Characterisation of the epigenome of an *in vitro* model of chondrogenesis

Kathleen Cheung

Thesis submitted for the degree of Doctor of Philosophy

Institute of Genetic Medicine
Newcastle University

March 2018

Abstract

Chondrogenesis, the differentiation of mesenchymal progenitor cells from the mesoderm germ layer during embryonic development, is partly regulated by epigenetic mechanisms such as histone modifications and DNA methylation. Histone proteins possess protruding N-terminal tails which may be post-translationally modified to alter the structure of chromatin resulting in a change in the accessibility of genes to the transcription machinery. In the genome, histone modifications mark *cis*-regulatory elements such as gene promoters and enhancers while DNA methylation occurs on cytosine residues at CpG sites and typically leading to transcriptional repression.

The aim of this project was to characterise the epigenome during *in vitro* differentiation of human mesenchymal stem cells (hMSCs) into chondrocytes. Chromatin immunoprecipitation followed by next-generation sequencing (ChIP-seq) was used to assess genome-wide a range of N-terminal post-transcriptional modifications (marks) to histone H3 lysines (H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3) in both hMSCs and differentiated chondrocytes. Chromatin states were characterised using the software ChromHMM and *cis*-regulatory elements were identified. Integration of DNA methylation data with chondrogenesis chromatin states revealed that enhancers marked by H3K4me1 and H3K27ac were hypomethylated during *in vitro* chondrogenesis. Similarity analysis between chondrogenesis chromatin states with epigenomes of cell types defined by the Roadmap Epigenomics project revealed that enhancers are more distinct between cell types compared to other chromatin states. SOX9 is regarded as the master transcription factor for chondrogenesis. An external mouse Sox9 ChIP-seq dataset was used to identify super enhancers in chondrocytes. Luciferase reporter assays showed that selected regions of super enhancers exhibit independent enhancer activity.

In conclusion, we observed that CpG sites within enhancers are de-methylated during hMSC differentiation into chondrocytes and propose that gene transcription during chondrogenesis is regulated by epigenetic changes at enhancers. Epigenetic changes have been implicated in cartilage diseases and greater understanding of the chondrocyte epigenome may have potential therapeutic value.

Acknowledgements

Many thanks to my supervisors Prof. David Young, Dr. Matt Barter and Dr. Carole Proctor for their expert guidance and advice throughout this project. Without them, I would never have got this far. I would also like to thank Dr. Louise Reynard for her helpful input into my project. I also express my gratitude to the Bioinformatics Support Unit, in particular Andrew Skelton, for technical expertise and computational resources. A huge thank you to the entire research group for their support and encouragement these past four years. It has been a fantastic experience.

Above all, I would like to thank my parents for teaching me the value of education and the importance of pursuing aspirations. They worked hard and sacrificed a great deal so my brothers and I could have a better life. Last but not least, I would like to thank Noodle the cat. He didn’t provide any scientific expertise nor was particularly supportive but has been a comforting companion during the joys and hardships of the past few years.

Data provision

Chondrogenesis cDNA microarray data was kindly provided by Dr. Matt Barter (Newcastle University). Chondrogenesis DNA 450K array data was generated by Dr. Matt Barter and Dr. Catherine Bui (Newcastle University). Adipogenesis and osteoblastogenesis DNA 450K array data was generated by Dr. Ruddy Gomez-Bahamonde (Newcastle University).

Funding

This PhD project was funded by the Centre for Integrated research into Musculoskeletal Ageing (CIMA).

Contents

List of tables

Chapter 1. Introduction

Chapter 2. Materials and methods

Chapter 3. Histone ChIP-seq quality control, alignment and peak calling

Chapter 4. Integration of chondrogenesis histone ChIP-seq to RNA-seq data

List of figures

Chapter 1. Introduction

Chapter 2. Materials and methods

Chapter 3. Histone ChIP-seq quality control, alignment and peak calling

Chapter 4. Integration of chondrogenesis histone ChIP-seq to RNA-seq data

Chapter 5. Correlation of DNA methylation to chromatin states

Chapter 6. Identification and characterisation of super enhancers in chondrogenesis

Appendix I

Appendix II

Appendix III

List of abbreviations

5hmC 5-hydroxymethylcytosine

5mC 5-methylcytosine

AD-MSC Adipose derived MSC

AP-1 Activator Protein 1

APS Ammonium persulphate

ARUK Arthritis Research UK

BM-MSC Bone marrow derived MSC

BMP Bone Morphogenetic Protein

bp Base pairs

ChIP Chromatin immunoprecipitation

ChIP-seq ChIP coupled with high throughput sequencing

ChIP-qPCR ChIP coupled with qPCR

CRISPR Clustered Regularly Interspaced Short Palindromic Repeats

dH2O Distilled H2O

DMB Dimethyl-Methylene Blue

DMEM Dulbecco's Modified Eagle's Medium

DNA Deoxyribonucleic acid

dNTP Deoxyribonucleotide triphosphate

ECL Electrochemiluminescence

ECM Extracellular Matrix

EDTA Ethylenediaminetetraacetic acid

ENCODE Encyclopaedia of DNA Elements

EtBr Ethidium bromide

ESC Embryonic Stem Cell

FBS Foetal bovine serum

FDR False discovery rate

FGFs Fibroblast Growth Factors

Fig Figure

FRiP Fraction of reads in peaks

GAG Glycosaminogycan

GAPDH Glyceraldehyde 3-phosphate dehydrogenase

GLI glioma-associated oncogene

GO Gene ontology

HAT Histone acetyltransferases

HBB Human Haemoglobin

HCL Hydrochloric acid

HDAC Histone Deacetylase

HDACi Histone Deacetylase inhibitor

HEK293T Human Embryonic Kidney cells 293

HMG High mobility group

HSC Hematopoietic stem cell

IDR Irreproducible discovery rate

IHEC International Human Epigenome Consortium

IGF Insulin Growth Factor

IgG Immunoglobulin G

IGV Integrative Genomics Viewer

IP Immunoprecipitation

LB Luria-Broth

LGB Lower gel buffer

LncRNA Long non-coding RNA

miRNA Micro-RNAs

M-MLV Moloney Murine Leukaemia Virus

MMP Matrix metalloproteinase

MSC Mesenchymal Stem Cells

NSC Normalized Strand Cross-correlation coefficient

OA Osteoarthritis

PBC PCR bottleneck coefficient

PBS Phosphate Buffered Saline

PCA Principal Component Analysis

PCR Polymerase Chain Reaction

PEV Position-effect variegation

QC Quality Control

qPCR Real time PCR

RLU Relative Light Unit

RNA Ribonucleic acid

RNAPII RNA Polymerase II

RNA-seq RNA sequencing

RPM Reads per million

RRBS Reduced representation bisulfite sequencing

RSC Relative Strand Cross-correlation coefficient

SDS-PAGE Sodium Dodecyl Sulfate - **Polyacrylamide** Gel Electrophoresis

siRNA small interfering RNA

TAE Tris-Acetate-EDTA

TEMED Tween-20, N,N,N’,N’-Tetramethylethylenediamine

TGF-β Treansforming Growth Factor- beta

TPM Transcripts per million

TSS Transcription Start Site

TTS Transcription Termination Site

UCSC University of California, Santa Cruz

UK United Kingdom

UTR Untranslated Region

Wnt Wingless-related integration site

Chapter 1. Introduction

1.1 Articular cartilage

Chondrogenesis is the differentiation of mesenchymal progenitor cells into chondrocytes from the mesoderm germ layer during embryonic development. Chondrocytes are the only cell type found in articular cartilage within synovial joints such as the knee or hip and are responsible for the secretion of extracellular matrix (ECM) and homeostasis of cartilage. Articular cartilage is the hyaline cartilage found at bone ends in the synovial joint (Fig. 1.1) and aids in movement of the joint and protects bone ends from degradation.

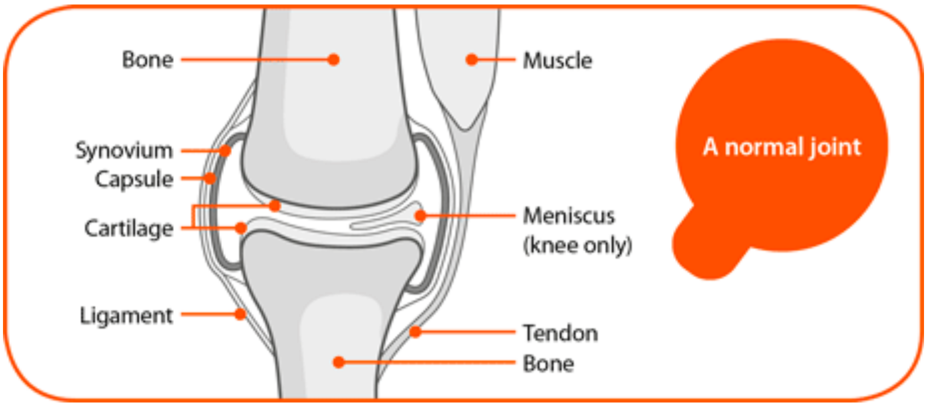


Figure 1.1 – Structure of a healthy synovial joint. Articular cartilage covers the end of bones within the synovial joint, protecting the bone ends and aiding in lubricating and movement of the joint. Chondrocytes are present only within the articular cartilage. Synovial fluid is present in the cavity between bone ends, covered by the synovial membrane. The articular capsule surrounds the synovial joint. Image from ARUK.

Chondrocytes synthesise, secrete and maintain the ECM proteins in their surrounding environment. Type II collagen is the main collagen present in articular cartilage and is highly expressed by chondrocytes. A network of collagen fibrils run through articular cartilage, along with proteoglycans such as perlecan which is required for fibrillogenesis (Kvist et al, 2006). The main proteoglycan in cartilage is aggrecan, which exists in large aggregates and is responsible for the hydrated gel structure seen in articular cartilage (Watanabe et al, 1998). As well as secreting ECM proteins, chondrocytes also produce and secrete enzymes responsible for the degradation of the ECM such as matrix metalloproteinases (MMPs). There are four main zones in articular cartilage (Fig. 1.2). The composition and properties of cartilage varies between each zone. Water is a major component of articular cartilage; up to 80% of articular cartilage is water, with the highest concentration of water in the tangential zone. The morphology of chondrocytes also changes depending on which zone they are found in, with cells becoming larger and more spherical in the deeper zones.

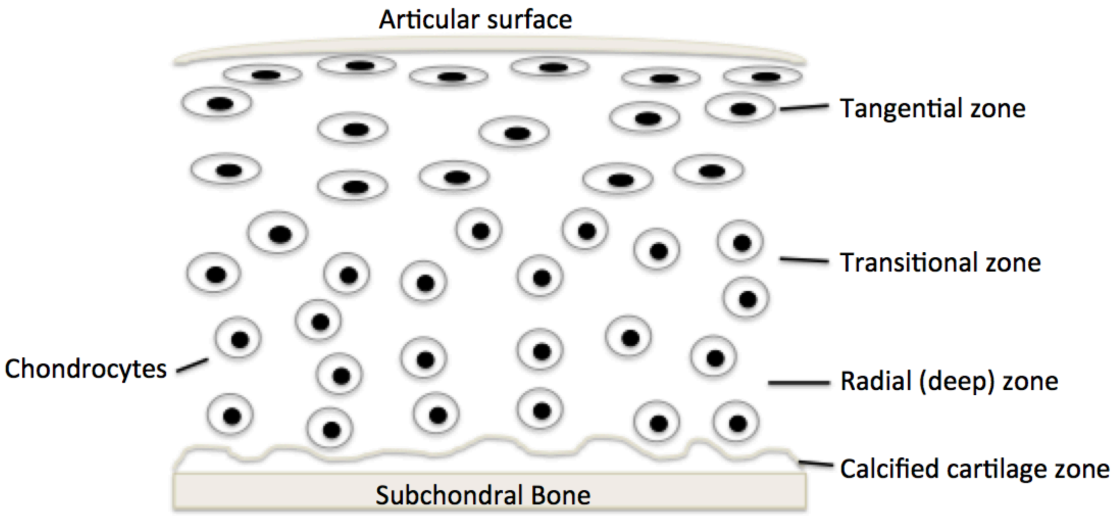


Figure 1.2 - Schematic cross section through articular cartilage. There are four main zones in articular cartilage, each with different ECM proteins and water compositions. The morphology of chondrocytes differs depending on where they are found within the cartilage, with spherical cells found near the subchondral site in the deep zone and becoming flatter towards the articular surface. The tangential (aka superficial) zone is near the articular surface and is in contact with synovial fluid in the joint. This zone acts as a defence for lower zones and is the site of most mechanical stress. The tangential zone consists of densely packed collagen type II and type IX fibres which contributes to the tensile properties of cartilage and allows it to resist the forces and stresses caused by joint movement. The transitional or middle zone contains thicker collagen fibrils and chondrocytes are sparse throughout this zone. The radial zone has the greatest resistance to mechanical loading and consists of the thickest collagen fibrils and more proteoglycan content than other zones. Chondrocytes are arranged in columns in this zone.

Articular cartilage is an avascular tissue and ECM proteins have a long half-life, factors which contribute to the low turnover and repair capacity of cartilage. Chondrocytes in adult cartilage are therefore comparatively inert and repair is a slow process. Healing in cartilage is incomplete and results in a fibrocartilage scar which is inferior to normal cartilage (Gomoll and Minas, 2014). Studies into the regeneration of articular cartilage have focused on the chondrocyte as they are responsible for the maintenance of cartilage integrity. Although articular cartilage in the adult body has a limited potential for regeneration, during development it is formed and remodelled on a large scale (Goldring, 2012). The limited capacity of mature chondrocytes for healing makes cartilage injury and disease extremely difficult to treat. However, during development, cartilage is extremely amenable to repair and remodelling. At this time, cartilage is able to successfully heal injuries without leaving a fibrocartilage scar (Namba et al, 1998).

1.2 Cartilage development

During development, chondrocytes are derived from the skeletal blastoma, a dense cluster of mesenchymal progenitors found in the mesoderm germ layer. Chondrogenesis is a multi-step tightly regulated process mediated by different transcription and growth factors at each stage (Fig. 1.3). Each stage is characterised by gene expression and cell morphology changes. The first step in chondrocyte differentiation is the migration of progenitor cells to the site of chondrogenesis, followed by interactions of progenitors to epithelial cells and the third stage is the condensation of chondrocyte progenitors. The critical stage in the development of mesenchymal tissues is condensation. This step involves the aggregation of progenitor cells that accumulate together to form a specific tissue type. Signalling molecules and cellular interactions are crucial for development of cartilage and other tissues. The lineage of mesenchymal progenitors is determined by this stage and cells express markers and genes specific for their intended cell type. Chondrogenic progenitors express transcription factors such as SOX9 pre-condensation which determine them for the chondrocyte cell lineage (Lorda-Diez et al, 2011). Mesenchymal stem cells (MSCs; discussed in section 1.4) found in the adult body can differentiate into chondrocytes and many *in vitro* studies elucidating the mechanisms of chondrogenesis have been performed using MSC chondrogenesis.

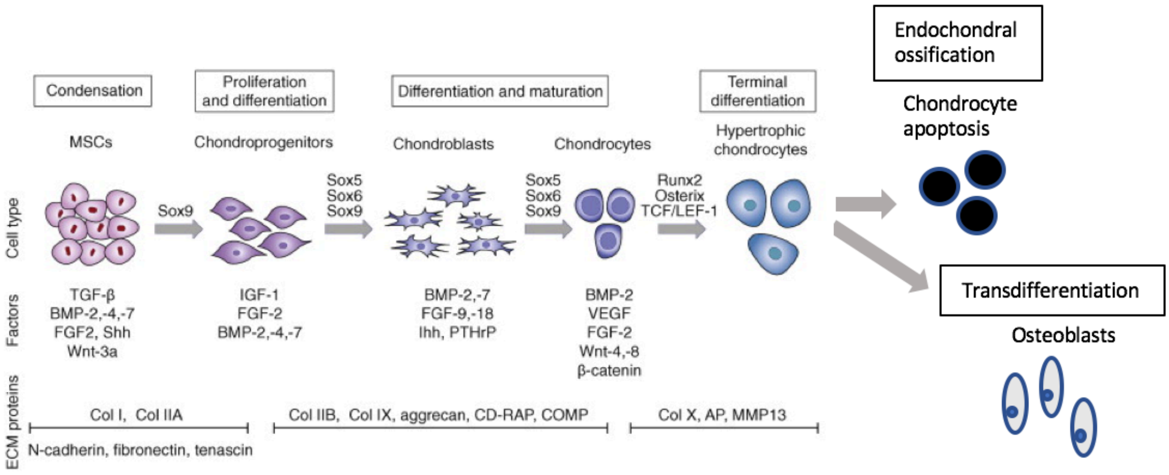


Figure 1.3 – Stages of chondrogenesis. Following MSC condensation, cells determined for the chondrogenic lineage undergo further differentiation steps. SOX9 is required for the initiation of MSC differentiation into chondrocytes and is involved throughout chondrogenesis. Each step is characterised by the engagement of various growth factors and changes in gene expression and cell morphology. The final step of chondrogenesis is the hypertrophy of chondrocytes. After this, the majority of chondrocytes apoptose and endochondral ossification occurs. However, some chondrocytes may transdifferentiate into osteoblasts. Image adapted from Vinatier et al, 2009.

The master transcriptional regulators for chondrogenesis are the SOX (**S**RY-related high-mobility-group (HMG) b**ox**) trio of transcription factors, *SOX9*, *SOX5* and *SOX6* (Akiyama, 2008). These transcription factors activate expression of genes that drive MSCs down the chondrocyte cell fate. *SOX9* in particular is important for chondrogenesis and acts upstream of *SOX5* and *SOX6*. Without *SOX9*, chondrogenesis cannot occur (Mori-Akiyama et al, 2003). The exact molecular mechanisms of *SOX9* are not fully understood but it acts in concordance with multiple signalling pathways. For example, the TGFβ signalling pathway in particular is important in early chondrogenesis (Kawakami et al, 2006). The SOX9 protein binds to DNA to regulate transcription of genes either as a homodimer or a heterodimer with different binding partners. At different stages of chondrogenesis, SOX9 binds as a heterodimer to other transcription factors to regulate different genes and may positively or negatively regulate transcription depending on the binding partner. SOX9 partners with GLI proteins to repress *COL10A1*, a gene up-regulated in hypertrophic chondrocytes (Leung et al, 2011). However, SOX9 also co-binds with the AP-1 transcription factor to promote chondrocyte hypertrophy (He et al, 2016).

Chondrogenesis requires a balance of pro- and anti-chondrogenic factors for normal development. For example, BMP signalling is necessary for differentiation of MSCs into chondrocytes. However, the BMP antagonist Noggin is also required; lack of Noggin leads to abnormal joint development (Brunet et al, 1998). Homozygous knockout of Sox9 is embryonic lethal in mice (Chaboissier et al, 2004) whereas heterozygous knockout in mice leads to severe abnormal cartilage development and perinatal lethality (Bi et al, 2001). Overexpression of SOX9 also induces abnormal cartilage and skeletal development. SOX9 interacts with β-catenin and the canonical Wnt pathway to decrease levels of cyclin D, inhibiting chondrocyte proliferation (Akiyama et al, 2004). SOX9 therefore affects chondrogenesis in a concentration dependent manner. SOX9 is also involved in other developmental processes such as sex determination in mammals, neurogenesis, lung, pancreas and liver development (Koopman 2001; Jo et al, 2014).

The initial steps of chondrogenesis and osteoblastogenesis are identical and MSCs differentiate into either chondrocytes or osteoblasts depending on the signals received from the interaction with epithelial cells (Hall and Miyake, 2000). Furthermore, during the early stages of skeletal development, most of the skeleton is formed from the cartilage anlagen, a template for bone formation. As development progresses, the transitory cartilaginous skeleton is gradually replaced by bone, a process called endochondral ossification. Most of the cartilage is slowly replaced except articular cartilage in locations such as synovial joints which are left intact and remain present in the adult body. During endochondral ossification, chondrocytes enter a hypertrophic stage, the terminal stage of chondrocyte differentiation. The ECM calcifies and hardens and chondrocytes undergo apoptosis, leaving cavities that osteoblasts and osteoclasts migrate into (Mackie et al, 2008). There is also evidence that chondrocytes can transdifferentiate into osteoblasts. Although the majority of chondrocytes die during endochondral ossification, a subset of chondrocytes are able to transdifferentiate into osteoblasts (Thesingh et al, 1991; Yang et al, 2014; Park et al, 2015). This transdifferentiation process is also seen during the healing of bone fractures in adults (Zhou et al, 2014).

