seq data, without access to the original algorithm or reimplementation from details in the publication it is difficult to evaluate this particular method.

Third, the peak calling strategy implemented in the Daim software package was only tested on mammalian transcription factor binding. These proteins often have a small

footprint, hence there is no evidence that Daim can be used to detect large interactions such as nuclear lamin binding. In theory, this should not be problematic because peak calls are made by combining neighbouring restriction fragments – increasing this distance should allow chaining of restriction fragments into much larger regions representative of the actual binding. To promote the use of Daim it is important to show that DamID-seq data from other organisms can also be analysed. The technology itself was originally developed for *Drosophila* and is where the application of DamID-seq is currently most popular. However, no problems are currently anticipated because the peak calling strategy is agnostic and only requires the reference genome sequence to generate the restriction fragment annotation used for read counting. In fact, the DamID-seq signal-to-noise ratio in *Drosophila* is reported to be much higher than in mammalian cells due to technical differences related to the optimum expression of Dam, which suggests that peak calling should be more effective.

6.6 Future work

Although the results presented in this thesis demonstrate that Daim can be used to analyse DamID-seq data from minimal numbers of cells, there are several follow-up experiments which are required:

First, a comparative analysis of other DamID-seq peak callers would help establish which algorithmic features are best suited to this type of sequencing data. There are currently four different published analysis pipelines for DamID-seq data, all of which implement novel approaches but none of which have been benchmarked (see Table 6.1 and Subsection 2.2.7 for details about each pipeline). The current advantage of Daim is that it incorporates methods for identifying both transcription factor binding and chromatin accessibility sites, whereas the other four only detect binding. However, most analyses of DamID-seq data in the literature use ad-hoc analyses catered to the

Author	Pubmed ID	Availability
Li et al	25785608	Available upon request from author
Marshall et al	26112292	https://owenjm.github.io/damidseq_pipeline
Gutierrez-Triana et al	27707796	https://bitbucket.org/juanlmateo/idear
Maksimov et al	27766446	https://github.com/Vift/DamID-Seq

TABLE 6.1. Publicly available DamID-seq analysis pipelines.

This table contains a list of all publicly available DamID-seq analysis pipelines at the time this work was published (see Subsection X for an extensive description of each pipeline).

experiment and are rarely released as a fully functioning software package. It would take a great deal of development time to reimplement these described methods so a comparison of officially published analysis pipelines would be better suited. The main difference between these four peak callers and Daim is that they do not account for replicate variability in the estimation of peak significance. This approach is much more similar to conventional ChIP-seq peak callers and it would not be surprising if a larger number of peaks overlapped. Whether this represented more accurate peak calling or simply increased false positive rates on par with current ChIP-seq algorithms would be an interesting avenue of investigation.

Second, a number of computational methods have been developed for Hi-C data which theoretically could be adapted to DamID-seq data. In a typical Hi-C experiment, chromatin is crosslinked with formaldehyde, digested using restriction enzymes, and interacting restriction fragments are ligated together and sequenced. The aligned reads are then counted into restriction fragments and this measurement is used to quantify interacting regions of the genome. The quantification of restriction fragments is what makes DamID-seq and Hi-C analytically similar, and a number of approaches to removing sequence composition biases and detecting enriched restriction fragments have already been developed. Most notably, the use of hidden Markov models (HMM) to detect the edges of topologically-associated domains (TAD) could

directly be adapted to identify differentially methylated regions. A measurement called the directionality index (DI) is used to quantify the extent of upstream and downstream bias in the sequencing data, this produces a fold change value for each restriction fragment along the genome. The fold change values are then used with the HMM to assign upstream and downstream bias calls to each restriction fragment. If the DI values were replaced with the fold change values from the Dam and Dam-fusion binding this approach would theoretically be usable, however its accuracy and sensitivity would have to be determined.

Third, an implementation of a non-parametric or single replicate peak-calling algorithm for DamID-seq data would be advantageous at lower cell numbers. Observations of the sequencing data indicated that library complexity decreased with cell number and this produced stark differences between the replicates which was difficult to overcome with normalisation. A restriction fragment in one replicate would have a read count of zero compared to a second replicate with a high read count – the difference between these two replicates was presumably the chance amplification of the restriction fragment in one replicate from the minimal number of cells. Genuine binding sites detected using ChIP-seq and larger cell number DamID-seq experiments were missed at lower cell numbers due precisely to this read count variation. Imputation methods for single cell RNA-seq which try to replace the missing value would not be helpful in this scenario because they require a large number of independent replicates to accurately model the drop-out rate. By calling peaks from multiple replicates independently, a flexible rule may instead be applied to retain peaks which are in a certain number of experiments. The likely increase in false positive and true positive rates would however have to be examined to determine whether this approach is beneficial.

Appendix A

Primers for qDamID experiments

Target fragment	Forward primer	Reverse primer
chr1:3253452-3254235	TCTGATTGTTGAATGGGAAGCC	GGAAATGAGTTCTTCCAGGAGG
chr1:4493131-4494215	CAAGTAAACAGAGCTGTGTCCC	TTGCACAGAGATTGTCTTAGCC
chr1:4531728-4532469	CATCCATCTTAAGGGAACTGGC	GAGATACACAGTGACCCAGAGG
chr1:4756249-4756898	GGTTCCATATTAAACCTGGAGGC	TTCTCAGTGTGAAGCAATCAGC
chr1:5103390-5105015	AGGGAGTACTTTAGCCTGTACC	AGGCAAGAGATTCCTGAGTTGG
chr1:5313921-5314691	GTACAAGCAGCTGTGACTAGC	CAGTCCTGATGTTCATTAGGCC
chr1:5973619-5974882	TTGTTATCTGACTACGTGCTGC	GTCATCTTTCTGAAACCCTCGG
chr1:7430889-7433398	TTATCGCAATCGTGACTTGAGG	CTAGCACAAAGTAACCGGATGG
chr1:13161338-13163819	GGTACTTAATCTGCTTCCTGCC	CGCAAGTTACAGTGTGTCACC
chr1:14213373-14215170	GTGGGACAGACAGTCTAGACG	TTTGCTGGGTATGCTGGC
chr1:20112703-20113882	TACCTAGATTGCGTGTGTTTCG	TGCAACAGACACAGAATCATCC
chr1:22428392-22430482	AGTCGATTCCAAGTCCTCTAGG	AAAGGCTGTTCTAAGCTTAGGG
chr1:24390214-24392382	CTTTGCTGAAGGCTACGTCC	GAGACCTCTGACTGTGAAGATGG
chr1:31709927-31710857	ACCTGGTGAACTATGTGGAGG	CCGTACCTTGGTATTTGGCG
chr1:38815572-38817400	CCTTTCAGAACGATATGGCAGG	GACCTAAGACTTTGACACAGCC

Target fragment	**Forward primer**	**Reverse primer**
chr1:40486125-40489188	CTAAGTGTTGGGAAGATAAGGGC	CTCCTGCAGCATTTACAGTACC
chr1:42904575-42906636	CTGTGTGCCTGTCTTGTAAGG	GGTAGTCCTTCTCACAGTAGGG
chr1:54923879-54925278	GTTTATCCTGCTGTGACTTCCC	CTTAGCTTGACAGTACAGCGG
chr1:55940730-55943982	GAGGAAATCACAAGCAGTCTGC	AGAGATTCAGGGAATCCACTCC
chr1:62871403-62874319	TCTGAGCTCACTTAGAGTCTGG	AAGACCACAAGATTCCAAAGCC
chr1:63963195-63965397	TTAAAGCGAAAGTGAGTCTGGG	CAGAAATAAACACCCGTAGCCC
chr1:64936451-64938194	GGTCTGCTGAGTTAAAGGAACG	CAAGCTGGTTTCTCTTAGGTCC
chr1:68040882-68042402	TTCAGCACTCTCAGAGATAGGG	CTGAAGGGTTTGACTAATGGCC
chr1:72554310-72555554	AAACTCCTCAAACCCTTTCACG	TTACCTCATAGAAGGGACAGCC
chr1:75315625-75316487	TTACTGCATTTCAAGGGAGTCG	CTGCATCTCTAAAGCAACTCCG
chr1:76035979-76036949	GCTTGAACTAGAGTGAACAGCC	GGGTTCTACAGTTTAAGCACGG
chr1:82386392-82388475	AGTGGGTATTGTTAGACCAGCC	TTTGTAGGAACTCTCTGCAACC
chr1:84116661-84118119	ACCAGTTCTCTTGTCAACTTGC	TGTGGCTGTTGTAATAAAGGCC
chr1:119082974-119085311	CTACTTGTGGACAATGCCTGC	TTGGCTGACTTACTGAACTTGC
chr1:119441511-119443155	GAAATGAGGCTAGCAAGGTTGG	GCATTCAAGGAAATAGCGTTGG
chr1:125677451-125678688	TGAAGGAGAACACAATGAGAGG	ATAGGAAAGAATGACAGTGGCC
chr1:133692940-133695586	CACTGAGCATGTAGTTGTCAGG	TCATGTGTTAAGTCAGCTTCCC
chr1:140434347-140435631	GGTGGTTTCCTGTGATTTGACC	AGGCTCTTGTGTACTCTGAGC
chr1:146425263-146426280	CTCACCTGCCTAATGAACTTCG	TGAGGGTTCTAGAATACCTCGG
chr1:153091676-153093993	TTGTTAGCCTCAGTACTAGCCC	CTATGTGTACCCAGCACTGTGG
chr1:158722229-158724089	GAAGAGAAGGTTTGGTGTGTCG	TTGTTGTTTCTTGCACAAACCG
chr1:163683033-163684928	TGACCAAATTCTTGTGAGGCG	TGACTCAATAAATGCCCTTTGC
chr1:165174347-165175474	TGGTCGTCACTGGCGTATCC	GCATCTCCAGGCTAGCTTCC
chr1:165634885-165636170	CTTGCTCTTTCCCTACTTAGGC	AGTTGTTACCTAGCTGCTATGC
chr1:172127217-172129869	TACCCTGCCCTAACTAAAGACC	CCTACGAGGAGAAGTAAGGACC