During development, large scale changes in gene expression occur in a temporal manner. Gene expression during chondrogenesis is in part regulated by epigenetic mechanisms such as DNA methylation and histone modifications (Furumatsu and Ozaki, 2010; Hata, 2015). MicroRNAs (miRNAs) and long non-coding RNAs (lncRNAs) also play a role in chondrogenesis (Yang et al, 2011; Barter et al, 2017; Tian et al, 2016). For example, the miRNA miR-140 was found to be important for regulating chondrogenesis related genes during differentiation (Barter et al, 2015). Furthermore, genome wide histone modification changes have been observed during *in vitro* differentiation of MSCs into chondrocytes (Herlofsen et al, 2013).

The epigenetic changes that occur during chondrogenesis are not yet fully understood. Understanding how genes are regulated is important for understanding how they become dysregulated during disease processes.

1.3 Cartilage diseases

Mutations in chondrogenesis-related genes and aberrant gene expression can lead to congenital cartilage disorders. Chondrodysplasias describe a group of skeletal conditions characterised by abnormal bone and cartilage development. Chondrodysplasias can be caused by mutations in regulatory regions of cartilage genes as well as the gene itself. Deletions in a distal regulatory region of the *SOX9* gene and within the SOX9 HMG domain both lead to campomelic dysplasia in humans (Meyer et al, 1997; Wunderle et al, 1998). Mutations in collagen genes *COL2A1* and *COL9A1-3* can also lead to some forms of chondrodysplasia (Tiller et al, 1995; Paassilita et al, 1999).

Cartilage diseases can also be acquired post-development. Osteoarthritis (OA) is an age-related cartilage disease characterised by progressive degradation of the articular cartilage in synovial joints such as the hip and knee joints (Fig. 1.4). Symptoms of OA can include pain, stiffness, lack of flexibility and loss of mobility. The most common joints affected are the knee, hands and hips (Wood et al, 2013). Gene expression changes are seen in OA chondrocytes. One of the major changes seen in the progression of OA is the upregulation of *MMP13*, which encodes for a matrix matalloproteinase that degrades type II collagen (Wang et al, 2013). The composition of articular cartilage also changes in OA; the amount of water present increases to over 90%, affecting the load bearing properties of the cartilage (Bhosale and Richardson, 2008).

Although age is a major risk factor, OA is distinct from normal synovial joint ageing and is not an inevitable result of age (Loeser et al, 2016). Genetic susceptibility loci that predisposes individuals to developing OA have been identified (Zeggini et al, 2012). However, OA is a complex multifactorial disease and more research is necessary to elucidate why and how individuals develop OA. There are currently no disease modifying drugs available for OA and patients are advised to manage symptoms using pain, anti-inflammatory medicine and lifestyle changes. End stage OA may be treated using joint replacement surgery which replaces the affected joint with an artificial prosthesis. Although joint replacements are generally successful, the artificial joint has a limited lifespan of 15-20 years (Arthritis Research UK). This means a patient may need to undergo multiple joint replacements depending on their lifespan. Furthermore, as OA is typically related to age, patients are often elderly which is associated with increased risk of complications when undergoing surgical procedures (Turrentine et al, 2006). OA presents a significant economic burden to the UK; joint replacements cost the NHS £852 million in 2010. Indirect costs of OA, such as loss of productivity, unemployment and disability payments, exceed £2.4 billion (Chen et al, 2012). Increases in global life expectancy have led to an increase in age-related diseases such as OA. As lifespan is set to increase further, there is a crucial need to develop better healthcare solutions to improve the quality of life for the elderly population (Jin et al, 2015).

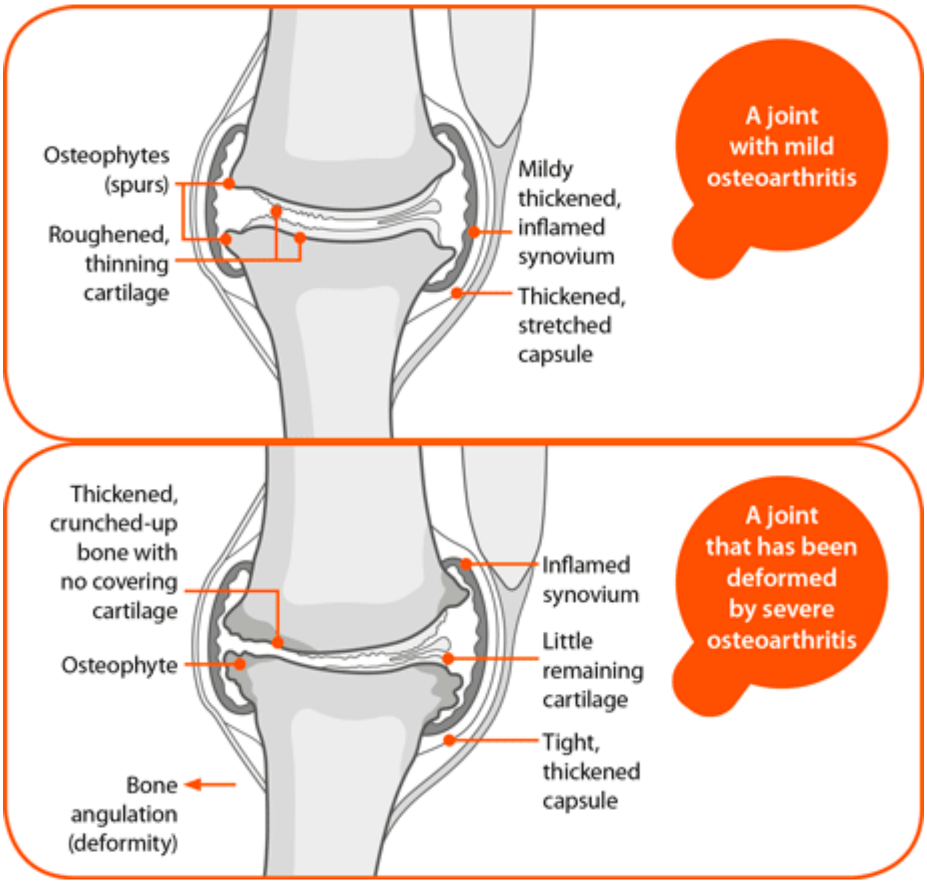

Osteophytes (spurs)
Roughened, thinning cartilage
Mildy thickened, inflamed synovium
Thickened, stretched capsule
A joint with mild osteoarthritis
Thickened, crunched-up bone with no covering cartilage
Osteophyte
Bone angulation (deformity)
Inflamed synovium
Little remaining cartilage
Tight, thickened capsule
A joint that has been deformed by severe osteoarthritis

Figure 1.4 – Synovial joint with OA. Articular cartilage covering the bone ends is progressively degraded until the bone is exposed. In severe OA, the unprotected exposed bone is also degraded. Although OA was traditionally thought to be non-inflammatory, it is now known that affected joints may be characterised by a low level of chronic inflammation, particularly in the synovium (Sokolove and Lepus, 2013). Image from ARUK.

Cartilage diseases such as OA are difficult to treat due to the limited responses of chondrocytes and the low capacity for regeneration in cartilage tissue. Studies show that the pathways involved in OA development are also involved in chondrogenesis (Goldring, 2012; Pitsillides and Beier, 2013). Epigenetic changes have also been implicated in OA. Changes in DNA methylation have been characterised in patients with hip and knee OA compared to non-OA controls (Rushton et al, 2012). Loss of DNA methylation in the enhancer of the inducible nitric oxide synthase gene is associated with gene expression changes in human OA chondrocytes (de Andrés et al, 2013). Loss of miR-140 in mice leads to impaired cartilage development and an OA-like phenotype (Miyaki et al, 2010). Long non-coding RNAs (lncRNAs) also play roles in chondrogenesis, cartilage homeostasis and disease (Huynh et al, 2017). These studies show that non-coding regions of the genome and regulatory elements can be just as important as coding genes during the development of cartilage diseases.

Congenital cartilage diseases such as chondrodysplasias are caused by abnormal development whereas acquired cartilage diseases such as OA have shown a link to developmental pathways. Therefore, investigations into the epigenome during chondrogenesis may lead to a better understanding of OA and other cartilage diseases. Due to the inherently low regeneration potential of cartilage in the adult body, tissue engineering and stem cell transplantation methods have been explored as a way to replace damaged or diseased cartilage.

1.4 Mesenchymal stem cells: use in research and regenerative medicine

Stem cells, also known as progenitor cells, are defined as undifferentiated or unspecialised cells that are able to self-regenerate as well as differentiate into multiple more specialized cell types. In mammals, there are many different stem cells characterised by where they originate and the cell types they can differentiate into. For example, embryonic stem cells (ESCs) are pluripotent stem cells capable of differentiating into any cell type in the body and are therefore the starting point of development of all tissue types in the body. All cells from the three primary germ layers - the ectoderm, endoderm and mesoderm, ultimately derive from ESCs. ESCs only exist during the early stages of development and can be extracted from the blastocyst stage of embryogenesis, although this results in the destruction of the embryo which raises difficult moral and ethical questions (Lo and Parham, 2009).

Post development, maintenance of tissues and creation of new cells is orchestrated by adult or somatic stem cells. In recent years, adult stem cells have been a large focus of research into regenerative medicine. Stem cells can be used to replace or repair injured and diseased tissues, especially important in tissues with constrained regeneration capacities such as tendon, nerve tissue and cartilage. Human adult stem cells in particular have received a lot of attention due to fewer ethical concerns compared to hESCs and the potential for personalised, autologous cell implants. However, unlike ESCs, adult stem cells do not have unlimited differentiation potential and are typically tissue specific. There are many adult stem cells, each able to differentiate into a limited subset of cell types depending on their stem cell niche. One example is haematopoietic stem cells (HSCs), these give rise to cells of the haematopoietic system, including cells involved in the immune response such as lymphocytes and monocytes (Eaves, 2015). In adults, HSCs are found in bone marrow. Other adult stem cells include epithelial stem cells responsible for the maintenance of epithelium in organs such as skin and intestines (Blanpain et at, 2007), and neural stem cells which generate cells in the nervous system including neurons and glial cells (Kornblum, 2007).

Mesenchymal stem cells (MSCs) are multipotent adult stem cells able to differentiate into a variety of cell types, mainly those comprising the musculoskeletal system. MSCs are also known as mesenchymal stromal cells. MSCs are often described as having a fibroblast-like cell morphology with thin, elongated cell bodies (Haniffa et al, 2009). MSCs are derived from the mesoderm germ layer and in adults are found in a range of stem cell niches including bone marrow, synovium, umbilical cord, muscle and adipose tissue (Kolf et al, 2007). MSCs are a heterogeneous population of stem cells and are broadly defined as cells capable of self-renewal plus having the potential to differentiate into osteoblasts, chondrocytes and adipocytes *in vitro* (Nombela-Arrieta et al, 2011). Due to their heterogeneous nature, there is no single cell surface marker for MSCs. Instead, they are identified by the presence of multiple cell surface antigen markers such as CD90, CD73 and CD105 whilst lacking expression of markers of other stem cell types such as CD45, CD34, CD14 or CD11b, CD79-alpha or CD19, and HLA-DR surface molecules. MSCs must also adhere to standard tissue culture plastics. These criteria are stipulated by the International Society for Cellular Therapy (Dominici et al, 2006) and cells must meet these criteria to be classed as MSCs. Additional to the common characteristics mentioned, MSCs from different stem cell niches can show different antigen markers or vary in the level of cell marker expression which may affect differentiation potential of the MSC. For example, cells that have a high expression of CD105 have an increased osteogenic potential (Maleki et al, 2014). Bone marrow derived MSCs (BM-MSCs) are rare, only comprising around 1 in 10,000 cells in bone marrow whereas adipose tissue can yield 500 times this amount (Williams et al, 2013; Kolaparthy et al, 2015). Despite their rarity, BM-MSCs are relatively easy to isolate and are commonly used in stem cell research. BM-MSCs have been shown to have greater chondrogenic and osteogenic potential compared to adipose derived MSCs (AD-MSCs; Li et al, 2015). Therefore, BM-MSCs are the better choice for studies into cartilage and bone.

As well as the cell types already mentioned, MSCs have the potential to differentiate into other lineages including tenocytes (Wang et al, 2005), myocytes (Singh et al, 2016) and neural cells (Scuteri et al, 2011). They have been studied extensively for the purpose of tissue engineering and regeneration due to their vast differentiation potential. For example, studies investigating the use of MSCs for tendon repair have been promising. Tendon is a connective tissue that connects muscle to bone and consists of tightly packed collagen fibrils in parallel to each other. Tendon has an inherently low regeneration capacity and healed injured tendon leaves tough, fibrous scar tissue that affects the physical properties and function of the tendon. The impaired function of the tendon increases the risk of further injury (Butler et al, 2004). Incorporation of MSCs into injured tendon resulted in improved repair in a rabbit model of Achilles tendon regeneration (Young et al, 1998). Other studies have also shown use of MSCs increased tendon repair in mice (Hoffman et al, 2006) and horses (Carvalho et al, 2013).

MSCs have been used for regeneration of bone. In adults, the skeleton is slowly and continuously remodelled; a process mediated by osteoblasts and osteoclasts. These are bone forming and bone resorbing cells originating from MSCs and HSCs respectively. Comparative to other tissues, in healthy individuals bone heals itself adequately in most cases (although bone repair is a slow process). However, bone regeneration is impaired with age and in diseases such as osteogenesis imperfecta (Gil et al, 2017) and osteoporosis (Tarantino et al, 2011). MSC grafts to the site of damaged bone increased repair in rat calvarial bone (Agacayak et al, 2012).

MSCs are also extensively used in cartilage repair studies. Like tendon, cartilage is a connective tissue and is dense in ECM (Section 1.1). Similar to tendon, cartilage possesses very limited regeneration capabilities in adults. Adult cartilage forms a fibrocartilage scar upon injury healing. Accordingly, cartilage repair faces many similar challenges to tendon repair. Clinical trials have shown that application of autologous MSCs to injured cartilage improved repair (Wang et al, 2017). A review of multiple animal and human studies of MSC use in cartilage repair identified different outcomes depending on the methods and conditions used and a lack of standardisation in measuring outcomes (Goldberg et al, 2017).

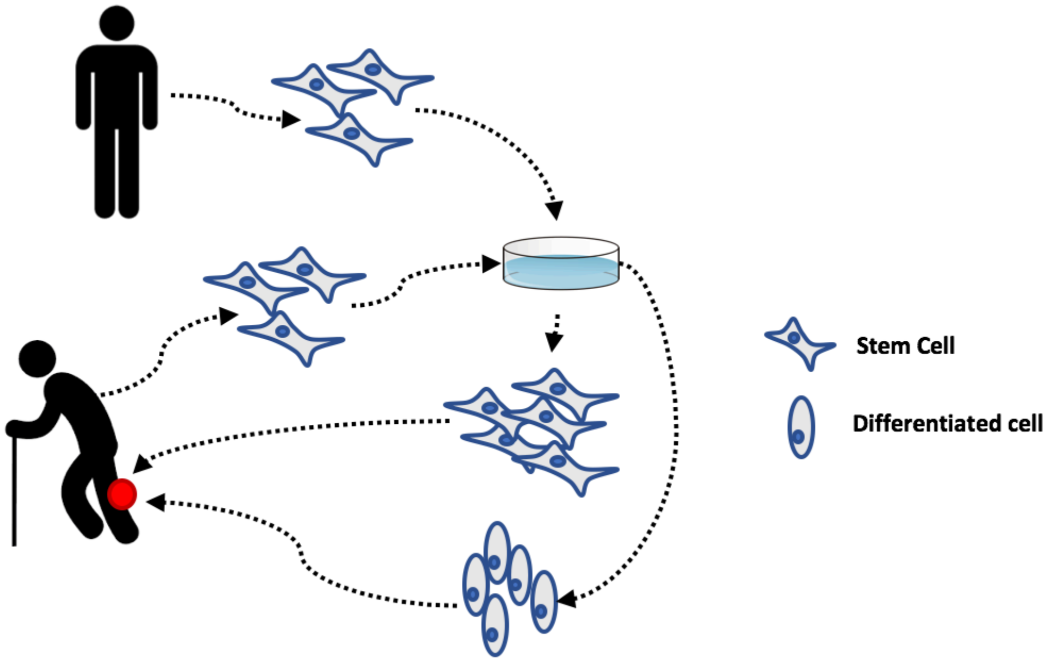


Figure 1.5 – Basic principle of stem cell transplantation. Stem cells can be extracted from the patient (autologous stem cell transplant) or a healthy donor, they can then be expanded or differentiated in vitro and transplanted to the site of injury or disease.

As well as directly using MSCs for tissue regeneration, MSCs can also be differentiated into the desired cell type to create artificial tissues *in vitro*. Tissue engineering approaches revolve around isolating MSCs from donors or patients and differentiating them under laboratory conditions before transplanting the differentiated cells back into the patient (Fig. 1.5). Differentiating isolated MSCs has advantages to simply implanting MSCs to the site of injury or disease. The injured or diseased tissue may not have conditions ideal for MSC differentiation into the required cell type, or MSCs may not stay at the site of injury (Haque et al, 2015). In this case, differentiating stem cells *in vitro* before transplanting may be the more effective method. Creating cells and tissues from isolated MSCs is a complex task. Depending on the tissue, it may be advantageous to grow cells in 3D scaffolds rather than monolayer cell cultures. Cells such as chondrocytes, tenocytes and osteoblasts exist *in vivo* surrounded by networks of extracellular matrix. They are sensitive to external stimuli such as weight loading which affects the integrity of cartilage, tendon and bone. Both over and under loading can negatively impact the structure and strength of the tissue (Sun, 2010; Galloway et al, 2013; Klein-Nulend et al, 2012). 3D scaffolds offer cells a more natural environment to grow compared to monolayer dishes; MSCs grown in 3D scaffolds show increased proliferation compared to those grown in monolayer cultures (Meng et al, 2014). Materials for stem cell scaffolds, particularly for bone and cartilage regeneration, have been widely researched. Scaffolds usually contain various growth factors and other proteins necessary for optimal attachment, proliferation and differentiation of MSCs. Collagen containing scaffolds have been shown to increase MSC proliferation (El-Jawhari et al, 2015). Scaffold-free methods have also been developed and these are reviewed in section 1.5.

As well as finding suitable materials for cell scaffolds, there has been research into engineering cultured MSCs to be more robust or more capable of differentiating into the desired cell type (Park et al, 2015). Stem cells proliferate at different rates and they may be more likely to differentiate into some cell lineages than others. Age is a major factor that negatively affects the rate of stem cell division and differentiation (Sharpless and DePinho, 2007). In healthy adults, MSCs divide and differentiate in both in response to tissue injury and as part of normal tissue homeostasis. With age, proliferation of MSCs decreases and MSCs tend to differentiate into adipocytes at the expense of osteoblastogenesis (Bethel et al, 2013). Studies into age related effects of MSC chondrogenesis have yielded conflicting outcomes with some showing that chondrogenesis is impaired with age (Kretlow et al, 2008) whereas others have shown no difference (Scharstuhl et al, 2007; Payne et al, 2010). Increasing the shelf life of extracted MSCs is also important. MSCs that have been cultured and passaged 7-12 times display reduced differentiation capacity and enter replicative senescence (Wagner et al, 2008). Similarly, MSCs extracted from aged mice show reduced function compared to young and adult mice (Bruna et al, 2016). Age was also found to negatively impact extracted human MSCs; genes related to oxidative stress were upregulated in aged MSCs (Peffers et al, 2016). Therefore, the age of stem cell donors is an important factor to consider when using MSCs in regenerative medicine and tissue engineering.

Regulation of gene expression in stem cells both *in vivo* and *in vitro* is mediated by multiple genetic, epigenetic and environmental factors. Understanding these mechanisms is key to developing new regenerative medicine solutions for injury and disease. The use of MSCs for tissue repair has been extensively studied and results have been positive. However, application of this technology to treat human patients is not widespread and is limited to clinical trials or research settings. Optimisation of such techniques is far from complete and there are still many questions to be answered before the use of stem cells to treat human injury and disease becomes routine. Further research is required to perfect existing techniques and develop new methods. Regenerative medicine using stem cells is a powerful tool to treat a wide range of diseases; in theory, stem cells can be used to regenerate any somatic cell type in the body. This is particularly paramount for cells and tissues that are not easily replaced or repaired by the body.