Target fragment	**Forward primer**	**Reverse primer**
chr1:177771938-177772947	ATTCTCCTCTGAGTCTTCTCGG	CTCTGATTAACACCACAGCAGG
chr1:194332168-194333861	AGAACACCTTTCCTCAAGTCCC	ACTCTGTTCCCTAGACAACTCC
chr2:4821035-4824136	GCGTTTGATAAGCTTGAACTGC	TACAATTAGTGCTTCTGTGCCG
chr2:10456022-10457492	TGTACTCGTAGCTTAGGTCACC	AGTCTAGAGGCTTCTTTCTGGG
chr2:25687921-25688545	GAGATGGTAACTACCTGGTTGC	TCACCCTTAATCATGGAGTCCC
chr2:37995275-37998526	AATCTCTACCCTGGTCAGTTCG	ATGTCACACCTTGAAGCATAGC
chr2:48609145-48610627	ACATAGATGCCGTTGATTCAGG	AAGCAGGATGTAGAAACCAAGC
chr2:51465368-51467137	TGAAGCTATTGTGCTCCTTTGG	CTTTCTGGGTTACAAGGCTGG
chr2:52903293-52904977	TGAGACTACTCATTGAGCTGCC	TATTCACAATTGCAGCCTACGG
chr2:55211351-55212752	CAGGAGGAAATGACAGCTTACG	CAAGTCAGAACAAGCAGGAAGG
chr2:67956580-67958146	ACCATAAAGCAATGTGGATGGG	CTTAAACTCCTGCTAAGTCCGC
chr2:69000950-69002670	GGCTACTGTACTGATGTTGACG	CACATCTATCCTTTCCACTGCC
chr2:112039529-112040787	CCAAGGAAGTTGAAGTAGTGCC	AATAGTTGACCTACCCACCTGC
chr2:125210888-125212989	TACCTCCATTTGGATTTCAGCC	TGCACAAGAGAGAATTAGAGCC
chr2:134289502-134290098	ACTTCATTGTTCCAGGTTTCCC	ATACAACACAATCCAGCTGACG
chr2:156727216-156728512	CGAGCAACAAAGCTAAGGAGC	CTATAGCTCTGTCCCTCAGACC
chr2:163052023-163053396	ACCAGATTCCTTGTGACACTGG	CAGTTAAACAGGACACCAGAGC
chr3:5300379-5302422	AAGCAGCTCTGACTAACAAAGC	TGGCTTACTACAGTTCTCCACC
chr3:51233022-51234887	TAACATGAGAGCCAACTTTCGC	TGTTAGCTGGCTTCAGAATTCC
chr3:69315549-69316381	AAAGGCTGCTGATAACCAATCC	GGGAGGGCTCTTATAATCCTGC
chr3:82438239-82439288	TTCCTTTAGATGTGACACAGCC	GAAAGAAGGGAAATGTGGACGG
chr3:83211319-83213668	CCACCAAGGACTCAACTTAACC	AGTCTGATGTAACTGCTCTGGG
chr3:118434052-118435589	TGAAACGGTTTCACTTCTGTGC	ATGTGAACATGGCTGCTTATCC
chr3:137691430-137695285	TTTGTGTTCAAACCCAAATGCC	TCATCTAAGGACACAAGATGGC
chr4:53558459-53560999	ACATTTAACCAACTCCTATGGCC	ACCTGTTTGAACTAGTCTTGTGC

Target fragment	**Forward primer**	**Reverse primer**
chr4:57779931-57782838	TCCAGAGAATGACTACAGACCC	TCTTCCGTATTGAGAGGAACCC
chr4:72167497-72168473	CACCCTCCCAGTTTATTTCACC	CCAAGGAGGATTAAAGAATGAGCC
chr4:82534483-82536959	AGGCTCTGTCAACATTTCAAGG	GAGGATGACAAGATGTTGGAGC
chr4:84222878-84225102	ACAGCTAGTGTGTGCTTTATGC	GCAACATAAACCTGTCTTTCGC
chr4:91368985-91372725	ATGCGGTCCTAATTAACCTTGC	GTAGCTAGTGCACTAACATGCC
chr4:98815280-98817538	AGTTAACCTCCCTAAGCAACCC	CCTCTGACAGCTCTGACTAGG
chr4:118428385-118430610	AACCTGGAACCATAACTTGTCG	ACTCTGCACATAATCCCATTCG
chr4:130306738-130308873	GAGGTCGGGTTTAGTTTCATGG	ATTCTAAGTCAGACCACAGCCC
chr4:133531465-133533488	AACTCAACAGTATCAGTGCAGC	ATGTCCCAGAACCTAGTAGTGC
chr4:134107522-134109295	TCCATGAATTTCTCTCTCGTGC	CTTACCTCTTAAGCTGTGCTGG
chr4:134752002-134754986	TCTGGCTTTCAAAGTCACATGG	GCAAAGCCCATTAGAATCTCCC
chr4:136175513-136177367	TGAAGCAGACATTATCAACCGG	TCTAACGGAGACCTCATTCTGG
chr4:141313622-141316758	AGCAAAGAGAAATAGGCCTAGC	CCCTTACGTCGAACTATTGAGC
chr4:149625131-149629555	CCGTTAGTAGGTCGGCTTCC	CAGATAGCCTGAGTCCTCTCG
chr5:14447995-14448945	AAACTGGAGAACAAGTCCATGC	GGGAGAAATGAAGCTGCTACC
chr5:18560576-18561466	GGTCCCAAGATTTATCTGCACC	CAGTCTGGAGAAACACTTCTGG
chr5:24401315-24405702	ACTAGCTGCTTCCAAATGAACC	CAAACTCTAGGATTCTCACCAGG
chr5:35883971-35885891	GATGTGTCTGTCCTTTGATGGG	CTCAGCATCTGTGTCTTGTTCC
chr5:92952892-92954791	AACTATAGTCACCACTCCTCCC	AATTCTAGCTAGGTCTGGAGGG
chr5:115247013-115248334	TCATTATTGCCACTCAACTGCC	GGTGTCTCTGGTACTAAGTCCC
chr5:115550133-115552970	GTACACAGGCTTCAAGTACACC	AATAAAGCGAAAGAAGGTTGCC
chr5:122451992-122455731	GGGTGTTTCACTATTGTAGGCG	TTAGGAACAGTGAACGGAAACC
chr5:131447548-131450388	CTAGCTGCCTTAAACTATGCCC	AAGAGATTCTCTAGGTGTGCGG
chr5:135404400-135405646	TTTGCCGAGTATCTGTCTTTGC	CTCAGAGGTTAGGGCAAGATGG
chr6:5565922-5569601	TAGCATTCCATTCGTGTAAGGC	GCAAGTACTGTGAATCTGCTGG

Target fragment	**Forward primer**	**Reverse primer**
chr6:54847808-54848857	AAAGGATGTGGAGTTGAACTGC	ACGTGTACATACCATGTTGTCG
chr6:77208387-77209214	GATGAAGCTCTCATGTAGTGGC	ACTGGCATACGATGTTTAGTGC
chr6:96284472-96285955	TATTCAGTGAGCTTTGGAACCG	AAGCTAAATGTCAGAACCTCGG
chr6:135680948-135684382	AGAGAGATACCTATGACGCTGC	AGGGTCTGGCTTTCTATAGAGC
chr7:34879169-34882516	AGACTTTCCCGATAGTCGTTCC	CTCTGCATTAAAGAGCCAGCC
chr7:36897457-36898498	ACAAATGAAATCACCCAGGTCG	AAATTTCCCTGCCGAGTTAAGG
chr7:54400976-54403661	GATGTGTATGACGAGCAGATGG	ATAAGTCCCGTTGTCAAACACG
chr7:69322257-69323206	CTTCCAGAGAAATGTTGGAGCC	GCCTCTGTTTATGGATGTCACC
chr7:78895867-78897740	GATGATAGCATTGGCATCTCCC	ATTGACTCTGAAGCCGATTTCG
chr7:130253489-130254433	CTGGGTGCCTCTATTTACTTGG	CAAGGGTCAATCTTCAAGGTGC
chr7:133041820-133044547	CATTTCTCAGCTGATTCCCAGG	CATATCAGGTGCAAGGGTAAGG
chr7:138065291-138065643	ATTCCCTTGTCATAGTCCTCCC	TTCTACCACATACTAGGGCAGG
chr8:10664942-10666821	ACACCTGTCATACCTCATGTCC	GTGAATTCAGTGGCCTGATTCC
chr8:11556937-11557652	CGATGTTCTTAAGTCTCTGGCC	TTTAGAAAGAGTTGCTTCCGGC
chr8:12499834-12502301	GCTCTGGAAGATTTGCTGAGG	CCCTAGCAACTTTCTAACTGGC
chr8:18688902-18692067	ATCTCATTTCACACAGCATCCG	AGGTTGCCATGGTTACAAATGG
chr8:33914668-33916112	TAGGAAGCCATGCAGACTTGG	TAAACTTCCAAGCCAAGAAGCC
chr8:47632447-47633584	GACACTTTCAGCCATGACTAGG	GCAGATGGAAAGAAGGGTAAGG
chr8:48579640-48580447	GCATAAGTGCGGATAGGAATGG	AGGCTCTTTAAGGAAGCATTGG
chr8:69885400-69886403	CCAATCCTGAGTATGACAAGCC	AGATGGGAAACAAGACACAAGC
chr8:69995116-69995965	CATCCAGTGGACTTTGACTTGG	AAGCCTCCACAGTTAAGAATGC
chr8:78663543-78665625	AAGCTTGGCAACTTTAGTCTCC	GAAACTACGTATGACACCCACG
chr8:83714525-83715988	TGAACTTGACACACTGAGAAGC	TGGACCACTATCCTTGAAGACC
chr8:87434061-87435122	CAGGCTGCATCCATACTATTCC	CAATGTCCAAACCCATCTGAGG
chr8:92046873-92047317	ACCTTCGTCTATTGCTCCTACG	CTAGGTTCCTTCATTCTGCTCG

Target fragment	**Forward primer**	**Reverse primer**
chr8:92071699-92072933	TCCTTCCAATCTCAAACTGTGG	TGGAAGAAGGTAGGGAATACGG
chr8:111787778-111789383	TAGGTCAAGTTTGAAGTGCTGC	AACAGCACCATCTAGAGGTAGC
chr8:112427972-112429683	TAAGCTGCCATCAGTAACTTGC	TAGTTGTGACATGATTCTGCCG
chr8:122408363-122411376	CTGGGTTTCAGATGACTCATGG	TGGCCATGCTTAAGTTTACAGG
chr9:8000380-8003661	TAAAGGAGTCAGACTTGGGAGG	GTGCGTCTTACTCAGCTTTAGG
chr9:43269694-43272175	GCACATTTGGTAAACGTGAACG	TGCTAGAGAGAGTAAAGCCACC
chr9:45381414-45383039	AAAGATGCTGCATTAAGGTGCC	GGCCAGATTAACCAGATTGTCC
chr9:58456834-58458521	ACTAAATCATGCAGCCCATACC	TGGATTGATTACTTCGTTGGCC
chr9:74860196-74861312	CATGTGCAGTTGTTATCTTGGC	TTGCTTGGAATTAACAGGACCC
chr9:85323857-85324854	TCATTTGTTCTGATGCCATGGG	TAACTGACAAGGTAGACTCGGG
chr10:8881728-8883835	GTTGGCTAGATTCAAACCTCCC	GAATTCCAGGGAAGTTCAGTGG
chr10:39131482-39134226	GGGTCCACAGGAATTATTTGGG	CTCATCTCTTAACCCAAGTGCC
chr10:59558145-59558845	CTGCAGCAATGAAGAATTGTGG	GGTAGTTGAGCAGCATAATCCC
chr10:62290564-62292758	TTTCTTGAGGTGTGGCTTATGC	TAGTGTCAAAGGAAGGGTCAGG
chr10:81177681-81179592	TGACAGCACAGTTACTAACAACC	CTAGTGAGATGCAGCTTTCAGG
chr10:82742922-82744187	AACAATCCCTTCTCAGAGACCC	GCATGTCTAGTGCTCTATTCGC
chr10:121854223-121854766	AGTGGGCTGTTGTAGAGATAGG	GCTACATCTCCAATCTTCACGG
chr11:6416828-6419305	CACTAAGTGGCCATTCATCTAGG	GCAAAGAAAGCAATTCGTGAGG
chr11:11810527-11812335	CTTTGATTTCCCACCTTGTTGC	TCGTTAGGTCTTCCTTTCCTCG
chr11:11972726-11973919	GGAATGAATGCTCATGAGGAGG	AACAGCAGCTAAAGTAAGTGCC
chr11:40688698-40690983	GCTTTGTCCCAAGATAGACTGG	TAAACTTCACTTGCCATCAGGG
chr11:54522513-54525383	TCTCACATTGGTTCCACTCTGG	GAGGTGATGAGGATTTCAAGGC
chr11:58918809-58920660	TCTTTCTTGTAGCTCTCCACGG	AATCCAGTGACATCATTCTGGG
chr11:76803502-76804176	TTTCTGCTCCACATCAAAGACC	GCAATCAGAGCACCTAGATAGC
chr11:81515233-81516449	CTTAACTCCAAAGCTGGGATGG	GTGAGACCATGCTCTAAGATGC

Target fragment	**Forward primer**	**Reverse primer**
chr11:81671790-81673234	GTGAACGGCGCTAAGTAAAGG	TTTCAACGTAGGAATGTGACGG
chr11:89985661-89986826	ATACCACCTTGATAATGACGCG	ATTCTTACGCTCAGGTGAAACG
chr11:97686494-97688112	AAATAAGAGGCTTTCCTTGGCC	CTTGGATGCTGACACTGATTCC
chr12:21440697-21442974	AGGACATCCTTTAACTGCATGG	CTTTCATCACCTGCTAGATGGC
chr12:73307126-73309821	ACTCCAAGCATGTTCCTTATGG	AGAGCTATGTCTCTCTGAAGGC
chr12:88459110-88459314	GGCCAGAATAGCAGTTTAAGGG	AGTTGTGTGACTGTTCAGATGC
chr12:95691212-95693125	AGATTTCCAGAACGTAAACCGC	TAATGAAGGGCAGTGAGACAGG
chr12:99303370-99305289	GCTTGGCTAAATGAGGACTAGG	GAGTCCCTTTGCAAGTAAGTCC
chr12:99777990-99778452	CTGCCATTCCTTAATGTCAGGG	GACTCCTCTGGCTAGGATTTCC
chr12:104315643-104316109	CACCCATAACCACTAAGACTGC	TCCTCCTATACATACTGCTGGC
chr13:15338501-15341053	CAGCTTTGTCCCATAGTACTGC	ATGTGCAGCAGCTAGTTAAAGG
chr13:17879943-17883018	GTAGTGTTGGTCCATTCCTGC	GTCCTAAAGTGCTTTCTGGACC
chr13:21528605-21529897	TGGTGTCCTTTATAACACTTGGC	GCACCTGGTTTAGAGTTAAGCC
chr13:24833780-24836006	CAAGCTGATTTCACACTTTGGC	TTTGCTGTCTTAGTTACTCCGG
chr13:63512984-63515457	TTCCACACCTACATGAATTGCC	TTCCCAGTCCTGAATAAGACGG
chr13:90010860-90011973	GAGTGTGTAGCTAGCAAGAACC	GTCTCAGTGGTGTTAGATGAGC
chr13:97358223-97359520	CATCATCCTGAGAAATCGGACC	TGGTAACTGACGGGTAACTAGG
chr13:97844864-97846023	GTTGGCAGACAGTCTAGAAACG	TAAATCCTGGGCAACATACAGC
chr14:9318954-9319856	AAAGTTAACAAGATGCCCGAGG	GTTGCTCCTTGTTTCACAATGC
chr14:25663235-25665616	CAATTAATTACGGGCATGCACC	CCTGACCAGATGTTCTTCAAGG
chr14:46787347-46789817	TGCTACCAGCTGATAGTCATCG	CTATGAAACTCACATCTGGGCC
chr14:70311209-70314396	TCAGGCAGACATTAACATCTGG	ACAAGTCTACCAGTTCCCTAGC
chr14:118137068-118138253	ATCACAAAGGGAATTGCTGAGG	CTTTAAGAGGCATGTCCCTTCG
chr14:118436903-118438597	AGCACTCATCCTTCCTTTAGGG	ATGTTGTGATGAACATGGACCC
chr15:7691981-7693463	AGCAACTGACTGAATTAGCAGC	TCAGCTCTCTGTCATTGTGACC

Target fragment	**Forward primer**	**Reverse primer**
chr15:10688904-10691229	AGCTTGGAGGATAGAGACTTGG	GGAGACTCAGTAATGCCAAGC
chr15:16639242-16640639	CTTACTGCATATGAGGCCAAGG	TTCCACCGTTCTGACTACTAGG
chr15:25761707-25762624	ATTCTGCTCTGAATTGGAAGGC	ACTTCACCATCATTACCCTCCC
chr15:39735015-39736414	ATCCCTTTCCATCTGTCCTACC	CCTGTGGGTAATGACATCTTGG
chr15:74008882-74009516	TCTAGTCAACAGAGTGGTCACC	AATGGACTTGAGCCTTTGTTCC
chr15:77335966-77338119	TTAGGACCACAGGAAGAACACC	AACACAAGGGCTCAAGACTACC
chr15:79923243-79924489	AGCAATTTAGAGCAAGCAAAGC	TCTTAATCTTTATGCAGGCGCC
chr15:80080853-80082600	TAACATGATGGCTCTCATTGGC	TTAAGCAAACCCAACCACTAGC
chr16:22425658-22427816	GTAGAGACCATAGGTGAGGTGG	TTAGGGATGTGTTTAGTGCACG
chr16:45836549-45839314	GATGAGGACAGCTTACTTTCCG	CTTACGCTTGTACAATCCACCC
chr16:46981937-46983049	TGGTAGTTGGAGATTCACCTGG	TGCTCCTGGTACATATCCTAGC
chr16:77399623-77400674	TGCCTTTAGATTCTGCCAATCC	ACACATAGGATAAAGTGAGCGC
chr16:84769061-84770355	CAAAGGCTTCTGTTTATGCTGC	TCTTCCAGGTGAACTTTCTTCG
chr16:97712182-97712804	ATGTCTACAAAGATGGCTTCCC	TCTGGTACAAATGTGAACAGCC
chr17:15680762-15681579	CTATGTAAATCCACAGCCTGGC	CTTCTGAGTGAACTACCCTTGG
chr17:25264844-25265992	ATCTAGGCTGTGCAATCTAAGC	AATGCATTGAACTTTGAGCTGC
chr17:85126974-85129955	GTAGATGGTATGAGAGCCAACC	GGCTATTGAGTGCATATCACCC
chr18:6636783-6639051	TCCCAGTAATGGTGATGTCTCC	ACAGAAGGGAAGGTGAGTTTCC
chr18:38583827-38585550	GTTCCTTCCTCAACATGTCTCG	TGAGAAGCCAAGATAGGGTCC
chr18:54718826-54719731	GCAAACAGATGATTCCCTGACC	ATAGCCTAAGCCATGACAGACC
chr18:80980782-80984780	ATGCTTGAAGGCACTGATTAGG	AGATGCAGTGATAGCTGAGAGC
chr18:82513958-82515741	TGAAGTCACAGCGAATGAATGC	TGGCATATCTAGGAGGACATGC
chr18:84966262-84969810	GCAGTGCTCTAGGATTAAAGGG	ATCCACATGAAGACCTCTGAGC
chr19:23496296-23499191	CTTCAGCTCTGAGAGACTCTCC	GTGAGGAGAGGATTCCAACAGG
chrX:20797333-20799055	TCTTACTACCTCTATGCCTGGG	AAGTGAAGCAGTTGGTATGAGG

Target fragment	**Forward primer**	**Reverse primer**
chrX:140693290-140694391	CAAAGGGTGAACTCCAAATCCC	TTTCTGTCTGCATTGCCTTCC
chrX:166640464-166642127	ACTGTGACACTTATAAAGCGCC	CTTGAGAATGTGATGGCAGGC