1.5 Scaffold-free *in vitro* models of chondrogenesis

To study chondrogenesis, many *in vitro* models have been developed (Yu et al, 2012). Chondrogenesis can be studied in animal models but it is not feasible to investigate human chondrogenesis *in vivo*. Therefore, *in vitro* models offer a valuable method of studying chondrogenesis of human cells. They are also a step towards developing usable cartilage for transplantation.

hMSCs can be extracted from various tissue sources such as adipose tissue and bone marrow. BM-MSCs are superior to AD-MSCs for chondrogenesis (Li et al, 2015). Once extracted, BM-MSCs proliferate well in culture and can be stimulated to differentiate into its cell lineages by the addition of growth factors into the cell media (Fig. 1.6). MSCs can also be maintained in the stem cell state by culturing in media containing fibroblast growth factor 2 (FGF2), ascorbic acid and platelet-derived growth factor (PDGF; Gharibi and Hughes, 2012). FGF2 promotes the proliferation of MSCs while inhibiting differentiation (Lai et al, 2011). Cultured MSCs can be induced to differentiate by replacing MSC maintaining media with differentiation media. Differentiation media is designed to simulate the changes in the environment that occur *in vivo* to drive MSCs down a particular cell lineage. For example, supplementation with BMP2 induces osteoblastogenesis (Westhrin et al, 2015) whereas indomethacin and 3-isobutyl-1-methylxanthine (IBMX) induces adipogenesis (Scott et al, 2011). For chondrogenesis to occur in cultured MSCs, TGFβ supplementation is essential (Worster et al, 2000; Bian et al, 2011). Addition of other supplements such as dexamethasone, ascorbic acid and insulin aids the *in vitro* differentiation process (Solchaga et al, 2011).

Successful differentiation of MSCs into the desired cell type is usually measured by the upregulation of cell type specific genes and the downregulation of genes involved in other cell lineages, including MSC markers. Phenotypic and cell morphology changes can also be measured using other methods such as cell staining and flow cytometry.

Osteoblastogenesis results in the upregulation of osteoblast markers such as *RUNX2*, *BGLAP* and *COL1A1* (Huang et al, 2007). Positive adipocyte markers include the *ADIPOQ* and *LEP* genes (Houde et al, 2014). MSC differentiation into chondrocytes is marked by induction of a chondrogenesis specific gene set such as *SOX9*, *COL2A1* and *ACAN* (Solchaga et al, 2011) and downregulation of osteoblast genes such as *RUNX2* before the hypertrophic stage (Lengner et al, 2005). Quantification of such markers allows for the assessment of differentiation.

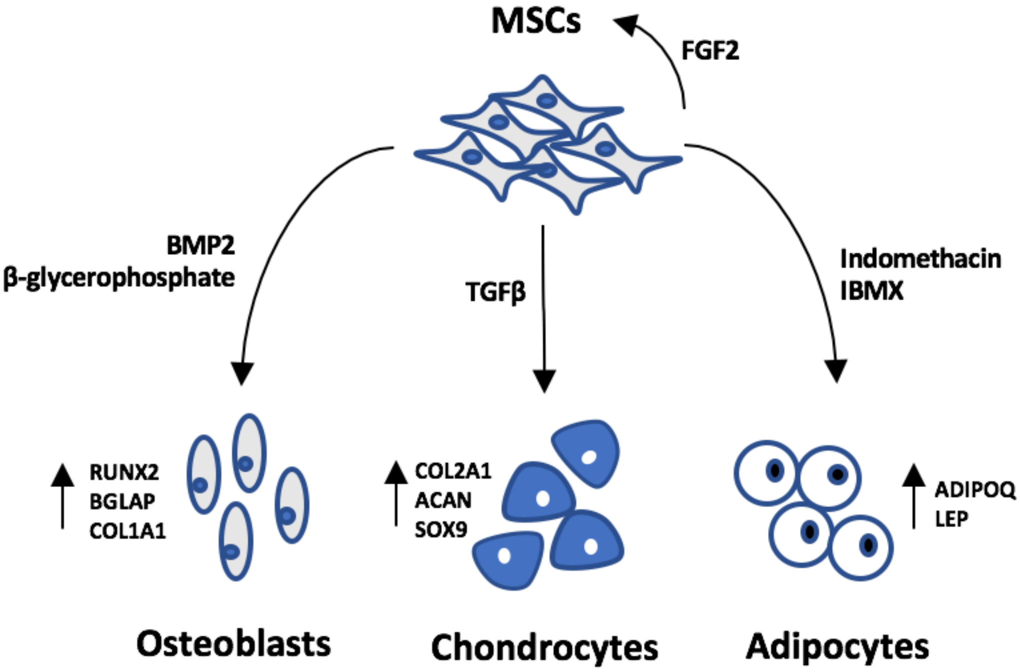


Figure 1.6 – In vitro MSC differentiation into osteoblasts, adipocytes and chondrocytes. Cultured MSCs can be stimulated to differentiate into osteoblasts by adding growth factors such as BMP2 and β-glycerophosphate into the cell media. Gene markers of osteoblasts include RUNX2, BGLAP and COL1A1. Addition of indomethacin and IBMX in cell media promotes adipogenesis; characterised by expression of ADIPOQ and LEP. For chondrogenesis to occur, MSCs must be grown in the presence of TGFβ. Chondrocyte markers include an upregulation of COL2A1, ACAN and SOX9

In vitro models of chondrogenesis vary and there are methods that utilise scaffolds and scaffold-free methods. Scaffold-free methods have advantages over scaffold based techniques as they are relatively simple to execute and do not rely on artificial scaffold material of varying quality. Furthermore, scaffold-free chondrogenesis mimics more closely the process that occurs during native development; the chondrocytes are seeded at high density which resembles the condensation step and the cells themselves create their own scaffold through expression and secretion of ECM proteins (Whitney et al, 2012).

The simplest scaffold-free chondrogenesis model involves differentiation of high density MSCs in pellet culture. During incubation of MSCs in chondrogenic cell media, cells differentiate into chondrocytes and form a pellet of ECM resembling cartilage. The micromass model is an adaptation of the pellet culture model whereby MSCs are differentiated into chondrocytes within the wells of 24 48 or 96-well plates without centrifugation of MSCs. The two chondrogenesis models are similar although the micromass model was found to be more cartilage-like compared to the pellet model (Zhang et al, 2010). The transwell model of chondrogenesis is another scaffold-free model of chondrogenesis (Murdoch et al, 2007). This model involves seeding MSCs at high density into tissue culture treated transwell inserts containing a porous membrane. Chondrogenic differentiation media is added into and around the transwell. Over time, a cartilaginous disc forms at the base of the transwell insert. The ECM within the transwell chondrogenesis model was found to be more homogenous than pellet cultures (Murdoch et al, 2007). Previous studies using the transwell chondrogenesis system has yielded gene expression changes similar to *in vivo* chondrogenesis (Murdoch et al, 2007; Barter et al, 2015).

The stages of chondrogenesis can be investigated by stopping *in vitro* differentiation at different time points. The challenges involved in differentiating MSCs into chondrocytes *in vitro* include avoiding the terminal hypertrophic fate of chondrocytes. In development, the majority of chondrocytes apoptose after hypertrophy during endochondral ossification. Articular cartilage is formed from chondrocytes that do not undergo this stage. Whilst cultured MSCs can produce a cartilage-like tissue with similar gene expression profiles, a model that can result in cartilage exactly like native cartilage has yet to be developed. Articular cartilage in the body is characterised with zones at different depths in the cartilage. At each zone, the ECM composition and chondrocyte morphology varies. *In vitro* models struggle to replicate the complex layers found in endogenous cartilage. Furthermore, MSCs extracted from donors are heterogeneous and can exhibit varied proliferation and differentiation potential (Somoza et al, 2014). Selecting MSCs for *in vitro* experiments can therefore be biased towards those that grow well in culture and can successfully differentiate in the model used by the researcher.

It is important to develop better cartilage constructs with the properties of endogenous articular cartilage for tissue regeneration purposes. It has been proposed that MSC chondrogenesis may be enhanced by using epigenetic modulators (Patel et al, 2015; Yapp et al, 2016). Therefore, improvement of cartilage constructs may be accomplished by engineering the cells themselves. Manipulating MSC differentiation can be achieved with the use of epigenetic modifying enzymes or using the CRISPR-cas9 system. However, more knowledge of how the differentiation process is regulated is required.

1.6 Epigenetics

1.6.1 Definition of epigenetics

Epigenetics (from the Greek prefix “epi” meaning above or over plus the word genetics) describes the changes and alterations around the genome that have an effect on transcriptional activity without changing the underlying nucleotide sequence of the genome itself. Epigenetic modifications can be hereditary and some definitions state that an epigenetic trait must be heritable, either to daughter cells or progeny (Berger et al, 2009). However, some traits which are non-heritable may also be classed as epigenetic (Bird, 2007). In this project, epigenetics is taken to mean any heritable or non-heritable change which does not change the genome sequence.

Definitions of epigenetics do not distinguish between mitotic and meiotic inheritance. Transgenerational epigenetic mechanisms are important for controlling normal cell processes such as development whereas many epigenetic marks can also be influenced by environmental factors and are not passed on to the next generation (Slatkin, 2009). This is an important distinction as acquired epigenetics can contribute to the development of disease.

Epigenetic changes may regulate gene transcription and are therefore a powerful mechanism for modulating gene expression in disease. Epigenetic marks regulate gene transcription partly by altering the structure of chromatin into transcriptionally permissible or repressed states. Epigenetic marks such as histone modifications can also attract specific transcription factors to regulate gene expression. The two major epigenetic mechanisms regulating gene transcription are histone modifications and DNA methylation (Fig. 1.7). Non-coding RNAs (ncRNAs) are also classed an as epigenetic mechanism (Peschansky and Wahlestadt, 2011), as are prions (Halfmann and Lindquist, 2010) but these are beyond the scope of this project and are not discussed further.

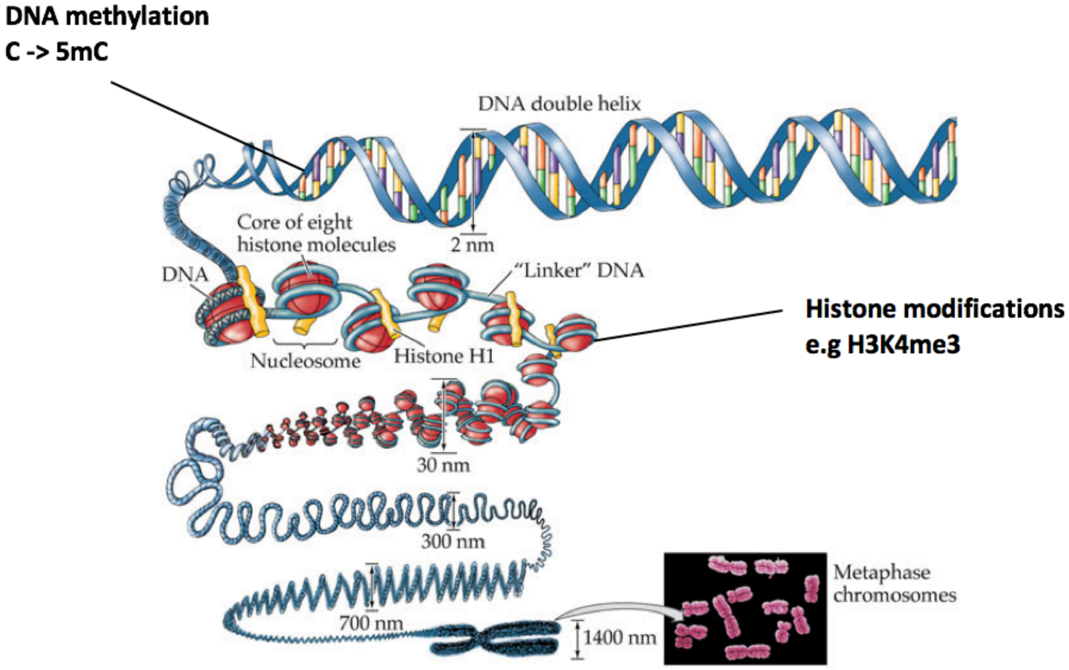


Figure 1.7 - Packaging of DNA into chromosomes in the nucleus. DNA wraps around histone proteins in a nucleosome unit to form chromatin. Chromatin is further condensed and packaged into a chromosome. Chromatin structure is determined by epigenetic marks such as DNA methylation and histone modifications. Histone tails may be modified by the addition of a chemical group, e.g. methylation. DNA methylation occurs on the cytosine base of CpG sites. These two epigenetic components can alter the structure of chromatin. (Image adapted from Sadava et al, 2012).

1.6.2 Histone modifications

1.6.2.1 Chromatin structure

DNA in the nucleus is compacted into chromatin by histones; chromatin is usually grouped into transcriptionally repressed heterochromatin or transcriptionally active euchromatin. Constitutive heterochromatin is found at centromeres and telomeres and consists of repeating sequences of DNA that are not transcribed. Facultative heterochromatin is usually transcriptionally silenced although may become active depending on signalling cues (Trojer and Reinberg, 2007). Transcriptionally active genes are typically found in euchromatin, which is not as tightly compacted as heterochromatin. The open structure of euchromatin is usually described as resembling beads on a string, with the string representing DNA and the histone proteins represented by beads. Not all genes in euchromatin are constitutively expressed and regions with inactive genes are closed to restrict accessibility to the transcription machinery. Regions of transcriptionally inactive euchromatin is condensed into facultative heterochromatin, with the process reversed upon activation of gene transcription. Repression of transcription exists on a continuum and genes may be partially or fully repressed. Studies in yeast show that the complete silencing of genes is a gradual process and occurs over multiple cell divisions (Katan-Khaykovich and Struhl, 2005). Furthermore, the spatial organisation of chromatin can affect gene transcription. Genes in euchromatin that are in close proximity to heterochromatin may also be silenced, a phenomenon termed position effect variegation (PEV). PEV was originally identified in *Drosophila* although it has also been demonstrated in mammals (Kleinjan and Heyningan, 1998). This demonstrates how the structure of chromatin can influence whether a gene is transcribed.

Post-transcriptional modifications to histone proteins provide a method of epigenetically altering gene expression. Histones are highly conserved proteins that DNA wrap around in a nucleosome, subsequently compacting it into chromatin. There are four core histones, H2A, H2B, H3 and H4. Each nucleosomes unit consists of 146bp DNA wrapped around two of each of the four core histones forming an octamer. Nucleosomes are joined by two linker histones H1 and H5.

1.6.2.2 Chromatin remodelling and histone modifications

The amino acids of core histone N-terminal tails can be chemically modified in a number of ways, the most common being methylation and acetylation. These marks have the potential to change the structure of chromatin and accessibility of the DNA to the transcriptional machinery. Histone marks regulate transcription by recruiting chromatin modelling complexes change the configuration of chromatin to allow or block the accessibility of the transcriptional machinery to genes. Specific histone modifications may be represented in text in short form. The nomenclature of histone modifications notation begins with the histone number followed by the modified residue and type and number of chemical groups added. For example, H3K4me3 conveys that histone H3 is modified on the fourth lysine residue with three methyl groups.

Besides methylation and acetylation, other histone modifications include ubiquitination, sumoylation and phosphorylation. This short review focuses on methylation and acetylation as other types of modifications are beyond the scope of this project. Histone marks that influence gene transcription are found throughout the genome including at regulatory elements and within gene bodies (Hon et al, 2009). Methylation of histones may lead to transcriptional repression or activation of genes depending on the specific mark. For example, H3K4me3 is associated with gene transcription whereas H3K27me3 is usually linked to transcriptional repression. Methylation of histones involves the covalent addition of a methyl group ($-CH_3$) to lysine or arginine residues on histone tails. S-Adenosylmethionine is the main methyl donor in many organisms including humans (Detich et al, 2003). Lysines may be mono-, di- or tri-methylated. Arginine residues may be mono- or di-methylated (Bannister and Kouzaridas, 2011). These modifications are mediated by either lysine or arginine methyltransferases. It was originally thought that histone methylation was an irreversible process. However, histone demethylation enzymes have been identified, the first of which was KDM1A (Shi et al, 2014). Histone demethylases act in complexes in order to demethylate nucleosomal histones.

Histone acetylation is associated with transcriptional activation (Eberharter and Becker, 2002). Acetylation of histones is mediated by histone acetyltransferases (HATs) and removed by histone de-acetylases (HDACs). HATs catalyse the addition of an acetyl group to lysines on histone tails, although sites within globular histone core can also be acetylated, such as H3K56 (Yuan et al, 2009). In contrast to methylation, lysines can only be acetylated once. Acetylation of histones leads to remodelling of chromatin into an open conformation. Acetylation is also proposed to neutralise the positive charge on the lysine, which weakens the bond between the histone and DNA. This opens the chromatin to transcription factors that can then bind to DNA to initiate transcription. HDACs remove acetyl groups from lysine, restoring the charge difference between histone and DNA (Bannister and Kouzarides, 2011).

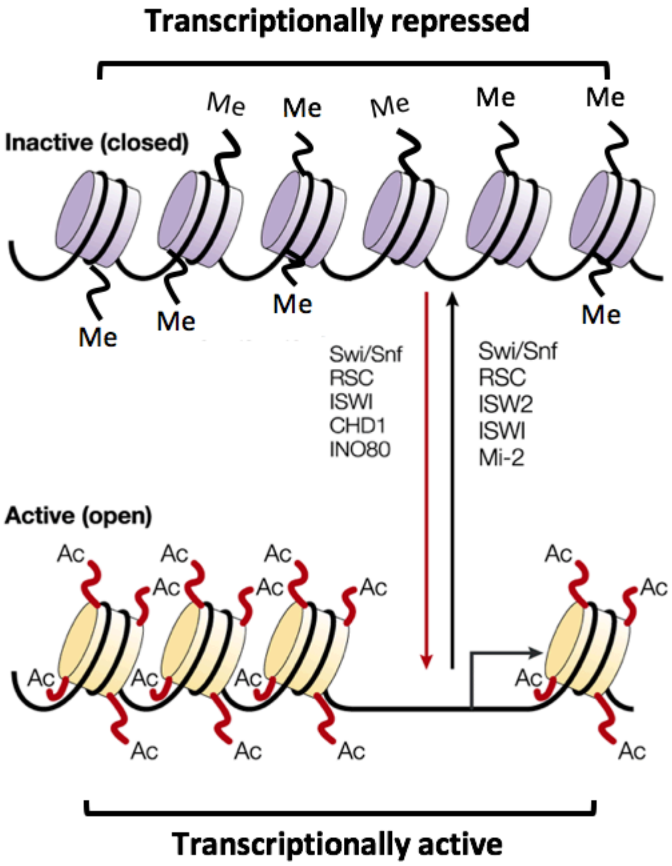


Figure 1.8 – Chromatin remodelling by ATP-dependent chromatin remodelers mediated by histone modifications. Specific histone modifications attract remodelers to alter the structure of chromatin. Some chromatin remodelers e.g. Swi/Snf are involved in both opening and closing chromatin whereas others only remodel in one direction. Image adapted from Tsukiyama, 2002.

Chromatin remodelling enzymes reposition histone proteins, changing the conformation of chromatin. A number of chromatin remodelling complexes have been identified (Fig. 1.8) but a comprehensive review is beyond the scope of this project. Chromatin remodellers interact with histone modifying enzymes to couple histone modifications and chromatin remodelling (Felisbino et al, 2013). Repressive histone marks are known to silence gene expression by recruiting polycomb group protein containing complexes (PRCs). Histone marks H3K9me3 and H3K27me3 are both found in heterochromatin. H3K9me3 marks are abundant throughout constitutive heterochromatin and H3K27me3 is common in facultative heterochromatin. The PRC1 complex is able to recognise and dock at H3K27me3 sites and subsequently ubiquitinate H2A, which in turn attracts PRC2. PRC2 catalyses the transfer of methyl groups to H3K27, creating a positive feedback loop. These two complexes work together to remodel chromatin into higher order structures to silence transcription (Blackledge et al, 2014).

1.6.2.3 Regulatory elements and the histone code

The histone code is the hypothesis that histone modification patterns can determine whether a gene is switched on or off. It is proposed that gene expression can be predicted by the histone modifications surrounding a gene (Jenuwein, 2001). Interpreting this code could help elucidate how transcription is regulated via histone modifications. The histone code hypothesis is more complex than originally thought. Histone modifications can indeed determine whether genes are transcribed. However, combinations of histone modifications often exist and different combinations have different meanings in the context of gene regulation.

Distinct histone modifications mark cis-regulatory regions in the genome such as gene promoters and enhancers. Different histone modifications may be site specific or found genome wide. The core or minimal promoter of genes generally consists of its transcriptional start site (TSS), binding sites for RNA polymerases and general transcription factors and a 5'-TATAAA-3' motif known as a TATA-box. The proximal promoter is the region outside of the core promoter that contains further transcription factor binding sites. Promoters are located upstream of genes but can also overlap the TSS and 5’UTR at the start of genes. The size of promoters varies but can range from 100bp to over 1kb (Kanhere and Bansal, 2005). Although histone methylation generally silences transcription of genes, some methylation marks can also be active. H3K4me3 is associated with the promoters of transcriptionally active genes. Regulatory elements can also contain bivalent domains, where both active and repressive marks are present. Bivalency is common in stem cell promoters, which often become monovalent after differentiation (Mikkelsen et al, 2007). H3K4me3 and H3K27me3 co-localisation are a bivalent state often seen in stem cell promoters. It is thought that H3K27me3 represses activation of lineage specific genes whereas H3K4me3 readies the gene for activation upon differentiation (Bernstein et al, 2006), however, others have shown that bivalently marked genes may be expressed at low levels (De Gobbi et al, 2011). Upon differentiation, promoters may lose either H3K27me3 or H3K4me3 marks, or remain bivalent (Cui et al, 2009). Bivalent modifications can also be established upon entering a differentiation lineage (De Gobbi et al, 2011). Genome wide histone marks H3K9ac and H3K14ac are usually associated with active and bivalent promoters alongside

H3K4me3, and show a strong correlation with the CpG content of promoters (Karmodiya et al, 2012).

H3K27me3 marks are associated with transcriptional repression and can be found throughout the genome. However, in rare cases H3K27me3 may also correlate with active transcription. H3K27me3 marks were enriched in the promoter of genes that were transcriptionally active in mouse embryonic and progenitor cells (Young et al, 2011). This adds an extra layer of complexity to the histone code hypothesis, as there is some degree of plasticity to whether a histone mark is repressive or active depending on the gene and/or cell type. H3K9me3 is another example of a transcriptionally repressive mark. Whilst typically found in constitutive heterochromatin in contrast to H3K27me3 which marks facultative heterochromatin, H3K9me3 also plays a role in cell type specific gene regulation (Becker et al, 2016).

Gene enhancers are distal regulatory elements that can regulate their target gene over long distances and do not necessarily act on its nearest gene. Enhancers can be found 1Mbp away from the gene it regulates (Cho, 2014). Sizes of enhancers show more variation than promoters and can range from 50bp to many thousands of bases. Enhancers are typically found upstream of genes but may also be downstream or within gene bodies. Active enhancers are proposed to interact with their target promoters through a chromatin looping mechanism (Fig. 1.9). The enhancer-promoter interaction initiates gene transcription by attracting the components of the transcription machinery such as RNA polymerase II. Like promoters, enhancers also contain binding sites for transcription factors and co-activators. H3K27ac is a marker of active promoters and active enhancers. H3K4me1 is associated with poised enhancers when alone and active enhancers alongside H3K27ac (Creyghton et al, 2010). As well as H3K27ac, the Mediator 1 coactivator protein (MED1) is often used as a marker for active enhancers. MED1 is a subunit of the Mediator transcriptional co-activator complex. The p300 co-activator is an active enhancer marker and can also be used to define super enhancers (Pott and Lieb, 2015). Enhancers tend to be cell type specific and regulate genes involved in cell type specific processes rather than genes involved in general maintenance of the cell (Ong and Corces, 2011).

Individual enhancers may regulate multiple genes and the same enhancer may regulate different genes in different cell types. There are far more enhancers present in the genome compared to genes. Redundancy of enhancers is proposed to aid in fine tuning the regulation of gene transcription (Guerrero et al, 2010).

A subgroup of enhancers called super enhancers have been identified. These are defined as multiple neighbouring enhancers that exhibit high levels of enhancer markers such as H3K27ac and MED1. Super enhancers are reported to increase the upregulation of genes compared to typical enhancers. Stretch enhancers, another subgroup of enhancers, are defined as long enhancers associated with cell type identity. Stretch and super enhancers share many similarities and may be largely synonymous (Pott and Lieb, 2015).

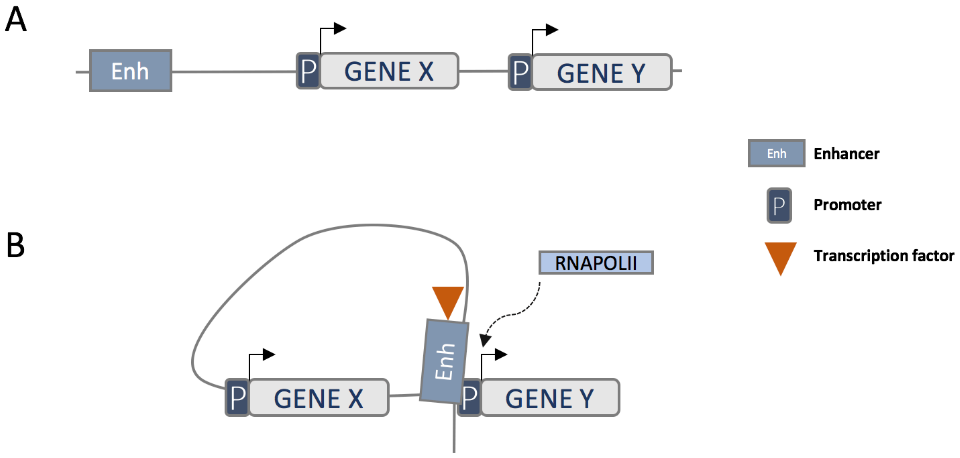


Figure 1.9 – Schematic of regulatory elements in the genome. (A) General layout of cis-regulatory elements surrounding genes; enhancers are typically located distal to the gene TSS whereas promoters are proximal to the gene TSS. (B) Enhancer-promoter looping model. Enhancers may be bound by transcription factors that facilitates enhancer binding to a target promoter to activate gene transcription by recruiting RNA polymerase II and transcription initiation factors. Enhancers do not necessarily target the nearest gene and can act on genes over 1Mbp away.

Histone modifications can occur both at regulatory elements of genes and within gene bodies. H3K36me3 is a transcriptionally permissive mark found in actively transcribed genes. The histone methyltransferase SET2 catalyses the addition of methyl groups H3K36 within genes and interacts with RNAPII, leading to the coupling of H3K36me3 marks and transcription elongation (Kizer et al, 2005). SET2 physically binds to gene bodies but not gene promoters (Schaft et al, 2003). This demonstrates that gene transcription is not only regulated at promoters and enhancers but relies on a co-ordinated effort across the gene.

Histone methylation and demethylation is proposed to affect diseases such as cancer and also in ageing and cell senescence by switching on and off relevant genes. This makes histone methyltransferases and demethylases attractive drug targets (Cloos et al, 2008). HDAC inhibitors (HDACi) are inhibitors of HDAC and prevent HDAC activity by binding to the catalytic domain of HDACs. Targeting of HDACs with HDACi has been used as a therapeutic for diseases such as cancer. Cancers caused by abnormal HDAC recruitment to promoters of oncogenes can potentially be treated with HDACi. Furthermore, HDACs are up-regulated in a number of cancers. Treatment with HDACi results in cell cycle arrest and apoptosis (Marks et al, 2000). Use of HDACi in cancer has been promising and a number of drugs have been licenced for this purpose. For example, vorinostat is a HDACi used to treat cutaneous T cell lymphomas and has also been used in HIV research (Archin et al, 2009). The use of HDACi’s has also been explored as therapeutics for OA. Recent studies have shown that treatment with HDACi can have a chondroprotective effect by repressing MMP expression induced by cytokines (Young et al, 2005; Culley et al, 2013). Despite encouraging outcomes so far, there are still some issues with specificity and cardiotoxicity (Gryder et al, 2012) associated with the therapeutic use of HDACi. The use of HDACi to modify gene expression in disease illustrates how important histone modifications are in regulating gene expression. HDACi’s may also be used to regulate chondrogenesis *in vitro*. An HDACi, trichostatin A, was shown to be able to inhibit hMSC chondrogenesis through suppression of the Sp1 transcription factor (Wang et al, 2011). These studies show that the epigenome can be modified to alter gene transcription.

As well as using enzymes to modulate epigenetic marks, the use of genome editing methods to alter epigenetic marks has been investigated. The versatile CRISPR-Cas9 technique can be used to engineer non-coding as well as coding regions of the genome. Epigenome editing can be achieved by targeting activator or repressor proteins to promoters or enhancers of genes (Hilton et al, 2015). Functional roles of enhancers have been elucidated by targeting Cas9 to transcription factor binding sites within enhancers (Korkmaz et al, 2016).

Groupings of histone modifications can define regulatory elements and regulate genes through modulating chromatin remodelling to allow or block access of transcription factors. However, histone modifications also rely on other epigenetic mechanisms such as DNA methylation. Histone modifications rely on DNA methylation and vice versa (Cedar and Bergman, 2009). Crosstalk between the two epigenetic mechanisms allows for greater control of gene transcription and it is important to consider histone modifications in the wider context of the whole epigenome.

1.6.3 DNA methylation

DNA methylation is an important epigenetic mark influencing gene transcription. Methylation of DNA is generally a repressive mark although there are exceptions (Halpern et al, 2014). DNA methylation is important in mammalian development in controlling the expression of imprinted genes and aiding in X-chromosome inactivation. Imprinted genes are genes that are expressed depending on their parental origin. Autosomal genes in diploid organisms possess two copies or alleles inherited from both parents and gene expression usually occurs equally from both. However, a small number of imprinted genes are only expressed from just one allele. For example, the IGF2 gene is only expressed from the paternally inherited allele and the maternal allele is silenced (Giannoukakis et al, 1993). Genomic imprinting also affects tissue specific genes (Gregg et al, 2010). During early embryonic development, the majority of DNA methylation is erased and re-established with imprinted genes programmed to only express one allele, in part due to DNA methylation of the silenced allele. In females, one X-chromosome is randomly inactivated so that genes on the X-chromosome have the same level of gene expression as the single X-chromosome present in males (Sharp et al, 2011). The *XIST* non-coding RNA gene initiates X-chromosome inactivation which is maintained by DNA methylation (Plath et al, 2002). More generally, DNA methylation acts alongside other epigenetic mechanisms to control gene transcription.

DNA may be methylated at CpG sites, typically clustered together to form CpG islands near gene promoters (Illingworth and Bird, 2009). Specifically, the cytosine base within CpG sites are methylated to form 5-methylcytosine (5mC). More rarely than cytosine, adenine bases can also be methylated (Wu et al, 2016). DNA methylation is mediated by DNA methyltransferase (DNMTs) enzymes which catalyse the addition of a methyl group to the cytosine base. In mammals, three active DNMTs have been found - DNMT1, DNMT3A and DNMT3B. DNMT3A and DNMT3B can target promoters as a complex along with HDACs and other transcriptional repressors, illustrating the crosstalk between histone modifications and DNA methylation (Arzenani et al, 2011; Smith and Meissner, 2013).

At CpG islands in promoters of actively transcribed genes, hypomethylation is achieved by preventing DNA methyltransferases from accessing the CpG sites. Housekeeping genes possess promoters that are constitutively hypomethylated, reflecting the continual transcription of these genes. Transcription factor binding at promoters help block DNA methyltransferases from methylating and thus repressing gene transcription. The Sp1 transcription factor was found to prevent methylation of CpG islands. Removal of transcription factor binding sites in gene promoters lead to higher methylation levels at CpG sites in the promoter, suggesting that DNA methylases and transcription factors compete for binding (Brandeis et al, 1994; Macleod et al, 1994). DNA methylation also plays a role in alternative splicing by inhibiting CTCF binding at exons; CTCF promotes the inclusion of exons by stalling RNA polymerase II (Shukla et al, 2011).

Methyl groups can be removed from DNA by demethylase enzymes in a multistep process (Pfeifer et al, 2013). 5mC can become oxidised to form 5-hydroxycytosine (5hmC), a process mediated by the TET family of proteins. 5hmC is an intermediate modification formed during the demethylation of 5mC. The existence of 5hmC was first discovered in bacteria and was confirmed to be present in mammalian cells in 2009 (Kriaucionis and Heintz, 2009; Tahiliani et al, 2009).

Interestingly, promoters of housekeeping genes and some development genes remain hypomethylated even when transcriptionally silent (Weber et al, 2007). This suggests other regulatory mechanisms are also involved in modulating gene expression. Although early studies on DNA methylation have focused on gene promoters, gene enhancers may also be methylated. In general, promoters show a lower level of DNA methylation compared to enhancers. DNA methylation at enhancers is associated with regulation of genes involved in cell specific processes. Hypomethylation of gene enhancers and super enhancers is correlated with increased cell specific gene expression in muscle cells (Ehrlich et al, 2016). It was shown in cancer cells that hypomethylation of enhancers more closely correlated with gene expression than hypomethylation of promoters (Aran et al, 2013). Therefore, associating DNA methylation at promoters to gene transcription without considering methylation at other regulatory elements may be overly simplistic.

1.7 Bioinformatics

1.7.1 High throughput sequencing

High throughput sequencing, also called next-generation sequencing (NGS; although this term has fallen out of use), is one of the methods commonly used in ‘omics studies. High throughput sequencing encompasses the newer, high throughput methods of DNA sequencing that have been developed after the original Sanger sequencing method. Sanger sequencing was developed in 1977 by Frederick Sanger and for many years it was the sole method of sequencing DNA. The dye terminator Sanger sequencing method determines the sequence of a DNA template using DNA polymerase, a DNA primer, deoxynucleosidetriphosphates (dNTPs) and di-deoxynucleotidetriphosphates (ddNTPs). Briefly, the primer binds to the ssDNA template and the DNA polymerase extends the primer using dNTPs. In current Sanger sequencing, the sequence is determined and reaction terminated when a fluorescently labelled ddNTP binds to the elongating strand and emits light; each of the four ddNTPs are labelled with fluorescent dyes that emit light at different wavelengths. Owing to the reaction being terminated by the incorporation of a ddNTP, a large amount of starting DNA template is required to fully determine the DNA sequence as many reactions are needed. This is one of the limitations of Sanger sequencing; it is not possible to sequence rare DNA samples using this method. Furthermore, it can be a slow process requiring many reactions to determine the full sequence of a length of DNA. Although Sanger sequencing for genomics applications has been largely replaced by newer methods, it is still widely used for small scale DNA sequencing such as determining the sequence of plasmids. It is also used when longer reads are required; one of the main advantages of Sanger sequencing is that it can generate reads of up to 1000bp. In contrast, many high throughput methods typically only manage 50-150bp reads (Hodkinson and Grice, 2015).

High throughput sequencing methods can sequence many DNA templates simultaneously multiple times. Massively parallel sequencing can generate millions of reads in a comparatively short time. This allows for deep sequencing of DNA where bases are sequenced multiple times, reducing the potential for sequencing errors. High throughput sequencing generates large amounts of data which presents challenges for storage and analysis.

There are a selection of high throughput sequencing methods developed by different life science companies including SOLiD sequencing, pyrosequencing, Ion Torrent sequencing and Illumina sequencing. Each has its own advantages and disadvantages and have enjoyed varying popularity over the years. As of 2015, Illumina's sequencing by synthesis platforms dominate the high throughput DNA sequencing market and their success is predicted to continue (Goodwin et al, 2016). Illumina's sequencing by synthesis method is an adaptation of dye terminator Sanger sequencing using non-terminating ddNTPs. The cost of high throughput sequencing has dramatically fallen in recent years, making it more accessible to researchers. Accordingly, increasingly sophisticated bioinformatics tools have been developed to cope with the large influx of sequencing datasets and the various needs of different research groups. In general, more bioinformatics resources are required to store and analyse high throughput sequencing data compared to array data.

From high throughput DNA sequencing, a number of derivative methods have been developed to assay various biological molecules and features. Originally, sequencing methods were used to determine the sequence of genomic DNA. Nowadays, there are adapted methods to assay the transcriptome, exome and epigenome.

1.7.2 ‘Omics methods

1.7.2.1 Transcriptome

The transcriptome is the full repertoire of RNA transcripts transcribed by the genome. This includes messenger RNA (mRNA), ncRNA, ribosomal RNA (rRNA) and small nucleolar RNA (snoRNA). The types of RNA have different functions; mRNA is important as it may be translated into protein. Other species of RNA may be involved in regulation of transcription or translation. To quantify the transcriptome, RNA is extracted from the cells or tissue of interest and cDNA synthesised from the RNA. The relative amounts of cDNA are used to assess transcript levels. There are two main methods of assaying the transcriptome, microarrays and RNA-seq. Gene expression microarrays consists of a DNA chip bound with nucleotide probes. Microarrays rely on complementary probes that bind fluorescently labelled cDNA to measure gene expression. The level of fluorescence is assessed to quantify relative abundance.

RNA-seq may also be used to determine the abundance of RNA transcripts in the transcriptome. The cDNA can be prepared into a cDNA library and sequenced in the same way as genomic DNA. Analysis of RNA-seq usually involves mapping reads to a reference genome, quantifying and normalising reads within a transcript and testing for differential gene expression between multiple conditions. RNA-seq has a larger dynamic range compared to DNA microarrays and can more accurately detect both very rare and highly abundant transcripts (Zhao et al, 2014). Gene expression microarrays have largely fallen out of fashion in favour of RNA-seq. Unlike DNA microarrays which depend on using probes to known transcripts to assay gene expression, RNA-seq has the ability to detect novel transcripts, splice sites and alternative splicing events.

1.7.2.2 DNA methylome

Whole genome methylation assays are used to characterise the DNA methylome during biological processes including disease and development. Like transcriptomics methods, the DNA methylome may be assayed using array based or sequencing based methods. To determine methylation, DNA is extracted and denatured and treated using sodium bisulfite which leads to the deamination of unmethylated cytosines to uracil. Methylated cytosines, 5mC or 5hmC, remain intact. Subsequent PCR amplifies uracil bases as thymine and 5mC or 5hmC as cytosines thus allowing for discrimination between methylated and unmethylated cytosines bases.

DNA methylation arrays are a popular method of investigating the methylome and are based on DNA gene expression microarray technology. It is carried out using a microarray chip containing probes that bind to bisulfite converted DNA sequences. The array probes incorporate a fluorescently labelled ddNTP and the intensity of the fluorescent signal represents the quantity of methylation. Methylation probes are designed to bind to specific CpG sites in the genome. Initially, 27K probe methylation arrays were offered by Illumina but this has expanded to 450K and beyond - the number of CpG sites that can be quantified has increased up to more than 850,000 in Illumina's new MethylationEPIC BeadChip. Only CpG sites with probes designed can be quantified; the use of probes is a major limitation of array-based methods because only designated sites can be assayed, an issue which also affects gene expression microarrays. Although methylation arrays are commonly described as a genome wide method, they are not truly genome wide and are biased towards regions with a high density of probes. Despite this, DNA methylation arrays remain popular whereas gene expression microarrays have given way to RNA-seq.

Whole genome bisulfite sequencing can also be used to assay DNA methylation. In this method, a DNA sequencing library is prepared from bisulfite treated DNA and sequenced. Subsequent bioinformatics analysis differentiates between thymines, which may be derived from normal thymines or unmethylated cytosines, and cytosines, derived from methylated cytosines. This approach does not rely on probes and can quantify methylation anywhere in the genome. Like DNA methylation arrays, bisulfite sequencing is unable to distinguish between 5mC and 5hmC. One of the major challenges with bisulfite sequencing is the bisulfite conversion of DNA which can lead to excessive DNA degradation. Incomplete conversion can lead to false positives if unmethylated cytosines are not converted to uracil (Leontiou et al, 2015). Another issue is inaccurate alignment of reads after conversion, read aligners designed for bisulfite sequencing data must allow for mapping asymmetry and a degree of possible errors (Chatterjee et al, 2012).

1.7.2.3 Histone modification assays

Chromatin immunoprecipitation coupled with high throughput sequencing (ChIP-seq) can be used to determine all the sites a DNA binding protein binds to in the genome. Transcription factors are commonly assayed although any protein that binds to DNA can be investigated using this method (TF ChIP-seq). Briefly, this technique involves cross-linking DNA to proteins, extracting chromatin from cells or tissues and sonicating the DNA with endogenously bound proteins then performing an immunoprecipitation using an antibody to bind the protein of interest. A DNA sequencing library is then prepared using immunoprecipitated DNA fragments (Fig. 1.10).