Appendix B

Parameters for Primer3

Parameter	**Value**
PRIMER_PICK_LEFT_PRIMER	1
PRIMER_PICK_INTERNAL_OLIGO	0
PRIMER_PICK_RIGHT_PRIMER	1
PRIMER_PRODUCT_SIZE_RANGE	80-150
PRIMER_NUM_RETURN	500
PRIMER_MAX_END_STABILITY	9
PRIMER_MIN_SIZE	18
PRIMER_OPT_SIZE	22
PRIMER_MAX_SIZE	26
PRIMER_MIN_TM	56
PRIMER_OPT_TM	59
PRIMER_MAX_TM	62
PRIMER_PAIR_MAX_DIFF_TM	2
PRIMER_MIN_GC	39
PRIMER_OPT_GC	50
PRIMER_MAX_GC	61
PRIMER_MAX_SELF_ANY	4
PRIMER_MAX_SELF_END	3
PRIMER_MAX_NS_ACCEPTED	0
PRIMER_MAX_POLY_X	3
PRIMER_PRODUCT_OPT_TM	50
PRIMER_GC_CLAMP	2
PRIMER_MAX_LIBRARY_MISPRIMING	12
PRIMER_PAIR_MAX_LIBRARY_MISPRIMING	24
PRIMER_MAX_TEMPLATE_MISPRIMING	12
PRIMER_PAIR_MAX_TEMPLATE_MISPRIMING	24
PRIMER_LOWERCASE_MASKING	1
PRIMER_MAX_NS_ACCEPTED	0
PRIMER_FIRST_BASE_INDEX	0

Bibliography

Adli, Mazhar, Jiang Zhu, and Bradley E Bernstein (July 2010). "Genome-wide chromatin maps derived from limited numbers of hematopoietic progenitors". In: *Nat. Methods* 7, p. 615.

Aird, Daniel et al. (Feb. 2011). "Analyzing and minimizing PCR amplification bias in Illumina sequencing libraries". en. In: *Genome Biol.* 12.2, R18.

Aksoy, Irene et al. (Apr. 2013). "Oct4 switches partnering from Sox2 to Sox17 to reinterpret the enhancer code and specify endoderm". en. In: *EMBO J.* 32.7, pp. 938–953.

Aleksic, Jelena, Sarah H Carl, and Michaela Frye (June 2014). "Beyond library size: a field guide to NGS normalization". en.

Anaconda (2018). *Conda*.

Anders, Simon and Wolfgang Huber (Oct. 2010). "Differential expression analysis for sequence count data". In: *Genome Biol.* 11.10, R106.

Andrews, S (2010). "FastQC: a quality control tool for high throughput sequence data". In:

Ang, Yen-Sin et al. (Apr. 2011). "Wdr5 mediates self-renewal and reprogramming via the embryonic stem cell core transcriptional network". en. In: *Cell* 145.2, pp. 183–197.

Ason, Brandon and William S Reznikoff (Jan. 2004). "DNA sequence bias during Tn5 transposition". en. In: *J. Mol. Biol.* 335.5, pp. 1213–1225.

Aughey, Gabriel N, Seth W Cheetham, and Tony D Southall (Mar. 2019). "DamID as a versatile tool for understanding gene regulation". en. In: *Development* 146.6.

Aughey, Gabriel N and Tony D Southall (Jan. 2016). "Dam it's good! DamID profiling of protein-DNA interactions". en. In: *Wiley Interdiscip. Rev. Dev. Biol.* 5.1, pp. 25–37.

Aughey, Gabriel N et al. (Feb. 2018). "CATaDa reveals global remodelling of chromatin accessibility during stem cell differentiation in vivo". en. In: *Elife* 7.

Bannister, Andrew J and Tony Kouzarides (Mar. 2011). "Regulation of chromatin by histone modifications". en. In: *Cell Res.* 21.3, pp. 381–395.

Bansal, Vikas (Mar. 2017). "A computational method for estimating the PCR duplication rate in DNA and RNA-seq experiments". en. In: *BMC Bioinformatics* 18.Suppl 3, p. 43.

Bar-Nur, Ori et al. (May 2018). "Direct Reprogramming of Mouse Fibroblasts into Functional Skeletal Muscle Progenitors". en. In: *Stem Cell Reports* 10.5, pp. 1505–1521.

Baranello, Laura et al. (May 2016). "ChIP bias as a function of cross-linking time". en. In: *Chromosome Res.* 24.2, pp. 175–181.

Barski, Artem et al. (May 2007). "High-resolution profiling of histone methylations in the human genome". en. In: *Cell* 129.4, pp. 823–837.

Bemmel, Joke G van et al. (Nov. 2010). "The insulator protein SU(HW) fine-tunes nuclear lamina interactions of the Drosophila genome". en. In: *PLoS One* 5.11, e15013.

Benson, G (Jan. 1999). "Tandem repeats finder: a program to analyze DNA sequences". en. In: *Nucleic Acids Res.* 27.2, pp. 573–580.

Bergerat, A and W Guschlbauer (Aug. 1990). "The double role of methyl donor and allosteric effector of S-adenosyl-methionine for Dam methylase of E. coli". en. In: *Nucleic Acids Res.* 18.15, pp. 4369–4375.

Brandariz-Fontes, Claudia et al. (Jan. 2015). "Effect of the enzyme and PCR conditions on the quality of high-throughput DNA sequencing results". en. In: *Sci. Rep.* 5, p. 8056.

Bray, Nicolas L et al. (May 2016). "Near-optimal probabilistic RNA-seq quantification". en. In: *Nat. Biotechnol.* 34.5, pp. 525–527.

Brind'Amour, Julie et al. (Jan. 2015). "An ultra-low-input native ChIP-seq protocol for genome-wide profiling of rare cell populations". en. In: *Nat. Commun.* 6, p. 6033.

Broad Institute (2018). *Picard Tools*.

Buecker, Christa et al. (June 2014). "Reorganization of enhancer patterns in transition from naive to primed pluripotency". en. In: *Cell Stem Cell* 14.6, pp. 838–853.

Buenrostro, Jason D et al. (Dec. 2013). "Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position". en. In: *Nat. Methods* 10.12, pp. 1213–1218.

Buenrostro, Jason D et al. (July 2015). "Single-cell chromatin accessibility reveals principles of regulatory variation". en. In: *Nature* 523.7561, pp. 486–490.

Carroll, Thomas S et al. (Apr. 2014). "Impact of artifact removal on ChIP quality metrics in ChIP-seq and ChIP-exo data". en. In: *Front. Genet.* 5, p. 75.

Cassandri, Matteo et al. (Nov. 2017). "Zinc-finger proteins in health and disease". en. In: *Cell Death Discov* 3, p. 17071.

Castelo, Robert (2018). *GenomicScores: Infrastructure to work with genomewide position-specific scores*.

Chadwick, B P and H F Willard (May 2001). "Histone H2A variants and the inactive X chromosome: identification of a second macroH2A variant". en. In: *Hum. Mol. Genet.* 10.10, pp. 1101–1113.

Chen, Jiji et al. (Mar. 2014). "Single-molecule dynamics of enhanceosome assembly in embryonic stem cells". en. In: *Cell* 156.6, pp. 1274–1285.

Chen, Xi et al. (June 2008). "Integration of external signaling pathways with the core transcriptional network in embryonic stem cells". en. In: *Cell* 133.6, pp. 1106–1117.

Chen, Yunshun et al. (Nov. 2017). "Differential methylation analysis of reduced representation bisulfite sequencing experiments using edgeR". en. In: *F1000Res.* 6, p. 2055.

Chronis, Constantinos et al. (Jan. 2017). "Cooperative Binding of Transcription Factors Orchestrates Reprogramming". en. In: *Cell* 168.3, 442–459.e20.

Conesa, Ana et al. (Jan. 2016). "A survey of best practices for RNA-seq data analysis". en. In: *Genome Biol.* 17, p. 13.

Cooper, Geoffrey M (2000). *Regulation of Transcription in Eukaryotes*. Sinauer Associates.

Crawford, Gregory E et al. (Jan. 2006). "Genome-wide mapping of DNase hypersensitive sites using massively parallel signature sequencing (MPSS)". en. In: *Genome Res.* 16.1, pp. 123–131.

Cremer, Thomas and Marion Cremer (Mar. 2010). "Chromosome territories". en. In: *Cold Spring Harb. Perspect. Biol.* 2.3, a003889.

Creyghton, Menno P et al. (Dec. 2010). "Histone H3K27ac separates active from poised enhancers and predicts developmental state". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 107.50, pp. 21931–21936.

Dang, Chi V (Mar. 2012). "MYC on the path to cancer". en. In: *Cell* 149.1, pp. 22–35.

Das, Partha Pratim et al. (Jan. 2014). "Distinct and combinatorial functions of Jmjd2b/Kdm4b and Jmjd2c/Kdm4c in mouse embryonic stem cell identity". en. In: *Mol. Cell* 53.1, pp. 32–48.

Deng, Tao et al. (Sept. 2015). "Functional compensation among HMGN variants modulates the DNase I hypersensitive sites at enhancers". en. In: *Genome Res.* 25.9, pp. 1295–1308.

Devailly, Guillaume et al. (Dec. 2015). "Variable reproducibility in genome-scale public data: A case study using ENCODE ChIP sequencing resource". en. In: *FEBS Lett.* 589.24 Pt B, pp. 3866–3870.

Diaz, Aaron, Abhinav Nellore, and Jun S Song (Oct. 2012). "CHANCE: comprehensive software for quality control and validation of ChIP-seq data". en. In: *Genome Biol.* 13.10, R98.

Diehl, Alexander (Dec. 2017). *Cell Ontology*. `https://www.ebi.ac.uk/ols/ontologies/cl`.

Dieuleveult, Maud de et al. (Feb. 2016). "Genome-wide nucleosome specificity and function of chromatin remodellers in ES cells". en. In: *Nature* 530.7588, pp. 113–116.

Domcke, Silvia et al. (Dec. 2015). "Competition between DNA methylation and transcription factors determines binding of NRF1". en. In: *Nature* 528.7583, pp. 575–579.

Eckert, Alexander (2017). *parallelDist: Parallel Distance Matrix Computation using Multiple Threads*.

Edgar, Ron, Michael Domrachev, and Alex E Lash (Jan. 2002). "Gene Expression Omnibus: NCBI gene expression and hybridization array data repository". en. In: *Nucleic Acids Res.* 30.1, pp. 207–210.

ENCODE Project Consortium (Sept. 2012). "An integrated encyclopedia of DNA elements in the human genome". en. In: *Nature* 489.7414, pp. 57–74.

Ernst, Jason and Manolis Kellis (Dec. 2017). "Chromatin-state discovery and genome annotation with ChromHMM". en. In: *Nat. Protoc.* 12.12, pp. 2478–2492.

Ewels, Philip et al. (Oct. 2016). "MultiQC: summarize analysis results for multiple tools and samples in a single report". en. In: *Bioinformatics* 32.19, pp. 3047–3048.

Flaus, Andrew and Tom Owen-Hughes (Oct. 2011). "Mechanisms for ATP-dependent chromatin remodelling: the means to the end". en. In: *FEBS J.* 278.19, pp. 3579–3595.

Flynn, Ryan A et al. (Mar. 2016). "7SK-BAF axis controls pervasive transcription at enhancers". en. In: *Nat. Struct. Mol. Biol.* 23.3, pp. 231–238.

Furey, Terrence S (Oct. 2012). "ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions". In: *Nat. Rev. Genet.* 13, p. 840.

Galonska, Christina et al. (Oct. 2015). "Ground State Conditions Induce Rapid Reorganization of Core Pluripotency Factor Binding before Global Epigenetic Reprogramming". en. In: *Cell Stem Cell* 17.4, pp. 462–470.

Gao, Lei et al. (Mar. 2018). "Chromatin Accessibility Landscape in Human Early Embryos and Its Association with Evolution". en. In: *Cell* 173.1, 248–259.e15.

Garcia-Fernàndez, Jordi (Dec. 2005). "The genesis and evolution of homeobox gene clusters". en. In: *Nat. Rev. Genet.* 6.12, pp. 881–892.

Germann, Sophie et al. (Oct. 2006). "DamID, a new tool for studying plant chromatin profiling in vivo, and its use to identify putative LHP1 target loci". en. In: *Plant J.* 48.1, pp. 153–163.

Gilfillan, Gregor D et al. (Nov. 2012). "Limitations and possibilities of low cell number ChIP-seq". en. In: *BMC Genomics* 13, p. 645.

Gilmour, D S and J T Lis (July 1984). "Detecting protein-DNA interactions in vivo: distribution of RNA polymerase on specific bacterial genes". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 81.14, pp. 4275–4279.

– (Aug. 1985). "In vivo interactions of RNA polymerase II with genes of Drosophila melanogaster". en. In: *Mol. Cell. Biol.* 5.8, pp. 2009–2018.

Ginis, Irene et al. (May 2004). "Differences between human and mouse embryonic stem cells". en. In: *Dev. Biol.* 269.2, pp. 360–380.

Giresi, Paul G et al. (June 2007). "FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin". en. In: *Genome Res.* 17.6, pp. 877–885.

Görisch, Sabine M et al. (Dec. 2005). "Histone acetylation increases chromatin accessibility". en. In: *J. Cell Sci.* 118.Pt 24, pp. 5825–5834.

Grant, Charles E, Timothy L Bailey, and William Stafford Noble (Apr. 2011). "FIMO: scanning for occurrences of a given motif". en. In: *Bioinformatics* 27.7, pp. 1017–1018.

Greer, Eric Lieberman et al. (May 2015). "DNA Methylation on N6-Adenine in C. elegans". en. In: *Cell* 161.4, pp. 868–878.

Greil, Frauke, Celine Moorman, and Bas Van Steensel (2006). "DamID: mapping of in vivo protein-genome interactions using tethered DNA adenine methyltransferase". en. In: *Methods Enzymol.* 410, pp. 342–359.

Gu, Zuguang (2017). *rGREAT: Client for GREAT Analysis*.

Gutierrez-Triana, Jose Arturo et al. (Nov. 2016). "iDamIDseq and iDEAR: an improved method and computational pipeline to profile chromatin-binding proteins". en. In: *Development* 143.22, pp. 4272–4278.

Haberle, Vanja and Alexander Stark (June 2018). "Eukaryotic core promoters and the functional basis of transcription initiation". en. In: *Nat. Rev. Mol. Cell Biol.*

Hainer, Sarah J et al. (Mar. 2018a). "Profiling of pluripotency factors in individual stem cells and early embryos". en.

– (June 2018b). "Profiling of pluripotency factors in individual stem cells and early embryos". en.

Hammar, Petter et al. (June 2012). "The lac repressor displays facilitated diffusion in living cells". en. In: *Science* 336.6088, pp. 1595–1598.

Han, Xiaoping et al. (Feb. 2018). "Mapping the Mouse Cell Atlas by Microwell-Seq". en. In: *Cell* 172.5, 1091–1107.e17.

Hansen, Kasper D, Rafael A Irizarry, and Zhijin Wu (Apr. 2012). "Removing technical variability in RNA-seq data using conditional quantile normalization". en. In: *Biostatistics* 13.2, pp. 204–216.

Harikumar, Arigela and Eran Meshorer (Dec. 2015). "Chromatin remodeling and bivalent histone modifications in embryonic stem cells". en. In: *EMBO Rep.* 16.12, pp. 1609–1619.

Hass, Matthew R et al. (Aug. 2015). "SpDamID: Marking DNA Bound by Protein Complexes Identifies Notch-Dimer Responsive Enhancers". en. In: *Mol. Cell* 59.4, pp. 685–697.

Heinz, Sven et al. (May 2010). "Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities". en. In: *Mol. Cell* 38.4, pp. 576–589.

Henikoff, Steven and M Mitchell Smith (Jan. 2015). "Histone variants and epigenetics". en. In: *Cold Spring Harb. Perspect. Biol.* 7.1, a019364.

Herdman, Chelsea et al. (July 2017). "A unique enhancer boundary complex on the mouse ribosomal RNA genes persists after loss of Rrn3 or UBF and the inactivation of RNA polymerase I transcription". en. In: *PLoS Genet.* 13.7, e1006899.

Hicks, Stephanie C and Rafael A Irizarry (June 2015). "quantro: a data-driven approach to guide the choice of an appropriate normalization method". en. In: *Genome Biol.* 16, p. 117.

Hicks, Stephanie C et al. (Apr. 2018). "Smooth quantile normalization". en. In: *Biostatistics* 19.2, pp. 185–198.

Hoffman, Elizabeth A et al. (Oct. 2015). "Formaldehyde Crosslinking: A Tool for the Study of Chromatin Complexes". In: *J. Biol. Chem.* 290.44, pp. 26404–26411.

Horton, John R et al. (Apr. 2015). "Structures of Escherichia coli DNA adenine methyltransferase (Dam) in complex with a non-GATC sequence: potential implications for methylation-independent transcriptional repression". en. In: *Nucleic Acids Res.* 43.8, pp. 4296–4308.

Hu, Gangqing et al. (Feb. 2013). "H2A.Z facilitates access of active and repressive complexes to chromatin in embryonic stem cell self-renewal and differentiation". en. In: *Cell Stem Cell* 12.2, pp. 180–192.

Hurd, Paul J and Christopher J Nelson (May 2009). "Advantages of next-generation sequencing versus the microarray in epigenetic research". In: *Brief. Funct. Genomics* 8.3, pp. 174–183.

Ioannidis, John P A (Oct. 2014). "How to make more published research true". en. In: *PLoS Med.* 11.10, e1001747.

Ioannidis, John P A et al. (Feb. 2009). "Repeatability of published microarray gene expression analyses". en. In: *Nat. Genet.* 41.2, pp. 149–155.

Iwafuchi-Doi, Makiko and Kenneth S Zaret (June 2016). "Cell fate control by pioneer transcription factors". en. In: *Development* 143.11, pp. 1833–1837.

Jacinto, Filipe V, Chris Benner, and Martin W Hetzer (June 2015). "The nucleoporin Nup153 regulates embryonic stem cell pluripotency through gene silencing". en. In: *Genes Dev.* 29.12, pp. 1224–1238.

Jager, Sara B and Christian Bjerggaard Vaegter (July 2016). "Avoiding experimental bias by systematic antibody validation". en. In: *Neural Regeneration Res.* 11.7, pp. 1079–1080.

Jang, Hyonchol et al. (July 2012). "O-GlcNAc regulates pluripotency and reprogramming by directly acting on core components of the pluripotency network". en. In: *Cell Stem Cell* 11.1, pp. 62–74.

Jesus Domingues, António Miguel de et al. (Mar. 2016). "Identification of Tox chromatin binding properties and downstream targets by DamID-Seq". en. In: *Genom Data* 7, pp. 264–268.

Johnson, David S et al. (June 2007). "Genome-wide mapping of in vivo protein-DNA interactions". en. In: *Science* 316.5830, pp. 1497–1502.

Jupp, S et al. (2015). "A new Ontology Lookup Service at EMBL-EBI". In: *SWAT4LS*.

Karmodiya, Krishanpal et al. (Aug. 2012). "H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells". en. In: *BMC Genomics* 13, p. 424.

Karolchik, Donna et al. (Jan. 2004). "The UCSC Table Browser data retrieval tool". en. In: *Nucleic Acids Res.* 32.Database issue, pp. D493–6.

Karwacki-Neisius, Violetta et al. (May 2013). "Reduced Oct4 expression directs a robust pluripotent state with distinct signaling activity and increased enhancer occupancy by Oct4 and Nanog". en. In: *Cell Stem Cell* 12.5, pp. 531–545.

Kasinathan, Sivakanthan et al. (Feb. 2014). "High-resolution mapping of transcription factor binding sites on native chromatin". en. In: *Nat. Methods* 11.2, pp. 203–209.

Kawase, E et al. (June 1994). "Strain difference in establishment of mouse embryonic stem (ES) cell lines". en. In: *Int. J. Dev. Biol.* 38.2, pp. 385–390.

Kent, W J et al. (Sept. 2010). "BigWig and BigBed: enabling browsing of large distributed datasets". en. In: *Bioinformatics* 26.17, pp. 2204–2207.

Khan, Aziz et al. (Jan. 2018). "JASPAR 2018: update of the open-access database of transcription factor binding profiles and its web framework". en. In: *Nucleic Acids Res.* 46.D1, p. D1284.

Kharchenko, Peter V, Michael Y Tolstorukov, and Peter J Park (Dec. 2008). "Design and analysis of ChIP-seq experiments for DNA-binding proteins". en. In: *Nat. Biotechnol.* 26.12, pp. 1351–1359.