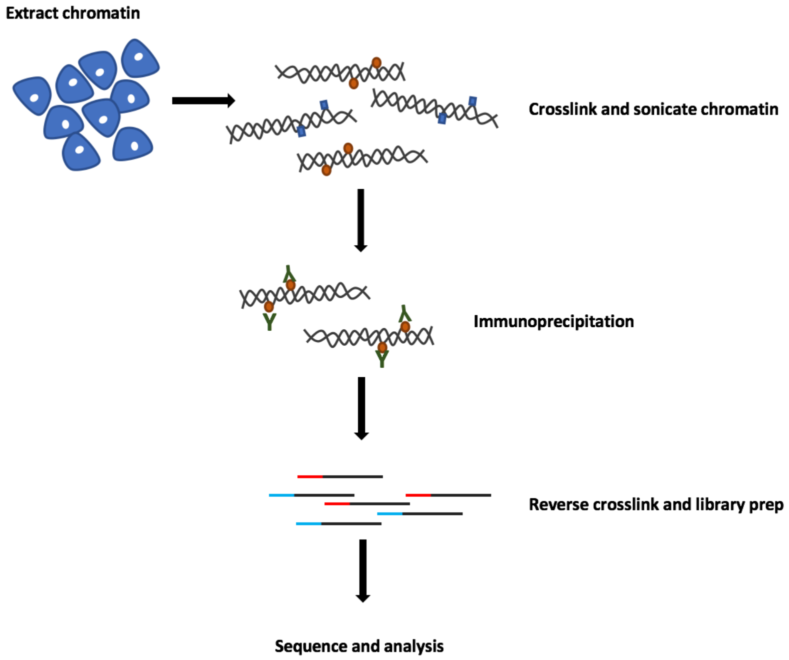

Extract chromatin
Crosslink and sonicate chromatin
Immunoprecipitation
Reverse crosslink and library prep
Sequence and analysis

Figure 1.10 – Overview of the ChIP-seq technique. Chromatin is extracted from cells or tissues, proteins are crosslinked with DNA before sonication of the chromatin. Chromatin fragments are immunoprecipitated using an antibody against a protein or histone modification. The immunoprecipitated chromatin fragments are reverse crosslinked and the DNA is prepared into a sequencing library.

Bioinformatics analysis is used to identify regions of the genome enriched with the protein after high throughput sequencing. Histone modifications in the genome can also be studied with this method (histone ChIP-seq) using antibodies against the histone modification of interest. Although similar, TF and histone ChIP-seq present with different challenges. Both rely on the availability of a good quality antibody for the protein or histone mark studied, however, the quality control and analysis of data post-sequencing is different. TF ChIP-seq produce short, sharp peak signals (narrow peaks) whereas histone ChIP-seq generally produce wider enrichment peak signals (broad peaks). This difference in signal length means different parameters are required when assessing the data and during peak calling.

Different ChIP-seq datasets can be analysed together to assess co-binding of transcription factors or generate chromatin states using multiple histone modifications. Co-occurrence of histone modifications may be used to define chromatin states (Ernst and Kellis, 2012). Promoters, enhancers, active and repressed regions can be determined using a panel of histone modifications assayed using ChIP-seq. The Encyclopaedia of DNA Elements (ENCODE) and Roadmap Epigenomics projects have used the ChIP-seq technique extensively to characterise the epigenomes of many human cell types and tissues. ENCODE, one of the early pioneers of large scale epigenomics, has published a set of guidelines for optimal ChIP-seq performance (Landt et al, 2012).

1.7.2.4 Chromatin interactome

Epigenomic marks influence the structure of the chromatin and chromatin interactions, these interactions can indicate target genes of regulatory elements. The 3D chromatin interactions in the genome can be assayed using chromatin conformation capture assays. Coupled with high throughput sequencing, methods such as Hi-C can detect all chromatin interactions present in the genome. 4C is a similar method whereby a region of the genome, e.g. a promoter or enhancer, is used as a bait to determine chromatin regions that interact with the bait region. Establishing the physical interactions in the genome can help to define enhancer-promoter pairings and identify trans-acting regulatory elements. Chromatin conformation assays can also be used to help define topologically associated domains (TADs). Regions of chromatin within the same TAD are likely to physically interact whereas separate TADs rarely interact (Pombo and Dillon, 2015).

Chromatin interactions differ between cell types and large-scale changes are seen during differentiation of stem cells (Dixon et al, 2015). TAD boundaries between cell types are largely conserved, however, interactions within and between TADs are cell type specific (Smith et al, 2016).

1.8 Bioinformatics and the ‘omics revolution

Modern bioinformatics is an interdisciplinary field that has developed over the past two or three decades to analyse and interpret biological data generated by new high-throughput technologies such as high throughput sequencing. Bioinformatics incorporates elements of computer science, mathematics and statistics and applies methods derived from these areas to biological research. In particular, *in silico* tools and methods have facilitated genome wide ‘omics research, allowing researchers to manage vast amounts of information at once. These computational approaches have greatly advanced our knowledge of the genome, transcriptome and the epigenome. The Human Genome Project (HGP; Lander et al, 2001), completed in 2003, signified the advent of the ‘omics era. Since then there have been many large-scale projects designed to interrogate various aspects of the human genome and its derivative ‘omes.

As well as the human genome, the genomes of model organisms and other species of interest have been determined. Since the HGP, thousands of prokaryote, viral and eukaryote genomes have been sequenced. The US National Centre for Biotechnology Information (NCBI), Ensembl consortium and University of California, Santa Cruz (UCSC) provide online public databases of annotated genomes and reference genome sequences. In the UK, the 100,000 genomes project orchestrated by Genomics England and the NHS aims to sequence the individual genomes of patients and their families with rare diseases and cancers.

Epigenomics projects usually incorporate DNA methylation, histone modification, transcription factor binding and chromatin conformation capture assays. The ENCODE project was initiated shortly after the conclusion of the HGP (Dunham et al, 2012). The ENCODE project is a worldwide consortium which aims to characterise all the regulatory elements in the human genome and to provide a public repository of data generated as part of this project. The project is currently ongoing and has progressed through three phases - the pilot phase, the production phase and the data analysis phase. As of 2017, the project has entered into its fourth phase which focuses on further data acquisition and analysis. The Roadmap Epigenomics project is a similar initiative to ENCODE and has so far characterised over 111 epigenomes of different human cell types and tissues (Romanoski et al, 2015). The Blueprint consortium is another large scale epigenomics project that has characterised over 100 epigenomes from haematopoietic cells. The epigenomes of some model organisms have also been investigated. For example, ModENCODE is a side project of ENCODE whose objective is to determine the functional elements within the *Caenorhabditis elegans* and *Drosophila melanogaster* genomes (Muers, 2011). The mouse also has a similar epigenome project, mouse ENCODE (Stamatoyannopoulos et al, 2012).

Sharing data is encouraged and some journals stipulate that data be made publically available once the manuscript has been published. Journals encourage authors to upload accompanying data into online repositories such as the NCBI databases sequence read archive (SRA) or gene expression omnibus (GEO). The European Bioinformatics Institute (EBI) hosts ArrayExpress, another data repository, which accepts sequencing as well as array data. The generation and sharing of large amounts of biological big data has led to new methods of analysing and interpreting results. Integration of ‘omics data allows researchers to bring together different aspects of biology to form a bigger picture. This is especially beneficial when investigating the interplay between multiple biological mechanisms that are assayed using different methods.

1.9 Rationale for PhD

Chondrocytes are the sole cell type in articular cartilage and are responsible for homeostasis of the surrounding ECM. Mutations in genes and regulatory regions of genes can lead to impaired chondrogenesis and cartilage abnormalities. It is becoming increasingly apparent that epigenetic regulatory mechanisms play important functions in all biological processes including chondrogenesis. Histone modification changes have been observed in MSC differentiation into chondrocytes in an alginate scaffold model (Herlofsen et al, 2013). Epigenetic mechanisms may be used to modulate gene expression in chondrogenesis for tissue engineering purposes and have therapeutic potential for treating OA (Huang et al, 2015; Yapp et al, 2016). Furthermore, OA chondrocytes undergo changes similar to hypertrophic chondrocytes in development (van der Kraan and van der Berg, 2012). Investigation of the epigenome during chondrogenesis will lead to a better understanding of gene transcription regulation in chondrocytes. This may facilitate engineering of chondrocytes and hMSCs better suited to tissue regeneration and implantation in diseased or injured cartilage. Epigenome editing is possible using modifying enzymes and the CRISPR-Cas9 technology. However, more knowledge of *cis*-regulatory regions and how they crosstalk with other epigenetic mechanisms such as DNA methylation during chondrogenesis is required.

Chondrogenesis in human cells is possible with the use of *in vitro* models. The scaffold-free transwell model of chondrogenesis has proven to be a good model for the study of hMSC differentiation into chondrocytes. Scaffold-free models are more likely to represent the chondrogenesis process that occurs during development. The initial objective of this project is to generate genome wide histone modification data during MSC chondrogenesis in the transwell model.

The use of histone ChIP-seq allows the genome wide investigation of multiple histone marks. By assaying a range of histone marks, we can observe patterns of co-occurrence and infer regulatory functions from existing knowledge. For this reason, we chose histone marks to reflect a range of regulatory states (Table 1.1). These histone marks were included in the ENCODE and Roadmap projects and are well characterised.

Table 1.1 – Histone modifications selected for ChIP-seq

Histone modification	Regulatory role
H3K4me3	Active promoter
H3K4me1	Active/poised enhancer
H3K27ac	Active promoter and active enhancer
H3K27me3	Transcriptionally repressive
H3K36me3	Transcriptionally permissive

Previously, our research group has generated microarray, RNA-seq and DNA 450k methylation array data using the transwell model of chondrogenesis. We observed gene expression and DNA methylation changes between day 0 and day 14 of MSC differentiation into chondrocytes. We hypothesise that by generating histone ChIP-seq data of the same model and integrating histone modification data with existing data, we will be able to define regulatory regions in the genome. DNA methylation and histone modifications often depend on each other and integrating these will lead to a greater insight into the interplay between these important epigenetic mechanisms. This will increase our understanding of chondrogenesis and identify potential targets for experimental validation. Furthermore, chondrogenesis ChIP-seq is a pre-requisite for performing similar studies in normal human articular chondrocytes, OA chondrocytes and other conditions. It has been proposed that epigenetic mechanisms may be used to enhance chondrogenesis. Therefore, better characterisation of the epigenome may lead to the development or improvement of *in vitro* models which could have therapeutic value for cartilage repair.

1.10 Project aims

- Generate histone ChIP-seq data for histone marks H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3 in an *in vitro* transwell model of chondrogenesis at day 0 and day 14. Make data publically available once published.

- Analyse histone ChIP-seq data and generate chromatin states from the histone marks assayed, identify regulatory regions such as enhancers and promoters.

- Integrate ChIP-seq data to chondrogenesis RNA-seq and DNA 450K methylation datasets.

- Identify epigenetic changes between hMSCs and differentiated chondrocytes.

- Experimentally test selected regulatory elements identified from epigenomic data.

Chapter 2. Materials and Methods

2.1 Materials

2.1.1 Equipment

ABI PRISM® 7900HT Sequence Detection System
Applied Biosystems™ QuantStudio 3
Applied Biosystems® Veriti® 96-Well Fast Thermal Cycler
Diagenode Bioruptor Standard
Diagenode Bioruptor Pico
Promega Glomax-Multi+ Detection System
NanoDrop 2000
Qubit™ 3.0 Fluorometer

2.1.2 Kits

iDeal Histone ChIP-seq kit (Diagenode, cat. no. C01010050)
PureYield™ Plasmid Miniprep System (Promega, cat. no. A1223)
PureYield™ Plasmid Midiprep System (Promega, cat. no. A2495)
Dual-Luciferase® Reporter Assay System (Promega, cat. no. E1960)
NucleoSpin Gel and PCR Clean-up (Macherey-Nagel, cat no. 740609)
Qubit dsDNA high sensitivity (HS) assay kit (Life technologies, cat. no. Q32851)

2.1.3 Antibodies

ChIP-seq grade polyclonal rabbit H3K4me1 (Diagenode, cat. no. C15410194)
ChIP-seq grade polyclonal rabbit H3K27ac (Diagenode, cat. no. C15410196)
ChIP-seq grade polyclonal rabbit H3K27me3 (Diagenode, cat. no. C15410195)
ChIP-seq grade polyclonal rabbit H3K36me3 (Diagenode, cat. no. C15410192)
SOX9 antibody, source rabbit (Millipore, cat. no. AB5535)
HRP-conjugated goat anti-rabbit secondary antibody (Millipore, cat. no. 12384)
IgG (included in Diagenode iDeal Histone ChIP-seq kit)

2.1.4 Enzymes

For cell culture: trypsin-EDTA (Life Technologies, cat. no. 25300054)
For digestion of cartilage discs: Hyaluronidase, trypsin, collagenase (all Sigma Aldrich)
RT-qPCR Taqman Gene Expression Master Mix (Applied Biosystems, cat. no. 4369510)
In-Fusion HD Enzyme Premix (Clontech, cat. no. CL639647)
Phire Hot Start II DNA Polymerase (Thermo Scientific)
Moloney Murine Leukaemia Virus (MMLV) reverse transcriptase (Life Technologies)

2.1.5 Immunoblotting reagents

TEMED (NNNN-Tetramethylethylenediamine; Sigma Aldrich)
Tween-20 (polyoxyethylene sorbitan monolaurate, Sigma Aldrich)
APS (Ammonium persulphate; BDH Chemicals)
37:5:1 Acrylaminde/bis-acrylamide (Amresco)
PVDF membrane (Immobilon-P polyvinylidene difluoride, 0.45μm, Millipore)
ECL (Enhanced chemiluminescence, Amersham Biosciences)
SeeBlue Plus 2 pre-stained protein standards (Invitrogen)
Instant dried skimmed milk powder (Tesco)
Triton X-100 (Sigma Aldrich)

2.1.6 General reagents, chemicals and alcohols

Molecular biology grade water (Sigma Aldrich, cat. no. W4502)
Formaldehyde (Sigma Aldrich, cat. no. F8775)
FuGENE® HD Transfection Reagent (Promega, cat. no. E2312)
Lipofectamine® 2000 (Life Technologies, cat. no. 11668019)
Agarose (Severn Biotech, cat. no. 301050)
Luria-Broth EZMix^{TM} powder (LB; Sigma Aldrich, cat. no. L7658)
Bacto-Agar (BD, cat. no. 214010)
Phenol:chloroform (Sigma Aldrich, cat. no. P2069)
Methanol, ethanol, isopropanol (Molecular biology grade, all Sigma Aldrich)
GlycoBlue (Invitrogen, cat. no. AM9616)

HyperLadder I and 5X sample loading buffer (Bioline, cat. no. BIO33053)
HyperLadder 100bp and 5X sample loading buffer (Bioline, cat. no. BIO33053)
Agencourt AMPure XP beads (Beckman Coulter, cat. no. A63881)
Random hexamers p(dN)$_6$ (Integrated DNA technologies, IDT-DNA)
Loading buffer (5X) and Hyperladder I (Bioline)

2.1.7 Tissue culture media/plastics

DMEM-12 by Gibco, life technologies (cat no. 11320-074)
DMEM by Gibco, life technologies (cat no. 41965-039)
MSCBM + hMSC SingleQuot kit (LONZA PT-3238 & PT-4105) + Fibroblast growth factor 2 (FGF2; R&D systems)
Chondrogenic media - made from DMEM medium with 100µg/ml sodium pyruvate (Lonza), 10ng/ml TGFβ3 (Peprotech), 100nM dexamethasone, 1x ITS-1 premix, 40 µg/ml proline, and 25 µg/ml ascorbate-2-phosphate (all from Sigma-Aldrich, Poole, UK), L-glutamine (Sigma Aldrich).
Tissue culture flasks/dishes, CORNING
Dulbecco’s phosphate buffered saline (PBS; Sigma Aldrich)

2.1.8 Antibiotics

Ampicillin, penicillin, streptomycin (all Sigma Aldrich)

2.1.9 Other

DH5α competent cells (Invitrogen, cat. no. 18265017)
Stellar competent cells (Clontect, cat. no. 636766)
PGL3-promoter vector (Promega)
Renilla pRL-TK (Promega)

2.2 Laboratory methods

2.2.1 Primary hMSC culture

Bone marrow aspirates were obtained from LONZA (Table 2.1) and hMSCs were cultured by Dr. Matt Barter (Newcastle University). Cells were phenotyped by flow cytometry and tri-linage differentiation potential was assessed by Dr. Matt Barter and Dr. Ruddy Gomez-Bahamonde (Newcastle University). Isolated hMSCs were cultured in LONZA MSCBM media supplemented with hMSC SingleQuot kit (containing L-glutamine, MSC growth supplement and Gentamicin/Amphotericin) + hFGF2. Cells were incubated at 37°C and passaged (using trypsin) every 3-4 days as necessary.

Table 2.1 – Primary hMSC donors from LONZA

Donor ID	**Sex**	**Age**
071508A	Female	22
2454e	Female	24
071671B	Female	24
071607A	Male	21

2.2.2 hMSC differentiation into chondrocytes

hMSCs were cultured in chondrogenic medium comprising of Dulbecco's modified Eagle's medium with 100µg/ml sodium pyruvate (Lonza), 10ng/ml TGFβ3 (Peprotech), 100nM dexamethasone, 1x ITS-1 premix (Insulin, transferrin, selenium+ Linoleic acid; CORNING), 40 µg/ml proline, and 25 µg/ml ascorbate-2-phosphate (all from Sigma-Aldrich, Poole, UK). For chondrogenesis, 5x10^5 or 1x10^6 hMSC were placed into 6.5mm diameter, 0.4-µm pore size polycarbonate Transwell filters (Merck Millipore) in 200µl media; the filters were placed into a 24-well plate, centrifuged at 200g for 5mins before 0.5ml chondrogenic medium was added to the well. Medium was changed every 2/3 days as necessary. hMSCs differentiated into chondrocytes and form a cartilaginous disc over time (Murdoch et al, 2007; Barter et al, 2015).

2.2.3 Cartilage disc digestion and isolation of chondrocytes

Cartilage discs were digested at day 14 of chondrogenesis. At 14 days, the chondrocytes were found to be fully differentiated in a pellet model of chondrogenesis (Johnstone et al, 1998). The following steps were optimised and these steps are the final optimised protocol. After transwell chondrogenesis and removal of cartilage discs from the transwells, the cartilage discs were chopped into ~2mm^3 pieces using a scalpel. Multiple discs (6-8) were digested in 2ml tubes simultaneously. Discs were digested first with 1000µl hyaluronidase (1mg/ml in sterile PBS) for 15mins at 37^{o}C and then centrifuged at 1500g for 5mins. The supernatant was discarded, and digested cartilage were washed with PBS before being centrifuged again as above. The discs were then enzyme digested with 1000µl trypsin (2.5mg/ml in sterile PBS) at 37^{o}C for 30mins. An equal amount of FBS containing media was used to inactivate the trypsin. The discs were centrifuged as above and the supernatant removed. The discs were finally digested with 1000µl collagenase (2mg/ml in DMEM media) for 1-1.5hrs at 37^{o}C until fully digested and matrix was no longer visible. The digested cartilage containing media was passed through a 100µm cell strainer to collect any remaining pieces of matrix. Cells were counted using a haemocytometer and centrifuged at 1500g for 5mins.

2.2.4 Chromatin extraction and sonication of hMSCs and differentiated chondrocytes

hMSCs were harvested from monolayer culture using trypsin, counted and 10 million cells were pelleted by centrifugation at 1500g for 5mins. Differentiated chondrocytes were isolated as above (section 2.23) and 10 million were pelleted by centrifugation at 1500g for 5mins. Chromatin from hMSC and MSC-derived chondrocytes were extracted using step 2a of the Diagenode iDeal ChIP-seq kit (Cat. No. C01010050). Briefly, 10ml ice cold lysis buffer iL1 was added to pelleted cells and resuspension achieved by pipetting, cells were incubated for 10mins at 4^{o}C in lysis buffer. Cells were centrifuged at 500g at 4^{o}C and the supernatant discarded. Cells were resuspended by pipetting in 10ml lysis buffer iL2 and incubated for 10mins at 4^{o}C. Cells were centrifuged as above and supernatant discarded. Protease inhibitor (200x) was added to shearing buffer iS1 and 1ml shearing buffer mix was added to chromatin, mixed by pipetting and incubated on ice for 10mins. Chromatin was stored at -80^{o}C until further use. Sonication of chromatin was performed on the same day as the immunoprecipitation.

Chromatin was sonicated using a Bioruptor Standard or a Bioruptor Pico by Diagenode. A range of cycles was used to optimise this step (1 cycle 30s on/20s off for Bioruptor Standard, 30s on/30s off for Bioruptor Pico). The number of sonication cycles ultimately used was 15 for both hMSC and chondrocytes. The size of chromatin fragments was assessed by reverse cross linking the chromatin by adding proteinase K and incubating samples at 65^{o}C for 4 hours or overnight and running the DNA fragments on a 1.5% (w/v) TAE-agarose gel (see section 2.2.8). Unused chromatin was stored at -80^{o}C.

2.2.5 Histone ChIP

Following chromatin extraction, the Diagenode iDeal ChIP-seq kit protocol step 3 was followed using H3K4me3 and IgG antibodies included with the kit. Diagenode's ChIP-seq grade H3K4me1, H3K27ac, H3K27me3 and H3K36me3 antibodies were also used. Total volume was 300µl per IP. Samples were incubated at 4^{o}C overnight on a rotating wheel (40rpm). For ChIP-seq, two IPs were performed for each antibody and immunoprecipitated DNA was pooled together post purification.

All ChIP-seq grade antibodies were validated for specificity and cross-reactivity by Diagenode using ChIP-qPCR, ChIP-seq, Dot Blot, Western blot and immunofluorescence. More information on antibody validation is available from the manufacturer.

2.2.6 Agencourt bead purification (for ChIP-seq)

Agencourt AMPure XP beads were used to purify immunoprecipitated DNA because the magnetic beads included in the Diagenode iDeal ChIP kit did not yield any DNA after purification. The volume of Agencourt beads used per purification was 45µl per 100µl of sample and a modified manufacturers protocol was followed to purify DNA. Briefly, 187.2µl Agencourt beads were added to each 416µl ChIP sample tube, mixed by pipetting and incubated at room temperature for 5min. Tubes were placed into a 6-tube Ambion magnetic stand for 2-5mins until the beads magnetized to the stand and the solution cleared. The cleared solution was discarded and the beads were washed twice with 200µl 70% (v/v) ethanol. Finally, immunoprecipitated DNA was eluted with 40µl molecular biology grade water (Sigma).

2.2.7 Phenol:chlorophorm DNA extraction (for ChIP-qPCR)

An equal volume of phenol:chlorophorm solution was added to immunoprecipitation samples and mixed by vortexing. The sample was then centrifuged for 5mins at 1500g. The aqueous layer of the sample was transferred to a new tube containing double the volume of 100% ethanol and mixed by vortexing. A 10µl aliquot of GlycoBlue Coprecipitant (Life Technologies) was added and samples were incubated at -20^{o}C for 30mins. The samples were then centrifuged at 1500g at 4^{o}C for 30mins. The supernatant was removed and the pellet resuspended in 70% (v/v) ethanol. Samples were centrifuged again as above and the supernatant discarded. Any remaining ethanol was left to evaporate and DNA was resuspended in 50µl molecular biology grade water.

2.2.8 Agarose gel electrophoresis

1-1.5% agarose gels were made by weighing out the appropriate amount of agarose and 1X TAE buffer. The solution was heated until the agarose was completely dissolved. 0.01% (v/v) ethidium bromide was added. Gels were electrophoresed at 50-100V until a sufficient level of nucleic acid separation was achieved. Loading buffer (5X, Bioline) was added to samples prior to loading on gel. Hyperladder I (Bioline) molecular weight marker was also loaded alongside samples.

2.2.9 ChIP-qPCR

To investigate H3K4me3 histone mark enrichment in promoter regions by qPCR, primers were designed that span the promoters of selected human genes (Table 2.2). The region 500bp upstream of the TSS of the gene was used to design promoter primers. Primers were also designed for the *COL2A1* intronic enhancer. Primer pairs were designed using the universal probe library assay design centre tool on the Roche life science website (https://lifescience.roche.com/en_gb/brands/universal-probe-library.html).

Table 2.2 – Gene promoter primers for ChIP-qPCR

Region	**Forward Primer (5'–3')**	**Reverse Primer (5'-3')**	**Probe**
COL2A1 promoter	tccgctgctcctttctacc	cctagaccaaggacggaaaa	52
TAGLN promoter	ccccctcttctcaaactcg	gaccctgcccggacttac	68
GAPDH promoter	caccagccatcctgtcct	cctgataattagggcagacaatc	23
HBB promoter	cagtggggctggaataaaag	tgtgagcttgcttctactctgtg	62
TNF promoter	taccgcttcctccagatgag	cattcaaccagcggaaaact	22
FN1 promoter	cttcgcttcacacaagtcca	cctttgcggtcatcaaactt	28
COL2A1 intron enhancer	gtgaggaaggtgtgggagag	gggtgggctctcctgtagt	19

2.2.10 qPCR analysis of immunoprecipitated DNA

TaqMan (Life Technologies) qPCR reactions with TaqMan Gene Expression Master Mix were performed according to manufacturer's instructions. An ABI PRISM® 7900HT Sequence Detection System was used with default cycle temperature settings; qPCR cycles were repeated 40-45 times. Enrichment of histone marks was calculated as a percentage of input as shown below:

Adjusted input = 1% input Ct – 6.644

100*2^(Adjusted input - Ct (IP) = percentage of input

A positive histone enrichment was defined as a percentage of input greater than 5% (Diagenode recommendation).

2.2.11 Quantification of nucleic acids

For expected DNA/RNA concentrations of > 10ng/µl, a NanoDrop 2000 spectrophotometer was used. A Qubit™ 3.0 dsRNA HS assay was used to quantify DNA samples prior to DNA library preparation and sequencing.

2.2.12 Histone ChIP-seq

Two different hMSC donors were used for ChIP-seq experiments, 017508A and 2454e. DNA library of ChIP-seq using donor 071508A was generated using Diagenode MicroPLEX v2 kit and single ended reads of 50bp length were generated on an Illumina HiSeq 2500. Each sample was sequenced on one lane on the flowcell. Library preparation and sequencing was performed by Diagenode, Belgium. For ChIP-seq replicate 2454e, DNA libraries were prepared using the NEBNext Ultra II kit and sequenced using an Illumina NextSeq 500, generating 75bp single ended reads. Each sample was sequenced over 4 lanes. For these samples, library preparation and sequencing was performed at the Genomics Core Facility at Newcastle University by Dr. Jonathan Coxhead. Fragment sizes of samples were assessed on a Bioanalyzer prior to sequencing. All samples were re-sheared after ChIP and before DNA library preparation due to fragment sizes being larger than expected. To achieve a sufficient number of reads, the following samples were sequenced twice: 071508A MSC input, 071508A MSC H3K36me3, 071508A CHON input, 071508A H3K36me3, 2454e CHON H3K4me3. Re-sequenced samples were merged with the original prior to alignment. IgG controls were also sequenced for donor 071508A but were not used in the final analysis.

2.2.13 Generation of chondrogenesis RNA-seq

hMSCs (donor 2454e) were differentiated into chondrocytes over 14 days. RNA was extracted from hMSCs (day 0) and differentiated chondrocytes at days 3 (not used in this project), 6 and 14. RNA was riboRNA depleted and library preparation performed after using the Ribo-Zero rRNA kit (Illumina). Paired end sequencing was performed on an Illumina HiSeq 2000 platform by GATC. Collection of RNA was performed by Dr. Matt Barter (Newcastle University).

2.2.14 Luciferase enhancer reporter assay

A selection of 6 potential super enhancers were selected for further validation; these were chosen based on their proximity to chondrogenesis related genes. Super enhancer regions were cloned into a PGL3-promoter vector (Fig. 2.1) to assess their enhancer activity. The PGL3-promoter vector is a plasmid construct containing the luciferase reporter gene (*luc*+) downstream of a minimal SV40 promoter. Suspected enhancers can be cloned upstream of the SV40 promoter, the construct is then transfected into cells and luciferase activity measured to assess the enhancer potential of the cloned insert.

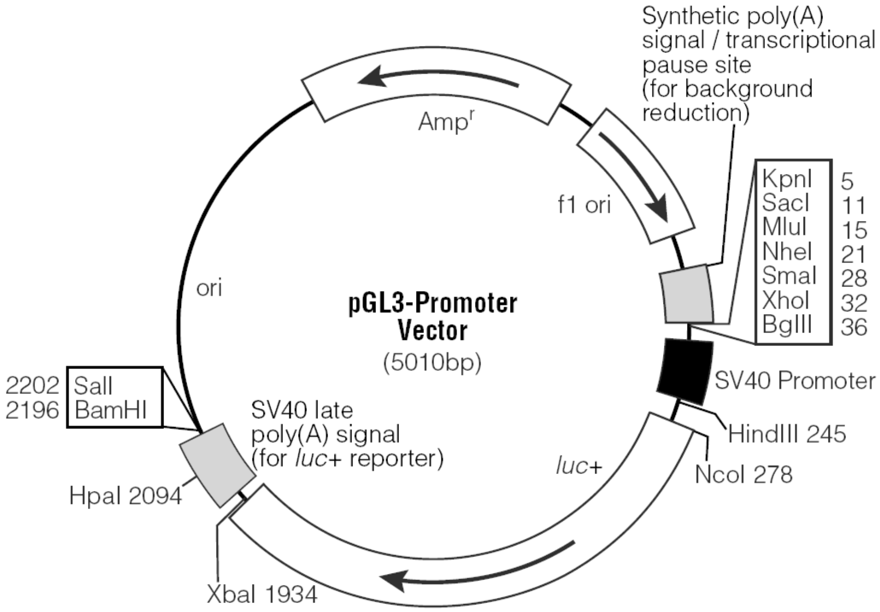


Figure 2.1 – Plasmid map of the PGL3-promoter vector comprising the luc+ gene, SV40 promoter, restriction enzyme sites and ampicillin resistance gene.

As only a limited size can be cloned into a PGL3-promoter vector, only a region containing a SOX9 peak within a chosen super enhancer was amplified and cloned (Table 2.1).

2.2.14.1 PGL3-promoter vector expansion

The PGL3-promoter vector was expanded by transforming into Subcloning Efficiency DH5α competent cells (Invitrogen). Briefly, 50µl DH5α cells were thawed on ice before 1ng PGL3 promoter vector was added and gently mixed. The sample was incubated on ice for 30mins and heat shocked at 37°C for 20secs. The sample was placed back on ice for 2mins before 950µl pre-warmed LB added and incubated at 37°C for 1hr. An aliquot of 150µl transformed bacteria was plated onto an LB-agar + Amp plate and incubated overnight at 37°C. The next day, one colony was picked and grown in 150ml LB broth + Amp. The plasmid was purified the subsequent day using the Promega Plasmid Midiprep System, quantified using a Nanodrop 2000 and stored at -20°C until further use. Primer extensions for In-Fusion cloning were designed using Clontech In-Fusion primer design web tool. The PGL3-promoter vector was inputted into the tool along with BglII restriction enzyme to generate primer extensions.

2.2.14.2 Enhancer PCR primers

Primers for regions containing SOX9 binding sites within a super enhancer were designed using Primer3web (version 4.0.0). Primer extensions for InFusion cloning into the PGL3-promoter vector designed previously were added to the forward and reverse primers for enhancer regions (Table 2.3). Phire hot start II DNA polymerase (Finnzymes) was used to amplify super enhancer regions. Each PCR reaction was prepared as follows: 0.4µl 5X Phire enzyme (1000 units/1.25ml), 50ng genomic DNA extracted from SW1353 cell line, 1µl forward primer (10µM), 1µl reverse primer (10µM), 1.6µl dNTP (2.5mM), Sigma water up to 20µl. A range of annealing temperatures were used – 61°C, 65°C, 69°C. PCR cycle conditions were as follows: 98°C for 30s, then cycle 35 times at 98°C for 5s – range of annealing temperatures for 5s – 72°C for 30s, final step 72°C for 1min. PCR products were electrophoresed on a 1% (w/v) agarose-TAE gel to check expected sizes of enhancer inserts.

Table 2.3 – Super enhancer In-Fusion primers for cloning into PGL3-promoter vector. InFusion primer extensions are highlighted in bold

SOX9 peak locus (hg19)	**Forward primer + extension 5’-3’**	**Reverse primer + extension 5’-3’**	**Predicted gene target**	**Insert ID**
Chr2:74810421-74811389	**CCCGGGCTCGAGATC**CCCCTTATAGTAGAGAACCAAGC	**TGCAGATCGCAGATC**gcCTCGGTTGCATTGCTTTA	LOXL3	1
Chr16:69957777-69959206	**CCCGGGCTCGAGATC**TGTTCCTTTGCCTCTGTTGC	**TGCAGATCGCAGATC**AGGGAATGCAGTGGGACTTT	WWP2	2
Chr16:69924387-69924932	**CCCGGGCTCGAGATC**GCTTTGTGTCCAGCTACTCC	**TGCAGATCGCAGATC**ACACCCTTCTCTGGACCATC	WWP2	3
Chr16:17452680-17453167	**CCCGGGCTCGAGATC**CATGGTCTGGGGAAGGTCTT	**TGCAGATCGCAGATC**TGAATGGCAGCTCACCTAGA	XYLT1	4
Chr12:104879495-104880612	**CCCGGGCTCGAGATC**TCCCTGACATTGCCAGTCTT	**TGCAGATCGCAGATC**CATGTTCAGCTGCAATGGGA	CHST11	5
Chr1:183922069-183922664	**CCCGGGCTCGAGATC**GGTCATGCCTCATCCCCTAA	**TGCAGATCGCAGATC**TCTCTCGGTTCCCTAGGTGA	COLGALT2	6

PCR products displaying the correct expected size were purified using the Nucleospin gel and PCR cleanup kit (Macherey-Nagel) following manufacturers protocol for PCR purification; PCR reactions from different annealing temperatures showing the same size were pooled prior to purification. Briefly, the volume of each PCR sample was adjusted to 100µl before 200µl buffer NT was added in a spin column and collection tube. Samples were centrifuged at 11,000g and the flowthrough was discarded, 650µl buffer NT3 was added to the spin column and samples were centrifuged again as above. The flowthrough was discarded and samples were centrifuged dry as above. Spin columns were placed into a new 1.5ml tube and 25µl buffer NE added directly to the spin column filter. Samples were incubated for at least 1min at room temperature before eluting by centrifuging for 1min at 11,000g. DNA was quantified using a Nanodrop 2000.

2.2.14.3 PGL3-promoter vector linearization and purification

The PGL3-promoter vector was linearised using restriction enzyme *BglII*. The reaction was prepared as follows: 5µg PGL3-promoter vector, 5µl buffer M (Roche), 2µl *BglII* (Roche), dH_2O (Sigma Aldrich) up to 50µl. The reaction was incubated at 37°C overnight. The total reaction was subsequently electrophoresed on a 1% w/v TAE agarose gel. The correct band size was excised under UV light and DNA was extracted using the Nucleospin gel and PCR cleanup kit following instructions for gel extraction. The protocol is the same for PCR extraction except for the initial step –the excised gel band was weighed and 200µl buffer NT1 was added for every 100mg gel, samples were incubated at 50°C until fully dissolved.

2.2.14.4 Cloning of enhancer insert into linearised PGL3-promoter vector

The Clontech InFusion cloning kit was used to insert an enhancer region into linearised PGL3-promoter vector. The reaction was prepared as follows: 2µl 5X InFusion HD enzyme premix, 100ng linearised PGL3-promoter vector, 100ng purified enhancer insert, dH_2O (Sigma Aldrich) up to 10µl. The reaction was incubated at 50°C for 15mins.

2.2.14.5 Expansion of cloned vector construct and purification

To expand the cloned vector, the construct was transformed into Stellar competent cells (Clontech) using manufacturers protocol. Briefly, an 50µl aliquot of Stellar competent cells were thawed on ice and mixed gently with 2.5µl InFusion reaction product and incubated on ice for 30mins. Cells were heat shocked at 42°C for 45secs and placed back on ice for 2mins before 450µl pre-warmed SOC (37°C) was added to the cells. Cells were incubated at 37°C for 1hr; after incubation, 100µl transformed cells were spread onto LB+ampicillin plates and incubated overnight at 37°C. The following day, 4-6 colonies were picked and each grown in 3ml LB broth plus ampicillin for 16 hours at 37°C in a shaking incubator at 200rpm. A 1.5ml aliquot was pelleted, plasmid extracted in a quick miniprep and restriction enzyme digested to check whether the enhancer insert was successfully cloned into the vector. Cells were pelleted by centrifuging for 15secs, the supernatant removed and pellet resuspended in 100µl Qiagen plasmid prep buffer P1, 200µl buffer P2 and 150µl buffer P3 were subsequently added to the sample. The sample was mixed by inversion (10X) and centrifuged at maximum speed for 3mins. The supernatant was transferred to a new tube with 1ml 100% ethanol and vortexed before centrifuged for 10mins. The ethanol was removed and sample left to air dry for approximately 10mins. The plasmid was then resuspended in 50µl water (Sigma Aldrich). Restriction enzyme digestion was performed as follows: 10µl miniprep plasmid, 1.5µl buffer 2.1 (NEB), 0.5µl *HindIII* enzyme (20,000 units/ml, NEB), 0.5µl *XhoI* enzyme (20,000 units/ml, NEB), 11.5µl water (Sigma Aldrich). The reaction was incubated at 37^{o}C for 2hr. The digested vector was electrophoresed on a 1% w/v TAE agarose gel for 30mins at 100v. Samples showing the correct insert size were purified from the original overnight culture using the Promega PureYield Miniprep system following manufacturers protocol. Briefly, 1.5ml bacterial culture was pelleted at maximum speed using a microcentrifuge, the supernatant was discarded and the pellet was resuspended in 600µl water (Sigma Aldrich). Cell lysis buffer was added at 100µl and mixed by inversion (6X), 350µl cold (4-8^{o}C) neutralisation solution and mixed by inversion. Samples were then centrifuged at maximum speed for 3mins, the supernatant was then transferred into a spin column plus collection tube and centrifuged again for 15secs and the flowthrough discarded. The samples were washed by adding 200µl Endotoxin Removal Buffer and centrifuging at maximum speed for 15secs, 400µl Column Wash Solution was added and the samples centrifuged again for 30secs. The spin column was then placed into a clean 1.5ml tube and 30µl water (Sigma Aldrich) added directly to the spin column filter. The samples were incubated at room temperature for 1min before eluting by microcentrifugation for 15secs at maximum speed. A Nanodrop 2000 was used to measure plasmid construct concentration. Samples were stored at -20^{o}C until further use.

2.2.14.6 Plasmid transfection

SW1353 cells were seeded into 96-well plates at a density of 5000-6000 cells per well in 100µl medium. The following day at approximately 50% confluency, cells were transfected as follows: 25ng enhancer PGL3 construct, 50ng *SOX9* overexpression plasmid (pUT-FLAG-SOX9; Lefebvre et al, 1997) or empty vector control, 2.5ng Renilla control plasmid and 3µl:1µg FuGENE transfection reagent to total DNA per 100µl total volume of DMEM (nil). The transfection mix was prepared and incubated at room temperature for 15mins prior to transfection. Transfected cells were incubated at 37^{o}C for 24hrs. Replicates were performed at n = 6 for each enhancer vector construct. The transfection experiment was repeated in HEK293T cells as above, but using transfection reagent Lipofectamine 2000.

To measure luciferase activity, the Promega Dual Luciferase Reporter Assay System kit was used. The buffers were prepared as follows: Luciferase Assay Reagent was prepared by resuspending luciferase assay substrate in 10ml luciferase assay buffer and Stop & Glo reagent was prepared by adding 1 volume of Stop & Glo substrate to 50 volumes Stop & Glo buffer. Reagents were stored at -80^{o}C when not in use and covered at all times to protect from light. To measure luciferase activity in transfected cells, media was aspirated away and 30µl 1X lysis buffer (Promege Dual Luciferase Reporter kit) was added to each 96-well and the samples placed on a rocker for 15mins. After lysing, 10µl of each lysed cell sample was transferred to a 96-well black and white domino plate. A Glomax luminometer was used to measure luminescence. The conditions were used as follows: dual luciferase two injections setting, 0.5s delay, speed 200µl/s, 40µl per injector. Injectors were primed one at a time before use and cleaned using default cycles after use. To normalise, sample values (measured by injector containing Luciferase Assay Reagent) were divided the background Renilla control (measure by the injector containing Stop&Glo reagent). The mean was calculated from replicate samples. A two sample T test with a Welch's correction was performed to test for significance (p < 0.05).

2.2.15 Western blot

A protein immunoblot was performed to assess SOX9 overexpression in transfected cells. To confirm overexpression of SOX9, cells were transfected as described in section 2.2.14.6. Cell numbers and volumes of reagents were scaled accordingly to 6-well plate format.

2.2.15.1 Cell lysis method

Medium was aspirated from transfected cells in 6-well plates. The plate was placed on ice and cold PBS was added to each well and the plate swirled. All PBS was then aspirated from the wells and 150µl magic lysis buffer ((50mM Tris-HCl, pH 7.4, 10% (v/v) glycerol, 1mM EDTA, 1mM EGTA pH 8.0), 1mM Na3VO4, 5mM NaF, 10mM β-glycerol phosphate, 5mM Na4P2O7, 1% (v/v) Triton X-100, 1µM microcystin-LF and 1 Complete protease inhibitor Mini tablet (Roche)) was added. Cells were scraped from the plate using a cell scraper and cell lysate was mixed by pipetting before transfer into a cooled 1.5ml tube and rotated at 4°C for 20mins. Lysates were centrifuged at 10,000g for 3mins at 4°C. The supernatant was then removed, snap frozen on dry ice and stored immediately at -80°C until further use.

2.2.15.2 Protein concentration quantification

A Qubit protein assay (Life Technologies) was used to quantify protein samples. Protein standards were prepared following manufacturers method. All solutions were prepared in Qubit Assay tubes. Qubit working solution was prepared by adding Qubit Protein Reagent to Qubit Protein Buffer at a ratio of 1:200. Each of the 3 standards weas prepared by adding 190µl Qubit working solution to 10µl standard. Protein samples were prepared by adding 10µl protein lysate and 190µl Qubit working solution. All prepared standards and samples were vortexed and incubated at room temperature for 15mins. The protein assay protocol on the Qubit 2.0 Fluorometer was selected and calibrated using the standards before measuring sample concentrations.