Kidder, Benjamin L, Gangqing Hu, and Keji Zhao (Sept. 2011). "ChIP-Seq: technical considerations for obtaining high-quality data". en. In: *Nat. Immunol.* 12.10, pp. 918–922.

Kim, Janghwan et al. (May 2011). "Direct reprogramming of mouse fibroblasts to neural progenitors". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 108.19, pp. 7838–7843.

Kind, Jop et al. (Sept. 2015). "Genome-wide maps of nuclear lamina interactions in single human cells". en. In: *Cell* 163.1, pp. 134–147.

King, Hamish W and Robert J Klose (Mar. 2017). "The pioneer factor OCT4 requires the chromatin remodeller BRG1 to support gene regulatory element function in mouse embryonic stem cells". en. In: *Elife* 6.

Kivioja, Teemu et al. (Nov. 2011). "Counting absolute numbers of molecules using unique molecular identifiers". en. In: *Nat. Methods* 9.1, pp. 72–74.

Klemm, Sandy L, Zohar Shipony, and William J Greenleaf (Apr. 2019). "Chromatin accessibility and the regulatory epigenome". en. In: *Nat. Rev. Genet.* 20.4, pp. 207–220.

Kodama, Yuichi et al. (Jan. 2012). "The Sequence Read Archive: explosive growth of sequencing data". en. In: *Nucleic Acids Res.* 40.Database issue, pp. D54–6.

Kolde, Raivo (2018). *pheatmap: Pretty Heatmaps*.

Kolesnikov, Nikolay et al. (Jan. 2015). "ArrayExpress update—simplifying data submissions". In: *Nucleic Acids Res.* 43.D1, pp. D1113–D1116.

Kolodziejczyk, Aleksandra A et al. (Oct. 2015). "Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation". en. In: *Cell Stem Cell* 17.4, pp. 471–485.

Kornberg, R D (May 1974). "Chromatin structure: a repeating unit of histones and DNA". en. In: *Science* 184.4139, pp. 868–871.

Köster, Johannes and Sven Rahmann (Oct. 2012). "Snakemake—a scalable bioinformatics workflow engine". In: *Bioinformatics* 28.19, pp. 2520–2522.

Krishnakumar, Raga et al. (Jan. 2016). "FOXD3 Regulates Pluripotent Stem Cell Potential by Simultaneously Initiating and Repressing Enhancer Activity". en. In: *Cell Stem Cell* 18.1, pp. 104–117.

Kuhn, Robert M, David Haussler, and W James Kent (Mar. 2013). "The UCSC genome browser and associated tools". en. In: *Brief. Bioinform.* 14.2, pp. 144–161.

Lambert, Samuel A et al. (Feb. 2018). "The Human Transcription Factors". en. In: *Cell* 172.4, pp. 650–665.

Landt, Stephen G et al. (Sept. 2012). "ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia". en. In: *Genome Res.* 22.9, pp. 1813–1831.

Lara-Astiaso, David et al. (Aug. 2014). "Immunogenetics. Chromatin state dynamics during blood formation". en. In: *Science* 345.6199, pp. 943–949.

Larsson, Johan (2018). *eulerr: Area-Proportional Euler and Venn Diagrams with Ellipses*.

Latchman, D S (Oct. 1993). "Transcription factors: an overview". en. In: *Int. J. Exp. Pathol.* 74.5, pp. 417–422.

– (Dec. 1997). "Transcription factors: an overview". en. In: *Int. J. Biochem. Cell Biol.* 29.12, pp. 1305–1312.

– (May 1999). "POU family transcription factors in the nervous system". en. In: *J. Cell. Physiol.* 179.2, pp. 126–133.

Lawrence, Michael, Robert Gentleman, and Vincent Carey (2009). *rtracklayer: an R package for interfacing with genome browsers*.

Lawrence, Michael et al. (2013). *Software for Computing and Annotating Genomic Ranges*.

Leinonen, Rasko et al. (Jan. 2011). "The sequence read archive". en. In: *Nucleic Acids Res.* 39.Database issue, pp. D19–21.

Lentini, Antonio et al. (July 2018). "A reassessment of DNA-immunoprecipitation-based genomic profiling". en. In: *Nat. Methods* 15.7, pp. 499–504.

Li, Dongwei et al. (Dec. 2017). "Chromatin Accessibility Dynamics during iPSC Reprogramming". en. In: *Cell Stem Cell* 21.6, 819–833.e6.

Li, Hao et al. (Sept. 2016). "Genome-wide identification and characterisation of HOT regions in the human genome". en. In: *BMC Genomics* 17.1, p. 733.

Li, Heng (Mar. 2013). "Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM". In: arXiv: `1303.3997 [q-bio.GN]`.

Li, Heng et al. (Aug. 2009). "The Sequence Alignment/Map format and SAMtools". en. In: *Bioinformatics* 25.16, pp. 2078–2079.

Li, Peipei et al. (Oct. 2015). "Comparing the normalization methods for the differential analysis of Illumina high-throughput RNA-Seq data". en. In: *BMC Bioinformatics* 16, p. 347.

Li, Renhua, Leonie U Hempel, and Tingbo Jiang (Mar. 2015). "A non-parametric peak calling algorithm for DamID-Seq". en. In: *PLoS One* 10.3, e0117415.

Liao, Yang, Gordon K Smyth, and Wei Shi (May 2013). "The Subread aligner: fast, accurate and scalable read mapping by seed-and-vote". en. In: *Nucleic Acids Res.* 41.10, e108.

Lie-A-Ling, Michael et al. (Sept. 2014). "RUNX1 positively regulates a cell adhesion and migration program in murine hemogenic endothelium prior to blood emergence". en. In: *Blood* 124.11, e11–20.

Liu, Tao et al. (Aug. 2011). "Cistrome: an integrative platform for transcriptional regulation studies". en. In: *Genome Biol.* 12.8, R83.

Liu, Xiaoyu et al. (Sept. 2016). "Distinct features of H3K4me3 and H3K27me3 chromatin domains in pre-implantation embryos". en. In: *Nature* 537.7621, pp. 558–562.

Liu, Yiyuan et al. (May 2017). "Widespread Mitotic Bookmarking by Histone Marks and Transcription Factors in Pluripotent Stem Cells". en. In: *Cell Rep.* 19.7, pp. 1283–1293.

Liu, Yue, Bin Zhao, and Yongqun He (Jan. 2016). *OGG: Ontology of Genes and Genomes*. `https://www.ebi.ac.uk/ols/ontologies/ogg`.

Liu, Ziying and W Lee Kraus (Feb. 2017). "Catalytic-Independent Functions of PARP-1 Determine Sox2 Pioneer Activity at Intractable Genomic Loci". en. In: *Mol. Cell* 65.4, 589–603.e9.

Local, Andrea et al. (Jan. 2018). "Identification of H3K4me1-associated proteins at mammalian enhancers". en. In: *Nat. Genet.* 50.1, pp. 73–82.

Lodato, Michael A et al. (Feb. 2013). "SOX2 co-occupies distal enhancer elements with distinct POU factors in ESCs and NPCs to specify cell state". en. In: *PLoS Genet.* 9.2, e1003288.

Love, Michael I, Wolfgang Huber, and Simon Anders (2014). "Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2". en. In: *Genome Biol.* 15.12, p. 550.

Luger, Karolin, Mekonnen L Dechassa, and David J Tremethick (June 2012). "New insights into nucleosome and chromatin structure: an ordered state or a disordered affair?" en. In: *Nat. Rev. Mol. Cell Biol.* 13.7, pp. 436–447.

Lun, Aaron T L and Gordon K Smyth (June 2014). "De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: controlling error rates correctly". en. In: *Nucleic Acids Res.* 42.11, e95.

– (Mar. 2016). "csaw: a Bioconductor package for differential binding analysis of ChIP-seq data using sliding windows". en. In: *Nucleic Acids Res.* 44.5, e45.

Ma, Wenxiu, William S Noble, and Timothy L Bailey (May 2014). "Motif-based analysis of large nucleotide data sets using MEME-ChIP". en. In: *Nat. Protoc.* 9.6, pp. 1428–1450.

Maksimov, Daniil A, Petr P Laktionov, and Stepan N Belyakin (Dec. 2016). "Data analysis algorithm for DamID-seq profiling of chromatin proteins in Drosophila melanogaster". en. In: *Chromosome Res.* 24.4, pp. 481–494.

Marinov, Georgi K et al. (Feb. 2014). "Large-scale quality analysis of published ChIP-seq data". en. In: *G3* 4.2, pp. 209–223.

Marshall, Owen J and Andrea H Brand (Oct. 2015). "damidseq_pipeline: an automated pipeline for processing DamID sequencing datasets". en. In: *Bioinformatics* 31.20, pp. 3371–3373.

Marshall, Owen J et al. (Sept. 2016). "Cell-type-specific profiling of protein-DNA interactions without cell isolation using targeted DamID with next-generation sequencing". en. In: *Nat. Protoc.* 11.9, pp. 1586–1598.

Marson, Alexander et al. (Aug. 2008). "Connecting microRNA genes to the core transcriptional regulatory circuitry of embryonic stem cells". en. In: *Cell* 134.3, pp. 521–533.

Martin, Marcel (May 2011). "Cutadapt removes adapter sequences from high-throughput sequencing reads". en. In: *EMBnet.journal* 17.1, pp. 10–12.

Mateo, Juan L et al. (Jan. 2015). "Characterization of the neural stem cell gene regulatory network identifies OLIG2 as a multifunctional regulator of self-renewal". en. In: *Genome Res.* 25.1, pp. 41–56.

Maza, Itay et al. (July 2015). "Transient acquisition of pluripotency during somatic cell transdifferentiation with iPSC reprogramming factors". en. In: *Nat. Biotechnol.* 33.7, pp. 769–774.

McLean, Cory Y et al. (May 2010). "GREAT improves functional interpretation of cis-regulatory regions". en. In: *Nat. Biotechnol.* 28.5, pp. 495–501.

Merkenschlager, Matthias and Elphège P Nora (Aug. 2016). "CTCF and Cohesin in Genome Folding and Transcriptional Gene Regulation". en. In: *Annu. Rev. Genomics Hum. Genet.* 17, pp. 17–43.

Meyer, Clifford A and X Shirley Liu (Nov. 2014). "Identifying and mitigating bias in next-generation sequencing methods for chromatin biology". en. In: *Nat. Rev. Genet.* 15.11, pp. 709–721.

Milagre, Inês et al. (Jan. 2017). "Gender Differences in Global but Not Targeted Demethylation in iPSC Reprogramming". en. In: *Cell Rep.* 18.5, pp. 1079–1089.

Miller, Anzy et al. (Sept. 2016). "Sall4 controls differentiation of pluripotent cells independently of the Nucleosome Remodelling and Deacetylation (NuRD) complex". en. In: *Development* 143.17, pp. 3074–3084.

Mistri, Tapan Kumar et al. (Sept. 2015). "Selective influence of Sox2 on POU transcription factor binding in embryonic and neural stem cells". en. In: *EMBO Rep.* 16.9, pp. 1177–1191.

Morgan, Martin (2017). *AnnotationHub: Client to access AnnotationHub resources*.

Morgulis, Aleksandr et al. (Jan. 2006). "WindowMasker: window-based masker for sequenced genomes". en. In: *Bioinformatics* 22.2, pp. 134–141.

Mortazavi, Ali et al. (July 2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq". en. In: *Nat. Methods* 5.7, pp. 621–628.

Moudgil, Arnav et al. (Feb. 2019). "Self-reporting transposons enable simultaneous readout of gene expression and transcription factor binding in single cells". en.

Müllner, Daniel and Others (2013). "fastcluster: Fast hierarchical, agglomerative clustering routines for R and Python". In: *J. Stat. Softw.* 53.9, pp. 1–18.

Murtha, Matthew et al. (Feb. 2015). "Comparative FAIRE-seq analysis reveals distinguishing features of the chromatin structure of ground state- and primed-pluripotent cells". en. In: *Stem Cells* 33.2, pp. 378–391.

Nebert, Daniel W (Dec. 2002). "Transcription factors and cancer: an overview". en. In: *Toxicology* 181-182, pp. 131–141.

Nègre, Nicolas et al. (June 2006). "Chromosomal distribution of PcG proteins during Drosophila development". en. In: *PLoS Biol.* 4.6, e170.

Neph, Shane et al. (July 2012). "BEDOPS: high-performance genomic feature operations". en. In: *Bioinformatics* 28.14, pp. 1919–1920.

Okashita, Naoki et al. (Dec. 2016). "PRDM14 Drives OCT3/4 Recruitment via Active Demethylation in the Transition from Primed to Naive Pluripotency". en. In: *Stem Cell Reports* 7.6, pp. 1072–1086.

Oshlack, Alicia and Matthew J Wakefield (Apr. 2009). "Transcript length bias in RNA-seq data confounds systems biology". en. In: *Biol. Direct* 4, p. 14.

Pagès, H et al. (2017). *Biostrings: Efficient manipulation of biological strings*.

Parekh, Swati et al. (May 2016). "The impact of amplification on differential expression analyses by RNA-seq". en. In: *Sci. Rep.* 6, p. 25533.

Pepke, Shirley, Barbara Wold, and Ali Mortazavi (Nov. 2009). "Computation for ChIP-seq and RNA-seq studies". en. In: *Nat. Methods* 6.11 Suppl, S22–32.

Perkins, James R et al. (2012). *ReadqPCR and NormqPCR: R packages for the reading, quality checking and normalisation of RT-qPCR quantification cycle (Cq) data*.

Pindyurin, Alexey V et al. (July 2016). "Inducible DamID systems for genomic mapping of chromatin proteins in Drosophila". en. In: *Nucleic Acids Res.* 44.12, pp. 5646–5657.

Pope, Scott D and Ruslan Medzhitov (Aug. 2018). "Emerging Principles of Gene Expression Programs and Their Regulation". en. In: *Mol. Cell* 71.3, pp. 389–397.

Quinlan, Aaron R (2002). "BEDTools: The Swiss-Army Tool for Genome Feature Analysis". In: *Current Protocols in Bioinformatics*. John Wiley & Sons, Inc.

R Core Team (2017). *R: A Language and Environment for Statistical Computing*. Vienna, Austria.

Ramírez, Fidel et al. (July 2016). "deepTools2: a next generation web server for deep-sequencing data analysis". en. In: *Nucleic Acids Res.* 44.W1, W160–5.

Ren, B et al. (Dec. 2000). "Genome-wide location and function of DNA binding proteins". en. In: *Science* 290.5500, pp. 2306–2309.

Risso, Davide et al. (Dec. 2011). "GC-content normalization for RNA-Seq data". en. In: *BMC Bioinformatics* 12, p. 480.

Ritchie, Matthew E et al. (May 2006). "Empirical array quality weights in the analysis of microarray data". en. In: *BMC Bioinformatics* 7, p. 261.

Ritchie, Matthew E et al. (Apr. 2015). "limma powers differential expression analyses for RNA-sequencing and microarray studies". en. In: *Nucleic Acids Res.* 43.7, e47.

Robinson, James T et al. (Jan. 2011). "Integrative genomics viewer". en. In: *Nat. Biotechnol.* 29.1, pp. 24–26.

Ross-Innes, Caryn S et al. (Jan. 2012). "Differential oestrogen receptor binding is associated with clinical outcome in breast cancer". en. In: *Nature* 481.7381, pp. 389–393.

Rotem, Assaf et al. (Nov. 2015). "Single-cell ChIP-seq reveals cell subpopulations defined by chromatin state". en. In: *Nat. Biotechnol.* 33.11, pp. 1165–1172.

Ruvinsky, Anatoly and Jennifer A Marshall Graves (2005). *Mammalian Genomics*. en. CABI.

Said, Harun M et al. (Jan. 2009). "Absence of GAPDH regulation in tumor-cells of different origin under hypoxic conditions in - vitro". en. In: *BMC Res. Notes* 2, p. 8.

Samuelsson, Johanna Kristina et al. (Apr. 2015). "Transposase-Assisted Chromatin immunoprecipitation (TAM-ChIP) as a tool for the simultaneous investigation of multiple targets in primary formalin-fixed, paraffin-embedded (FFPE) cancer tissues". In: *6th IMPPC Annual Conference Molecular Targets for Predictive and Personalized Medicine of Cancer*.

Scalia, Carla Rossana et al. (Jan. 2017). "Antigen Masking During Fixation and Embedding, Dissected". en. In: *J. Histochem. Cytochem.* 65.1, pp. 5–20.

Schick, Sandra et al. (Dec. 2015). "Dynamics of chromatin accessibility and epigenetic state in response to UV damage". en. In: *J. Cell Sci.* 128.23, pp. 4380–4394.

Schmidl, Christian et al. (Oct. 2015). "ChIPmentation: fast, robust, low-input ChIP-seq for histones and transcription factors". en. In: *Nat. Methods* 12.10, pp. 963–965.

Schmiedeberg, Lars et al. (Feb. 2009). "A temporal threshold for formaldehyde crosslinking and fixation". en. In: *PLoS One* 4.2, e4636.

Schuster, Eugene et al. (Aug. 2010). "DamID in C. elegans reveals longevity-associated targets of DAF-16/FoxO". en. In: *Mol. Syst. Biol.* 6, p. 399.

Sha, Ky et al. (Aug. 2010). "Distributed probing of chromatin structure in vivo reveals pervasive chromatin accessibility for expressed and non-expressed genes during tissue differentiation in C. elegans". en. In: *BMC Genomics* 11, p. 465.

Shen, Wei et al. (Oct. 2016). "SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation". en. In: *PLoS One* 11.10, e0163962.

Shen, Zuolian et al. (May 2017). "Enforcement of developmental lineage specificity by transcription factor Oct1". en. In: *Elife* 6.

Sherwood, Richard I et al. (Feb. 2014). "Discovery of directional and nondirectional pioneer transcription factors by modeling DNase profile magnitude and shape". en. In: *Nat. Biotechnol.* 32.2, pp. 171–178.

Shin, Jihoon et al. (Feb. 2016). "Aurkb/PP1-mediated resetting of Oct4 during the cell cycle determines the identity of embryonic stem cells". en. In: *Elife* 5, e10877.

Siepel, Adam et al. (Aug. 2005). "Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes". en. In: *Genome Res.* 15.8, pp. 1034–1050.

Silvester, Nicole et al. (Jan. 2015). "Content discovery and retrieval services at the European Nucleotide Archive". en. In: *Nucleic Acids Res.* 43.Database issue, pp. D23–9.

Simandi, Zoltan et al. (Aug. 2016). "OCT4 Acts as an Integrator of Pluripotency and Signal-Induced Differentiation". en. In: *Mol. Cell* 63.4, pp. 647–661.

Simon, Claire S et al. (Apr. 2017). "Functional characterisation of cis-regulatory elements governing dynamic Eomes expression in the early mouse embryo". en. In: *Development* 144.7, pp. 1249–1260.

Simon, Jeremy M et al. (Jan. 2012). "Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA". en. In: *Nat. Protoc.* 7.2, pp. 256–267.

Sims 3rd, Robert J and Danny Reinberg (Mar. 2009). "Processing the H3K36me3 signature". en. In: *Nat. Genet.* 41.3, pp. 270–271.

Skene, Peter J and Steven Henikoff (June 2015). "A simple method for generating high-resolution maps of genome-wide protein binding". en. In: *Elife* 4, e09225.

Skene, Peter J and Steven Henikoff (Jan. 2017). "An efficient targeted nuclease strategy for high-resolution mapping of DNA binding sites". In: *Elife* 6, e21856.

Sloan, Cricket A et al. (Jan. 2016). "ENCODE data at the ENCODE portal". en. In: *Nucleic Acids Res.* 44.D1, pp. D726–32.

Solomon, M J, P L Larsen, and A Varshavsky (June 1988). "Mapping protein-DNA interactions in vivo with formaldehyde: evidence that histone H4 is retained on a highly transcribed gene". en. In: *Cell* 53.6, pp. 937–947.

Soneson, Charlotte, Michael I Love, and Mark D Robinson (Dec. 2015). "Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences". en. In: *F1000Res.* 4, p. 1521.