2.2.15.3 SDS-PAGE and immunoblot

Cell lysates were thawed on ice and Laemmli sample buffer (0.1 M Tris-HCl, pH 6.8, 0.35M SDS, 20% (v/v) glycerol, 0.01% bromophenol blue and 10% (v/v) β-mercaptoethanol) added to 1/5 of the volume. The resulting mixtures were heated to 100ºC for 5 minutes, cooled on ice, and then electrophoresed on a 12.5% SDS (w/v) polyacrylamide gel. The separating gel was prepared as follows: 3ml dH_2O, 1.5ml lower gel buffer (1.5M Tris-Base pH8.8, 0.4% SDS (w/v)), 1.5ml acrylamide/bis 40% solution, 30µl APS, 10µl TEMED (added last). The stacking was prepared as follows: 5ml stacking solution (0.5M Tris- HCl pH 6.8, 0.4% (w/v) SDS), 30µl APS, 10µl TEMED (added last). Gels were set up in a Bio-Rad PROTEAN gel electrophoresis unit and electrophoresed for approximately 1hr at 120V. Samples were then transferred to PVDF membrane by electroblotting in a Scie-Plas V20-SDB 20 x 20cm semi-dry blotter for 1-1.5 hours at $1mA/cm^2$ in transfer buffer. Membranes were then covered in blocking buffer (TBS-T & 5% w/v non-fat dry milk powder (w/v)) for 1 hour at room temperature. The membranes were washed 3 times for 5mins each in TBS-T. Washed membranes were incubated overnight at 4ºC with SOX9 primary antibody, diluted 1:1000 in primary antibody dilution buffer (TBS-T & 5% w/v non-fat dry milk powder), with gentle agitation. The membrane was washed 3 times for 10mins each and then incubated with 5ml secondary antibody (HRP-conjugated) diluted 1/1000 in TBS-T and 5% milk for 1hr. After incubation, the membrane was washed again as above. The membrane was developed using ECL plus solution (Promega), made up by adding 25µl solution A to 1ml solution B. The prepared ECL plus solution was added to the membrane and incubated at room temperature for 5mins. ECL solution was removed and the membrane was visualised on a ChemiGenius II BioImager.

2.2.16 siRNA depletion of LOXL1-4 genes

2.2.16.1 Transfection of hMSCs

siRNAs targeting *LOXL1*, *LOXL2*, *LOXL3* and *LOXL4* were purchased from Dharmacon (ON-TARGETplus SMARTpool siRNA). SMARTpool siRNAs are a mix of 4 siRNAs designed to the region of interest to maximize potency and minimize off-target effects. siRNAs were prepared to a stock concentration of 20µM by adding 250µl siRNA buffer (Dharmacon). hMSCs (donor 2454e) were transfected with siRNAs for 48hrs before differentiation into chondrocytes. For each siRNA transfection, 250,000 cells were seeded onto 6cm tissue culture dishes in 3ml hMSC medium. After 24hrs, cells were transfected with 50nM siRNA using Dharmafect 1 transfection reagent according to manufacturers instructions. Briefly, for each siRNA pool, 50nM siRNA was added to serum free medium to a volume of 200µl. A siRNA control was included. In a separate tube, 4µl Dharmafect 1 was added to 196µl serum free medium. Both solutions were mixed, incubated at room temperature for 5mins and then combined for 20mins before 1600µl hMSC media added. hMSC medium was aspirated from the 6cm dishes containing seeded hMSCs and replaced with 2ml transfection media. After 48hrs, hMSCs were differentiated into chondrocytes using the transwell insert model of chondrogenesis. Prior to differentiation, 50,000 hMSCs from each siRNA experiment were pelleted and frozen in 1.5ml tubes at -20^{o}C to be processed at the same time as differentiated chondrocytes to reduce technical effects.

2.2.16.2 RNA extraction and reverse transcription

RNA from transfected hMSCs and differentiated chondrocytes at day 7 was extracted. RNA was extracted from cartilage discs and hMSCs (frozen previously) using TRIzol reagent (Invitrogen). Briefly, 250µl TRIZol was added to cartilage disc and hMSC samples in a 1.5ml tube and incubated at room temperature for 5mins. Cartilage discs were ground using a pestle until disintegrated, 50µl chloroform was added to each sample and mixed thoroughly by shaking. Samples were incubated at room temperature for 2 mins before centrifugation at 2,000g for 15mins at 4^{o}C. The aqueous phase was transferred to a fresh 1.5ml tube containing 125µl isopropanol and vortexed. Samples were incubated at room temperature for 10mins before centrifuging as above for 10mins. The supernatant was removed and the pellet washed with 250µl 75% (v/v) ethanol. Samples were centrifuged as above for 5mins, ethanol was removed and the pellet air dried until ethanol was fully removed. RNA was resuspended in 20µl water (Sigma Aldrich) and the concentration determined using a Nanodrop 2000.

RNA was reverse transcribed into cDNA using reverse transcriptase. The protocol per reaction was as follows: 500ng RNA in 8µl water (Sigma Aldrich), 3µl dNTPs (10mM) and 1μl random hexamers (1μg/μl; IDT-DNA) were heated to 70°C for 5mins before being chilled on ice. 4μl 5X RT buffer (Invitrogen), 2μl 0.1M DTT (Invitrogen), 0.25μl MMLV (200 units/l Invitrogen) and 1.75μl water (Sigma Aldrich) was added to each sample. Samples were heated to 37°C for 50mins and 75°C for 15mins before 30μl water (Sigma Aldrich) was added. Samples were frozen at -20°C until further use.

2.2.16.3 RT-qPCR

Gene expression assayed using TaqMan RT-qPCR with probe library primers (Table 2.4), normalized against 18S expression. RT-qPCR cycles were as follows: 95°C for 10mins, 40 cycles of 95°C for 30sec, 60°C for 30sec and 72°C for 30sec.

Table 2.4 – Roche universal probe library LOXL1-4 primers and ABI assays for SOX9 and ACAN used for TaqMan RT-qPCR

Gene Name	**Forward 5’ – 3’**	**Reverse 5’ – 3’**	**Universal probe library no. or sequence**
LOXL1	accagggcacagcagactt	gtggctgcatccagtaggtc	87
LOXL2	ggatctggcacgactgtca	accttggtgccattgagg	62
LOXL3	caggaccagcactcttctcc	cactgacaggtcgcatgg	15
LOXL4	ccagcttctgtctggaggac	aagttggcacatgcgtagc	43
SOX9	acttgcacaacgccgag	ctggtacttgtaatccgggtg	5'-FAM-TCTGGAGACTTCTGAACGAGAGCGA-IABkFQ-3'
ACAN	agcgagttgtcatggtctg	tgtgggactgaagttcttgg	5'-FAM-CTGGGTTTTCGTGACTCTGAGGGT-IABkFQ-3'
18S	cgaatggctcattaaatcagttatgg	tattagctctagaattaccacagttatcc	5’FAM-TCCTTTGGTCGCTCGCTCCTCTCCC0-TAMRA 3’

2.2.17 GAG assay

2.2.17.1 Pellet culture chondrogenesis

hMSCs were differentiated into chondrocytes using a pellet culture. hMSCs were seeded into wells of a 96-well tissue culture plate at a density of 50,000 cells per well in PBS. The 96-well plate was centrifuged at 200g for 5mins to pellet hMSCs in the wells. PBS was aspirated and replaced with 150µl chondrogenic media and the cells incubated at 37^{o}C. Medium was refreshed every 2-3 days as necessary.

2.2.17.2 GAG assay

Medium was aspirated from cartilage pellets at 7 days. To digest cartilage pellets, 70µl phosphate buffer and a pre-mixed solution of 20µl papain, 10µl cys-HCl and 10µl EDTA was added to each 96-well containing a pellet and incubated at 65^{o}C for 2-4hrs. Chondroitin sulphate standards at concentrations of 0-40µg/ml were prepared with phosphate buffer. Dimethyl-methylene blue (DMB) reagent was prepared as follows: 3.04g glycine, 2.37g NaCl, 95ml 0.1M HCl, up to 1l dH$_2$O and 16mg DMB. To a fresh 96-well plate, 40µl digested cartilage pellet or standard was added to each well in duplicate and 250µl DMB solution was added to each well. Absorbance was measured using a plate reader at 530nm. A standard curve was created using the known concentrations of chondroitin sulphate and the concentrations of samples were determined using the standard curve.

2.3 Bioinformatics Methods

2.3.1 Software

If the software has an associated published paper this is referenced, otherwise links to the main website/manual page are given. Website links current as of August 2017.

Scripts are deposited into GitHub at:
https://github.com/kathleencheung/PhD_Young_lab

2.3.1.1 List of software used in no particular order:

Fastqc v0.11.5 (http://www.bioinformatics.babraham.ac.uk/projects/fastqc)
MultiQC v0.8 (http://multiqc.info/)
FASTX-toolkit (http://hannonlab.cshl.edu/fastx_toolkit/)
Bowtie2 v2.2.4 (Langmead and Salzberg, 2012)
HISAT2 v2.0.4 (Pertea et al, 2016)
SAMtools v1.6 (Li et al, 2009)
BEDtools v2.26.0 (Quinlan and Hall, 2010)
UCSC utilities (https://genome.ucsc.edu/)
ENCODE tools (https://www.encodeproject.org/software/)
Ngs.plot v2.61 (Shen et al, 2014)
HOMER v4.7 (http://homer.ucsd.edu/homer/)
MACS2 v2.1.0.20150731 (Feng et al, 2012)
Salmon v0.7.2 (Patro et al, 2015)
ChromHMM v1.12 (Ernst and Kellis, 2012)
Picard tools v1.130 (https://broadinstitute.github.io/picard/)
PETModule (Zhao et al, 2016)
GREAT v3.0.0 (McLean et al, 2010)

2.3.1.2 Bioconductor packages:

DiffBind v 2.2.12 (Stark and Brown, 2011)
SPP v1.13 (Kharchenko et al, 2008)
Tximport v1.6.0 (Soneson et al, 2015)
Minfi v.1.20.2 (Aryee et al, 2014)
Missmethyl v0.99.0 (Phipson et al, 2016)
Limma v3.30.13 (Ritchie et al, 2015)
Sva v3.22.0 (Leek et al, 2012)

2.3.1.3 Genome browsers:

Integrative Genome Viewer v2.3.91 (Robinson et al, 2011)
UCSC genome browser (Kent et al, 2002)

2.3.2 QC and alignment of ChIP-seq reads

For quality control of raw sequencing reads, the FastQC tool was used. Aggregation and summarisation of multiple FastQC reports was performed using MultiQC. Alignment of sequencing reads to human reference genome hg38 was performed by bowtie2. A bowtie2 index for hg38 was generated using the bowtie2-build indexer with default settings. Reference genome hg38 in fasta format was downloaded from UCSC, as were chromosome size and annotation files.

2.3.3 Normalisation of aligned ChIP-seq reads

Output SAM alignment files from bowtie2 were converted to sorted BAM files using samtools *view* and *sort* respectively. For genome browser visualization, BAM files were then normalised to reads per million (RPM) scaled genome coverage files in bedgraph format using bedtools genomecov, using the -scale option with RPM values. RPM scales for each sample were calculated by finding the total aligned reads using samtools flagstat and dividing 1 million by this number. Bedgraph files were converted to bigwig format using the bedGraphToBigWig program from ENCODE.

2.3.4 Peak calling and peak annotation

Peaks were called using MACS2 peak caller using input samples as background controls with the --broad option turned on, a *q* value cutoff of 0.05 and the effective hg38 genome size given as 3.05x10^9. The effective genome size represents the portion of the genome to which sequencing reads can be mapped. This was calculated using UCSC tool faCount to determine the total number of mapped bases and unmapped bases (denoted by N) in hg38, the number of unmapped bases was subtracted from the total number of bases to give the effective genome size. For hg38, 3.05x10^9 represents 95% of the total genome size. Peaks were annotated by HOMER annotatePeaks.pl which associates peaks to the nearest gene and overlapping genomic feature.

2.3.5 ChIP-seq sample QC metrics

Quality metrics calculated for ChIP-seq samples were PCR bottleneck coefficient (PBC), fraction of reads in peaks (FRiP), normalized strand cross-correlation coefficient (NSC) and relative strand cross-correlation coefficient (RSC). PBC was calculated by dividing uniquely mapped reads by total mapped reads. Duplicated and unique reads were found using Picard tool MarkDuplicates. FRiP was calculated using the Bioconductor package DiffBind in RStudio. NSC and RSC were calculated using ENCODE tool Phantompeakqualtools and SPP.

2.3.6 Comparison and correlation of ChIP-seq replicates

Bioconductor package DiffBind was used to generate a correlation heatmap of ChIP-seq replicates from hMSC donor 2454e and 071508A. Overlap of peaks was also assessed using DiffBind and Venn diagrams created using the Venneuler package in RStudio.

2.3.7 RNA-seq expression quantification and analysis

QC of raw RNA-seq reads was performed using Fastqc and summarised using MultiQC. To improve quality scores, reads were subject to a hard trim (101bp to 90bp) using fastx_trimmer (Appendix ii, Fig. 1). For hMSCs at day 0, 52.8 million paired end reads were achieved and 47.5 million paired end reads were achieved for day 14 chondrocytes (Appendix ii, Table 1). Gene abundance in transcripts per million (TPM) was quantified using Salmon in quasi-mapping mode. A hg38 Salmon transcriptome index was generated using the Salmon indexer with default settings. Transcripts were summarized to gene level using the tximport Bioconductor package in RStudio. Hg38 reference genome files were downloaded from UCSC. TPMs were extracted using Bioconductor package tximport in R. Gene expression plots were generated using the ggplot2 package in R. Genes with a log2 fold change > 1.5 (actual fold change 2.83) were considered to be differentially expressed. Log2 fold change was calculated as follows: log2(day14 TPM + 1) – log2(day0 TPM + 1). A value of 1 was added to all TPM values to avoid negative logged values; this also removes undue influence of very small expression values. GO terms were found using DAVID gene list analysis. Log2 fold change TPM values were correlated with changes from chondrogenesis cDNA microarray data (Barter et al, 2015) and a Pearson's correlation test performed.

2.3.8 Correlation of histone mark enrichment to RNA-seq data

Read counts per million for each histone mark were plotted against subsets (high, medium and low expressed) of genes to explore the association of histone mark enrichment and gene expression. Genes with a TPM < 2 were considered low expressed; a density plot of TPM values was created and a gene expression cut off of 2 was chosen; this was the threshold where the highest density peak sharply decreases (Appendix ii, Fig. 2). Genes with medium expression had a TPM of between 2 and 16.7; 16.7 was the mean TPM in day 0 and day 14 samples. Genes with a TPM higher than the mean were considered highly expressed. Read counts per million mapped reads (reads per million; RPM) were plotted against genes in each expression level set. Plots were generated using ngs.plot. Histone peaks were associated to the nearest gene using HOMER annotatepeaks.pl. The BiomaRt package was used to query the BioMart database in order to annotate genes.

2.3.9 Differential histone binding analysis

Significant differential binding sites for H3K4me1 and H3K4me3 samples between hMSCs and differentiated chondrocytes were determined using the DiffBind package in RStudio. hMSC and differentiated chondrocyte samples from both replicates were combined in this analysis. The default DESeq2 method was used. More information is available in the DiffBind vignette.

2.3.10 Chondrogenesis DNA microarray

Chondrogenesis microarray data was kindly provided by Dr. Matt Barter (Newcastle University) who carried out the experiment and analyzed the data. Correlation plots and tests were performed using RStudio.

2.3.11 Chromatin State Learning using ChromHMM

BED alignment files were used as input into the software ChromHMM. ChromHMM uses a multivariate Hidden Markov Model (HMM) to compute emission probabilities of histone marks in a chromatin state by modelling the presence or absence of histone marks in 200bp bins across the genome. The collective ChIP-seq alignment tracks can be considered a multivariate sample i.e. each 200bp bin in the genome displays a set of observations, in this case the observations are the read counts from each histone mark sample. If available, the read counts are normalised using a control. ChromHMM trains a model using the input data to calculate the probability of a set of observations that are present in an unknown (hidden) state. The program takes the number of hidden states from the user and calculates the emission and transition probabilities for those states. A full description of the method is available in Ernst and Kellis, 2010.

To identify chromatin states, the following subcommands were performed. Briefly, the BinarizeBed command was used to convert aligned read coordinates into a binarised data format for chromatin state learning. Alignment files from all ChIP-seq samples (both replicates) were used along with their controls. The command LearnModel was then used to compute chromatin states from the binarised alignment data. The number of states specified to the program started at 8 and increased by increments of 2 until there was a sufficient separation of distinct states. The final state number used was 16. Chromatin states were annotated using information from the Epigenomics Roadmap project and the Reorder command was used to re-label, re-order and assign track colours to chromatin states. The MakeBrowserFiles, MakeSegmentation, NeighbourhoodEnrichment and OverlapEnrichment commands were then used to create associated plots and browser tracks with the annotated chromatin state names. Chromatin state browser tracks (BED format) were visualised in IGV genome browser.

2.3.12 Assessing chromatin state changes

Genomic co-ordinates were split into 200bp bins using BEDTools window. The hMSC and differentiated chondrocyte chromatin state at each 200bp bin was noted and the frequency of change between hMSCs and differentiated chondrocyte was calculated. A frequency plot was generated using ggplot2 and a corresponding Sankey plot created using the riverplot package in RStudio.

2.3.13 Analysis of DNA 450k methylation arrays

An Infinium HumanMethylation450 BeadChip array was used to quantify DNA methylation of known CpGs in the human genome. hMSCs from four donors were differentiated into chondrocytes (Table 2.5) using the transwell model of chondrogenesis. A total of 12 samples were included in the chondrogenesis 450K array, 7 hMSC samples and 5 chondrocyte samples. Adipogenesis and osteoblastogenesis DNA 450k methylation was performed at n = 3 (Table 2.6). Laboratory work and data collection was performed by Dr. Matt Barter (Newcastle University; chondrogenesis), Catherine Bui (Newcastle University; chondrogenesis) and Dr. Ruddy Gomez-Bahamonde (Newcastle University; chondrogenesis, adipogenesis and osteoblastogenesis).

Raw files were pre-processed using the Bioconductor package minfi. Data were normalised using the functional normalisation algorithm developed for 450k analysis (Fortin et al, 2014). Significantly differentially methylation CpGs positions (DMPs) between hMSCs and chondrocytes were determined using the lmfit() and ebayes() functions in limma. The experiment design included both paired and unpaired samples and this was reflected in the phenotable. DMPs were calculated using a mixed paired and unpaired analysis in limma after applying SVASeq batch correction (Leek et al, 2012).

Table 2.5 – hMSC donors and samples used for chondrogenesis DNA 450K methylation array.

Donor ID	Sample ID	Cell type	Sex	Age
071508A	MSC0_8A_1	hMSC	Female	22
071508A	MSC0_8A_2	hMSC	Female	22
071508A	MSC0_8A_3	hMSC	Female	22
071508A	MSC14_8A_1	Chondrocyte	Female	22
071508A	MSC14_8A_2	Chondrocyte	Female	22
071508A	MSC14_8A_3	Chondrocyte	Female	22
2454E	MSC0_ruddy_2454E	hMSC	Female	24
071508A	MSC0_ruddy_8A	hMSC	Female	22
071671B	MSC0_ruddy_1B	hMSC	Female	24
071607A	MSC14_7A	Chondrocyte	Male	21
2454E	MSC0_2454E	hMSC	Female	24
2454E	MSC14_2454E	Chondrocyte	Female	24

Table 2.6 – hMSC donor and samples used for adipogenesis and osteoblastogenesis DNA 450k methylation array

Donor ID	Sample ID	Cell type	Sex	Age
3728A	R01C01	hMSC	Female	19
3728A	R02C01	hMSC	Female	19
3728A	R03C01	hMSC	Female	19
3728A	R04C01	Adipocyte	Female	19
3728A	R05C01	Adipocyte	Female	19
3728A	R06C01	Adipocyte	Female	19
3728A	R01C02	Osteoblast	Female	19
3728A	R02C02	Osteoblast	Female	19
3728A	R03C02	Osteoblast	Female	19

2.3.14 GO term analysis

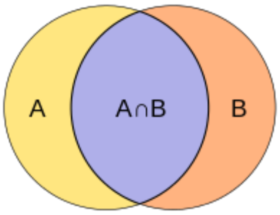
The Bioconductor missMethyl package was used to find GO terms for the top 500 hypomethylated CpG sites. The GREAT GO ontology webtool was used to associate GO terms to chromatin states after conversion from hg38 to hg19 using UCSC liftover. GREAT default settings associates genomic features to the nearest genes using a basal plus extension rule. Genes are assigned a regulatory feature area which is extended 5kb upstream and 1kb downstream of the gene, until it meets another gene. This area is extended up to a maximum of 1000kb if no nearby genes are found in the initial search area. This extended area is defined as the region where regulatory features are likely to be found and genes may share regulatory regions. The inputted chromatin states (or other regulatory features) are associated by overlap to the extended gene regulatory area. More than one gene may be assigned to a regulatory feature. Default basal plus extension settings were used to associate enhancer states to nearby genes and a simple nearest gene approach (within 1000kb) was used to find GO terms for all other chromatin states. Chromatin states may be associated with multiple genes if they overlap more than one gene regulatory region. The whole genome was used as a background control. GREAT output was viewed in the *Significant By Region-based Binomial* setting recommended for large datasets. The top 20 significant (p < 0.05) GO terms were selected for visualisation.