Song, Lingyun and Gregory E Crawford (Feb. 2010). "DNase-seq: a high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells". en. In: *Cold Spring Harb. Protoc.* 2010.2, db.prot5384.

Song, Lingyun et al. (Oct. 2011). "Open chromatin defined by DNaseI and FAIRE identifies regulatory elements that shape cell-type identity". en. In: *Genome Res.* 21.10, pp. 1757–1767.

Southall, Tony D et al. (July 2013). "Cell-type-specific profiling of gene expression and chromatin binding without cell isolation: assaying RNA Pol II occupancy in neural stem cells". en. In: *Dev. Cell* 26.1, pp. 101–112.

Steglich, Babett, Shelley Sazer, and Karl Ekwall (Sept. 2013). "Transcriptional regulation at the yeast nuclear envelope". en. In: *Nucleus* 4.5, pp. 379–389.

Steube, Arndt et al. (Nov. 2017). "High-intensity UV laser ChIP-seq for the study of protein-DNA interactions in living cells". In: *Nat. Commun.* 8.1, p. 1303.

Sun, Ling V et al. (Aug. 2003). "Protein-DNA interaction mapping using genomic tiling path microarrays in Drosophila". en. In: *Proc. Natl. Acad. Sci. U. S. A.* 100.16, pp. 9428–9433.

Swift, Joseph and Gloria M Coruzzi (Jan. 2017). "A matter of time - How transient transcription factor interactions create dynamic gene regulatory networks". en. In: *Biochim. Biophys. Acta Gene Regul. Mech.* 1860.1, pp. 75–83.

Takahashi, Kazutoshi and Shinya Yamanaka (Mar. 2016). "A decade of transcription factor-mediated reprogramming to pluripotency". en. In: *Nat. Rev. Mol. Cell Biol.* 17.3, pp. 183–193.

Tarailo-Graovac, Maja and Nansheng Chen (Mar. 2009). "Using RepeatMasker to identify repetitive elements in genomic sequences". en. In: *Curr. Protoc. Bioinformatics* Chapter 4, Unit 4.10.

Taslim, Cenny, Tim Huang, and Shili Lin (June 2011). "DIME: R-package for identifying differential ChIP-seq based on an ensemble of mixture models". en. In: *Bioinformatics* 27.11, pp. 1569–1570.

Team, Bioconductor Core and Bioconductor Package Maintainer (2016). *TxDb.Mmusculus.UCSC.mm10*
Annotation package for TxDb object(s).

Thomas, Reuben et al. (May 2017). "Features that define the best ChIP-seq peak calling algorithms". en. In: *Brief. Bioinform.* 18.3, pp. 441–450.

Thornton, Brenda and Chhandak Basu (Mar. 2011). "Real-time PCR (qPCR) primer design using free online software". en. In: *Biochem. Mol. Biol. Educ.* 39.2, pp. 145–154.

Tosti, Luca et al. (Mar. 2018). "Mapping transcription factor occupancy using minimal numbers of cells in vitro and in vivo". en. In: *Genome Res.*

Tremethick, David J (Feb. 2007). "Higher-order structures of chromatin: the elusive 30 nm fiber". en. In: *Cell* 128.4, pp. 651–654.

Tsompana, Maria and Michael J Buck (Nov. 2014). "Chromatin accessibility: a window into the genome". en. In: *Epigenetics Chromatin* 7.1, p. 33.

Tu, Shengjiang et al. (June 2016). "Co-repressor CBFA2T2 regulates pluripotency and germline development". en. In: *Nature* 534.7607, pp. 387–390.

Untergasser, Andreas et al. (Aug. 2012). "Primer3–new capabilities and interfaces". en. In: *Nucleic Acids Res.* 40.15, e115.

Valieris, Renan (2018). *parallel-fastq-dump*.

Van Steensel, B, J Delrow, and S Henikoff (Mar. 2001). "Chromatin profiling using targeted DNA adenine methyltransferase". en. In: *Nat. Genet.* 27.3, pp. 304–308.

Van Steensel, B and S Henikoff (Apr. 2000). "Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase". en. In: *Nat. Biotechnol.* 18.4, pp. 424–428.

VanInsberghe, Michael et al. (Jan. 2018). "Highly multiplexed single-cell quantitative PCR". en. In: *PLoS One* 13.1, e0191601.

Vaquerizas, Juan M et al. (Apr. 2009). "A census of human transcription factors: function, expression and evolution". en. In: *Nat. Rev. Genet.* 10.4, pp. 252–263.

Venkatesh, Swaminathan and Jerry L Workman (Mar. 2015). "Histone exchange, chromatin structure and the regulation of transcription". en. In: *Nat. Rev. Mol. Cell Biol.* 16.3, pp. 178–189.

Vogel, Maartje J et al. (Dec. 2006). "Human heterochromatin proteins form large domains containing KRAB-ZNF genes". en. In: *Genome Res.* 16.12, pp. 1493–1504.

Wang, Li et al. (May 2014). "INO80 facilitates pluripotency gene activation in embryonic stem cell self-renewal, reprogramming, and blastocyst development". en. In: *Cell Stem Cell* 14.5, pp. 575–591.

Wang, Su et al. (Dec. 2013). "Target analysis by integration of transcriptome and ChIP-seq data with BETA". en. In: *Nat. Protoc.* 8.12, pp. 2502–2515.

Wang, Ying, Xiaoman Li, and Haiyan Hu (Feb. 2014). "H3K4me2 reliably defines transcription factor binding regions in different cells". en. In: *Genomics* 103.2-3, pp. 222–228.

Wapinski, Orly L et al. (Oct. 2013). "Hierarchical mechanisms for direct reprogramming of fibroblasts to neurons". en. In: *Cell* 155.3, pp. 621–635.

Whyte, Warren A et al. (Apr. 2013). "Master transcription factors and mediator establish super-enhancers at key cell identity genes". en. In: *Cell* 153.2, pp. 307–319.

Wickham, Hadley (2009). *ggplot2: Elegant Graphics for Data Analysis*. Springer-Verlag New York.

Wilkinson, Adam C, Hiromitsu Nakauchi, and Berthold Göttgens (Oct. 2017). "Mammalian Transcription Factor Networks: Recent Advances in Interrogating Biological Complexity". en. In: *cels* 5.4, pp. 319–331.

Wolfram, Verena et al. (Aug. 2012). "The LIM-homeodomain protein islet dictates motor neuron electrical properties by regulating K(+) channel expression". en. In: *Neuron* 75.4, pp. 663–674.

Woodcock, C L, L L Frado, and J B Rattner (July 1984). "The higher-order structure of chromatin: evidence for a helical ribbon arrangement". en. In: *J. Cell Biol.* 99.1 Pt 1, pp. 42–52.

Wreczycka, Katarzyna et al. (Mar. 2017). "HOT or not: Examining the basis of high-occupancy target regions". en.

Wu, Feinan and Jie Yao (Aug. 2013). "Spatial compartmentalization at the nuclear periphery characterized by genome-wide mapping". en. In: *BMC Genomics* 14, p. 591.

Wu, Tao P et al. (Apr. 2016). "DNA methylation on N(6)-adenine in mammalian embryonic stem cells". en. In: *Nature* 532.7599, pp. 329–333.

Xiao, Chuan-Le et al. (July 2018). "N6-Methyladenine DNA Modification in the Human Genome". en. In: *Mol. Cell* 71.2, 306–318.e7.

Xu, Tianlei et al. (Mar. 2015). "Base-resolution methylation patterns accurately predict transcription factor bindings in vivo". en. In: *Nucleic Acids Res.* 43.5, pp. 2757–2766.

Yaffe, Eitan and Amos Tanay (Oct. 2011). "Probabilistic modeling of Hi-C contact maps eliminates systematic biases to characterize global chromosomal architecture". en. In: *Nat. Genet.* 43.11, pp. 1059–1065.

Yang, Shen-Hsi et al. (June 2014). "Otx2 and Oct4 drive early enhancer activation during embryonic stem cell transition from naive pluripotency". en. In: *Cell Rep.* 7.6, pp. 1968–1981.

Yardımcı, Galip Gürkan et al. (Oct. 2014). "Explicit DNase sequence bias modeling enables high-resolution transcription factor footprint detection". en. In: *Nucleic Acids Res.* 42.19, pp. 11865–11878.

Yin, Hang et al. (Dec. 2011). "A high-resolution whole-genome map of key chromatin modifications in the adult Drosophila melanogaster". en. In: *PLoS Genet.* 7.12, e1002380.

Young, Matthew D et al. (Feb. 2010). "Gene ontology analysis for RNA-seq: accounting for selection bias". en. In: *Genome Biol.* 11.2, R14.

Yue, Feng et al. (Nov. 2014). "A comparative encyclopedia of DNA elements in the mouse genome". en. In: *Nature* 515.7527, pp. 355–364.

Zhang, Guoqiang et al. (May 2015). "N6-Methyladenine DNA Modification in Drosophila". en. In: *Cell* 161.4, pp. 893–906.

Zhang, Yong et al. (Sept. 2008). "Model-based analysis of ChIP-Seq (MACS)". en. In: *Genome Biol.* 9.9, R137.

Zheng, Wei, Lisa M Chung, and Hongyu Zhao (July 2011). "Bias detection and correction in RNA-Sequencing data". en. In: *BMC Bioinformatics* 12, p. 290.

Zhou, Quan et al. (Apr. 2017). "A mouse tissue transcription factor atlas". en. In: *Nat. Commun.* 8, p. 15089.

Zhu, Yuelin et al. (Dec. 2008). "GEOmetadb: powerful alternative search engine for the Gene Expression Omnibus". en. In: *Bioinformatics* 24.23, pp. 2798–2800.

Zhu, Yuelin et al. (Jan. 2013). "SRAdb: query and use public next-generation sequencing data from within R". en. In: *BMC Bioinformatics* 14, p. 19.

Zilfou, Jack T and Scott W Lowe (Nov. 2009). "Tumor suppressive functions of p53". en. In: *Cold Spring Harb. Perspect. Biol.* 1.5, a001883.

Zwart, Wilbert et al. (Apr. 2013). "A carrier-assisted ChIP-seq method for estrogen receptor-chromatin interactions from breast cancer core needle biopsy samples". en. In: *BMC Genomics* 14, p. 232.