2.3.15 Comparison of chondrogenesis chromatin states to Roadmap states

Chondrogenesis chromatin state genome co-ordinates were converted from hg38 to hg19 using UCSC's liftover tool for this analysis. Equivalent chromatin states between Roadmap's extended 18 state model and chondrogenesis 16 state model were compared. To determine similarity between chromatin states, the Jaccard index was used. This calculates a similarity co-efficient between 0 and 1, with higher values indicating more similarity.

The Jaccard index is calculated as follows:

$$J(A,B) = \frac{|A \cap B|}{|A \cup B|} = \frac{|A \cap B|}{|A| + |B| - |A \cap B|}$$



J = Intersection between A and B
Union

All possible pairwise comparisons between the 98 Roadmap cell types and hMSCs and chondrocytes were performed for the eight equivalent chromatin states. The BEDtools Jaccard tool was used to calculate similarity coefficients. GNU Parallel was used to parallelise the process. Scripts from http://quinlanlab.org (lead developer for BEDtools) were adapted for this analysis. The princomp() function in R was used for principal component analysis (PCA).

2.3.16 Correlation of DNA methylation to chromatin states

Chromatin states for CpG sites in the 450K methylation array were found using BEDtools intersect. UCSC liftover was used to convert hg19 probe genome co-ordinates to hg38. Of the total 485513 probes in the 450K array, 485438 (99.99%) lifted over successfully including all significant CpG sites (q < 0.05 and 10% Δβ). Jaccard statistics was performed for significant CpGs and all CpGs in chondrogenesis chromatin states using BEDtools. The average length of each chondrogenesis chromatin state was found by dividing the total size by the number of states.

2.3.17 Generation of plots

Chromatin state plots were generated by ChromHMM. DNA methylation plots were generated using built in function in minfi. Violin and PCA plots were created using the ggplot2 package in RStudio. Correlation heatmaps with hierarchical clustering based on Euclidean distance were constructed from a matrix of Jaccard index values using the pheatmap() function; all heatmaps were drawn using the same scale. Pie charts and bar plots were created using Microsoft Excel. Data wrangling was performed using R and Shell languages.

2.3.18 Analysis of SOX9 and JUN ChIP-seq datasets

Mouse rib chondrocyte SOX9 and JUN ChIP-seq data were downloaded from the GEO database (accession GSE69109 and GSE73372 respectively). Fastq files were converted from sequence read archive (SRA) files using the fastq-dump tool from the SRA toolkit. Both datasets were aligned to mouse reference genome mm10 (downloaded from UCSC) using Bowtie2 (default settings) after generating a Bowtie2 index. Aligned reads were converted into hg38 coordinates using the UCSC liftover tool. Peaks were called using MACS2 peak caller using input samples as a control with a *q* value cutoff of 0.05 and the effective hg38 genome size given as $3.05\text{x}10^9$. *De novo* motif discovery was performed using the top 500 significant peaks in each peakset using the MEME ChIP tool (default settings); the TOMTOM utility within the MEME suite was used to compare discovered motifs to known human transcription factor binding motifs. An E-value (FDR) < 0.05 was considered to be significant.

2.3.19 Identification of super enhancers

Super enhancers in chondrocytes were identified in our dataset based on the method described in Pott and Lieb, 2014. Regions identified as strong enhancers (13_EnhS; chromatin states computed by ChromHMM) in differentiated chondrocytes were used. We stipulated that strong enhancers states in our chondrocytes must also be an enhancer state in Roadmap E049 chondrocytes (9_EnhA1 or 10_EnhA2). Strong enhancer states were stitched together if they were less than 12.5kb apart. Significant SOX9 and JUN peaks (*q* value > 0.05) were intersected with stitched enhancers and enhancers were ranked by SOX9 signal value (from MACS2 peak caller). A SOX9 signal value cutoff was defined by finding the point of the curve where the slope = 1. Enhancers with SOX9 peaks above this value were designated as super enhancers. Super enhancers must have an overlapping SOX9 and JUN peak. HOMER annotatepeaks.pl was used to associate super enhancers to the nearest gene. The GREAT GO ontology tool was used to retrieve GO terms for super enhancers.

Chapter 3. Histone ChIP-seq quality control, alignment and peak calling

3.1 Introduction

Histone modifications regulate gene transcription by recruiting chromatin remodellers to restructure chromatin to alter the accessibility of genes to transcription factors. Regulatory elements of genes may be marked by active or repressive histone modifications. Histone ChIP-seq is a high-throughput method of assaying genome wide histone modifications. A principal aim of this project was to generate a histone ChIP-seq dataset for the transwell *in vitro* model of chondrogenesis. We aimed to elucidate epigenomic changes important for chondrogenesis by investigating histone modifications at day 0 and day 14 of differentiation.

We selected five histone modifications to assay in our ChIP-seq experiment. These were H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3. These were selected to offer a broad range of representation of regulatory elements in the genome. H3K4me3 marks are enriched in active gene promoters (Guenther et al, 2007). H3K4me1 and H3K27ac both mark gene enhancers although H3K27ac is more indicative of an active enhancer (Creyghton et al, 2010) whereas high H3K4me1 levels denotes poised enhancers (Heinz et al, 2015). H3K27me3 is typically a repressive mark and is found within transcriptionally inactive genes and heterochromatin (Boros et al, 2014) athough a study found that H3K27me3 may be associated with active transcription in some genes (Young et al, 2011). H3K27me3 may also be present alongside H3K4me3 in gene promoters, which signifies a bivalent state. H3K36me3 modifications within genes are associated with active transcription (Vakoc et al, 2006). H3K36me3 can mark both introns and exons although at lower levels in introns and alternatively spliced exons, which suggests this histone mark could have a role in the regulation of splicing (Kolasinska-Zwierz et al, 2009). These five histone modifications were also included in ChIP-seq experiments in the ENCODE (Dunham et al, 2012) and Epigenomics Roadmap projects. H3K4me3, H3K4me1, H3K27me3 and H3K36me3 were included as core marks in the Epigenomics Roadmap project and H3K27ac was an additional mark in some datasets (Kundaje et al, 2015).

The inclusion of a negative control is highly important to ChIP-seq experiments in order to assess protein or histone mark enrichment and the signal to noise ratio. Options for ChIP-seq controls are a mock IgG control, input control and for histone ChIP-seq, a H3 antibody control may also be used. The IgG control involves using an IgG antibody to serve as a mock immunoprecipitation with the DNA fragments pulled down used as the background noise. An input control is the DNA used in the immunoprecipitation experiment but is not immunoprecipitated itself so it represents non-enriched DNA fragments. Although an IgG control mimics more closely an immunoprecipitation experiment, it can be difficult to recover enough DNA for sequencing and the limited material recovered may lead to PCR bias during the library preparation step (Kidder et al, 2011). Furthermore, input controls provide a greater coverage of noise over the genome (Kidder et al, 2011) and peak calling algorithms are usually designed with assumptions appropriate with using an input control (Boer et al, 2014). Therefore, input controls are preferred over IgG controls for ChIP-seq. For histone ChIP-seq a general H3 antibody can be used to generate the background noise. A study found that there was no difference in the use of an H3 control or an input control (Flensburg et al, 2014). In this project, an input control was generated.

The ENCODE project has published a set of guidelines for generating and assessing optimal ChIP-seq data (Landt et al, 2012). These guidelines were considered during the design and optimisation of the chondrogenesis ChIP-seq workflow. The differentiated chondrocytes form a cartilage-like disc in the transwell model of chondrogenesis. One of the challenges involved was the extraction of cells from the ECM-dense disc. Before ChIP-seq was attempted, the protocol was optimised using ChIP coupled with quantitative PCR (ChIP-qPCR). After histone ChIP-seq, quality control (QC) of the samples was performed and a range of quality metrics was calculated. These metrics were part of ENCODE guidelines for assessing the quality of ChIP-seq data. For each sample, we calculated the PCR bottleneck coefficient (PBC), normalised strand coefficient (NSC), relative strand coefficient (RSC) and the fraction of reads in peaks (FRiP). The PBC measures library complexity and whether duplicate reads are overrepresented. RSC, NSC and FRiP are measures of enrichment.

3.2 Aims:

- Differentiate hMSCs into chondrocytes over 14 days using the transwell model of chondrogenesis. Optimise chondrocyte isolation and chromatin extraction from cartilage discs.

- Assess histone mark enrichment for the *in vitro* model of chondrogenesis (hMSC at day 0 vs chondrocytes at day 14) using ChIP-qPCR.

- Generate a histone ChIP-seq dataset for the same model using antibodies against histone marks H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3 including an input control, and assess the quality of the data and reproducibility of biological replicates.

- Call peaks and annotate peaks in the ChIP-seq dataset.

3.3 Results

3.3.1 Sonication of extracted hMSC and chondrocyte chromatin

hMSCs were differentiated into chondrocytes in transwell inserts over 14 days and chromatin extracted at days 0 (hMSCs) and day 14 (differentiated chondrocytes). Chromatin was extracted from hMSCs and differentiated chondrocytes using Diagenode's iDeal histone ChIP-seq kit. Differentiated chondrocytes were first isolated from the cartilage-like disc formed over the chondrogenesis process using enzyme digestion before chromatin was extracted. After crosslinking with formaldehyde, the chromatin was sonicated. Sonication of chromatin is an important step in the ChIP protocol. Adequate fragmentation of chromatin is required for a high resolution of protein enrichment. Diagenode recommended an ideal average fragment size of 100-600bp for histone ChIP-seq. Sonication cycles were optimised for hMSC and chondrocyte chromatin using a Biorupter Standard (Diagenode). Chromatin from hMSCs was sonicated with a range of cycle numbers and the DNA fragment size (after reverse crosslinking) was visualised using agarose gel electrophoresis (Fig. 3.1A).

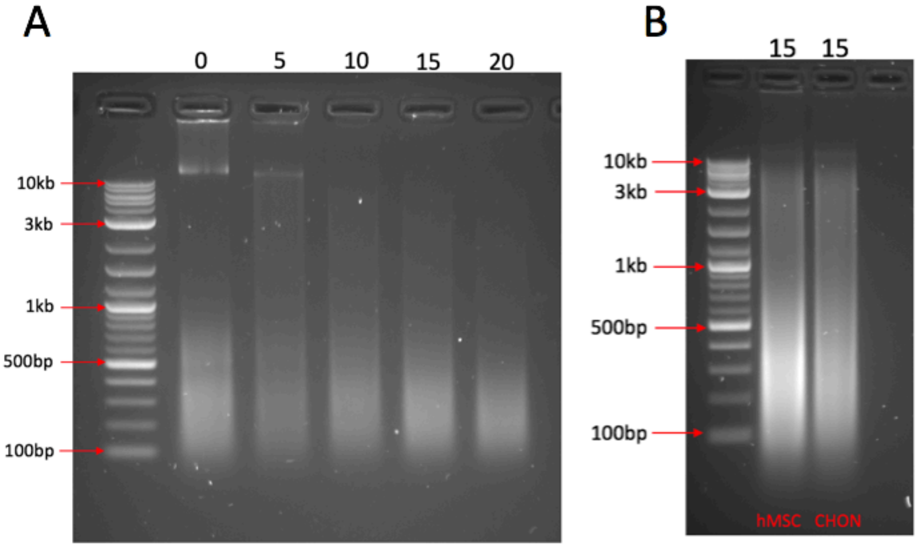


Figure 3.1 – (A) Sonication of hMSC chromatin using 0, 5, 10, 15 and 20 cycles (Diagenode standard; 30s on/20s off. (B) DNA fragment sizes from chromatin extracted from hMSCs and isolated differentiated chondrocytes sonicated at 15 cycles.

We observed that as the cycle number increased, the average DNA fragment size decreased. DNA fragments without sonication and at 5 and 10 sonication cycles were too large for ChIP as bands exceeding 10kb were present. We observed that 15 sonication cycles (30s on/20s off; Diagenode Standard) gave a DNA fragment size range of 150-800bp with the majority of the DNA at ~250bp. Chromatin extracted from differentiated chondrocytes was also sonicated for 15 cycles and this gave similar sizes to hMSC with an average of ~250bp (Fig. 3.1B). Whilst chromatin sonicated for 20 cycles also yielded fragment sizes of ~150-600bp, over sonication leads to reduced ChIP efficiency (Pschelintsev et al, 2016) so the lowest sonication cycle with the same size range was chosen. Thus 15 cycles were decided to be optimal.

For each ChIP-seq replicate experiment, chromatin from hMSC and differentiated chondrocytes was extracted and sonicated at the same time. Immunoprecipitations using different histone modification antibodies were carried out using chromatin from the same batch to minimise technical variation. Similarly, the input controls were from the same chromatin used for immunoprecipitations.

3.3.2 ChIP-qPCR

3.3.2.1 H3K4me3 ChIP-qPCR

ChIP-seq grade antibodies and histone mark enrichment of hMSCs and differentiated chondrocytes were assessed using ChIP-qPCR. The H3K4me3 antibody was tested using ChIP-qPCR prior to performing experiments for sequencing. H3K4me3 is found in the promoters of actively transcribed genes. Enrichment for H3K4me3 marks were assayed using primers designed to promoter regions of actively expressed genes in hMSCs and chondrocytes (Fig. 3.2). Gene expression abundance in transcripts per million (TPM) in RNA-seq data of the same *in vitro* model of chondrogenesis was used as a reference for expression levels. Genes that change during chondrogenesis were selected to elucidate whether the change in H3K4me3 enrichment was associated with gene expression change.

A positive enrichment was defined as a percent of input greater than 5% (Diagenode recommended threshold). Enrichment was normalised using an input control because at this stage we had determined that future ChIP-seq experiments would include an input control. However, IgG immunoprecipitations were included in initial ChIP-qPCR experiments and the percent of input of IgG calculated as an extra negative control.

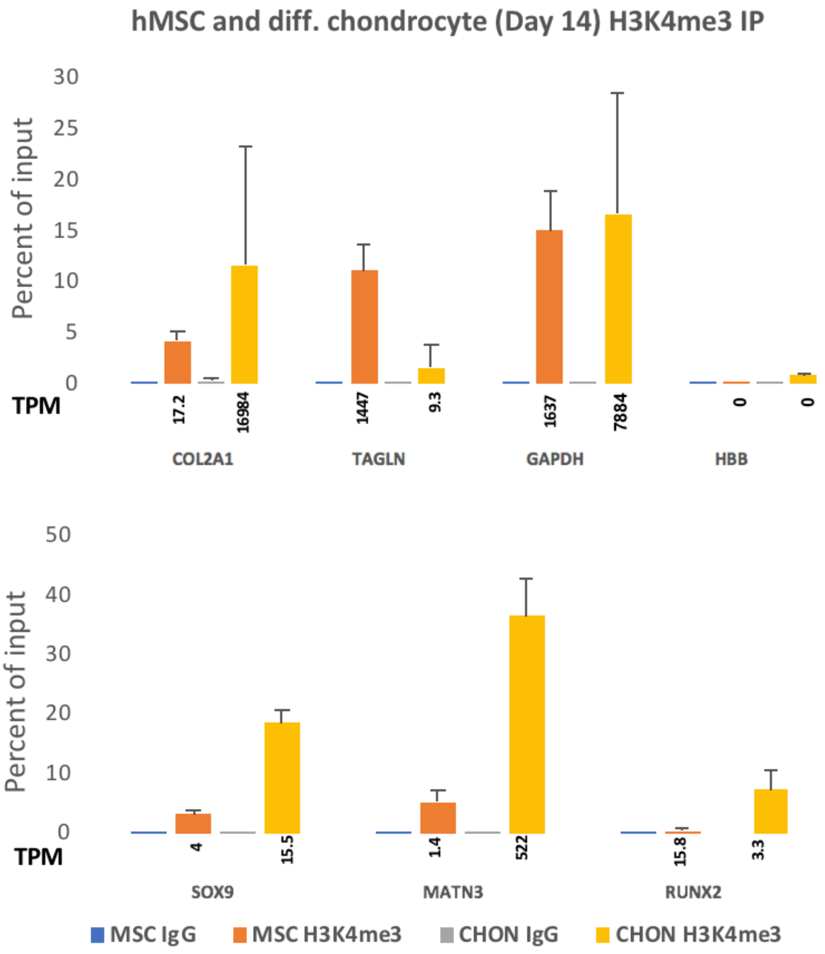


Figure 3.2 – H3K4me3 enrichment for hMSCs and day 14 differentiated chondrocytes (labelled CHON) at gene promoters. H3K4me3 enrichment at promoters of genes COL2A1, TAGLN, GAPDH, HBB, SOX9, MATN3 and RUNX2 were assayed using ChIP-qPCR (technical replicates n = 3). A positive enrichment was defined as a percent of input above 5%, an IgG control was included but this was not used in the final analysis. Error bars represent the standard deviation. TPM of genes assayed by RNA-seq are shown below the bars.

Increased enrichment of H3K4me3 at gene promoters was seen for genes that were upregulated from day 0 to day 14 of chondrogenesis based on TPM. *COL2A1*, *SOX9* and *MATN3* showed both increased gene expression and promoter H3K4me3 enrichment in chondrocytes compared to hMSCs. The opposite was seen for *TAGLN*, a gene which is down regulated during chondrogenesis. However, *RUNX2* which is a marker of osteoblastogenesis and is downregulated during chondrogenesis, shows an enrichment of H3K4me3 at the gene promoter that is absent in hMSCs. *GAPDH* was used as a positive control gene and H3K4me3 enrichment was observed as expected in both hMSCs and chondrocytes. As a housekeeping gene, expression of *GAPDH* was not expected to change during chondrogenesis but it appears to increase in chondrocytes. It is unknown whether this change is significant because differential testing could not be performed for RNA-seq samples due to low sample size (n = 1). However, chondrogenesis microarray data (kindly provided by Dr Matt Barter, Newcastle University) confirmed that GAPDH is upregulated (log2 fold change 1.4; *q* value < 2.37×10^{-10}) in differentiated chondrocytes compared to hMSCs. Nonetheless, GAPDH remains highly expressed in both cell types and this is reflected in the enrichment of H3K4me3. The negative control gene, *HBB*, was neither expressed nor was enriched for H3K4me3 in hMSCs and differentiated chondrocytes. These initial findings show that upregulation of gene expression during chondrogenesis is accompanied with an increase in H3K4me3 enrichment at promoters.

3.3.2.2 H3K4me1 and H3K27ac ChIP-qPCR

Designing primers for ChIP-qPCR to assay other histone marks proved challenging. H3K4me3 marks active promoters and was therefore a simple matter of designing primers upstream (0-500bp) of the TSS of genes. Other histone marks we were interested in included the poised and active enhancer marks H3K4me1 and H3K27ac respectively, active gene body mark H3K36me3 and repressive mark H3K27me3. Enhancers can be located long distances from genes and gene body modifications can be found anywhere within the gene body. For non-promoter histone modifications, we were limited to genomic regions that were known to be marked by specific histone marks. For this reason, gene body modifications were not assayed using ChIP-qPCR and enhancer modifications were only assessed in a known enhancer. H3K4me1 and H3K27ac histone modifications in the *COL2A1* enhancer located within the first intron of the *COL2A1* gene (Krebsbach et al, 1996) was assayed using ChIP-qPCR. In MSCs, there was a marginal positive enrichment of the poised enhancer mark H3K4me1 and no enrichment of the active enhancer mark H3K27ac. Both modifications are positively enriched in chondrocytes, defined as a percentage of input greater than 5% (Fig 3.3).

Overall, for the genes that were assayed, a change in gene expression occurred with the expected change in H3K4me3 enrichment at the gene promoters. Both enhancer marks were enriched in a known intronic *COL2A1* enhancer in chondrocytes whereas only the poised mark H3K4me1 was enriched in hMSCs, illustrating an activation of this enhancer in chondrogenesis.

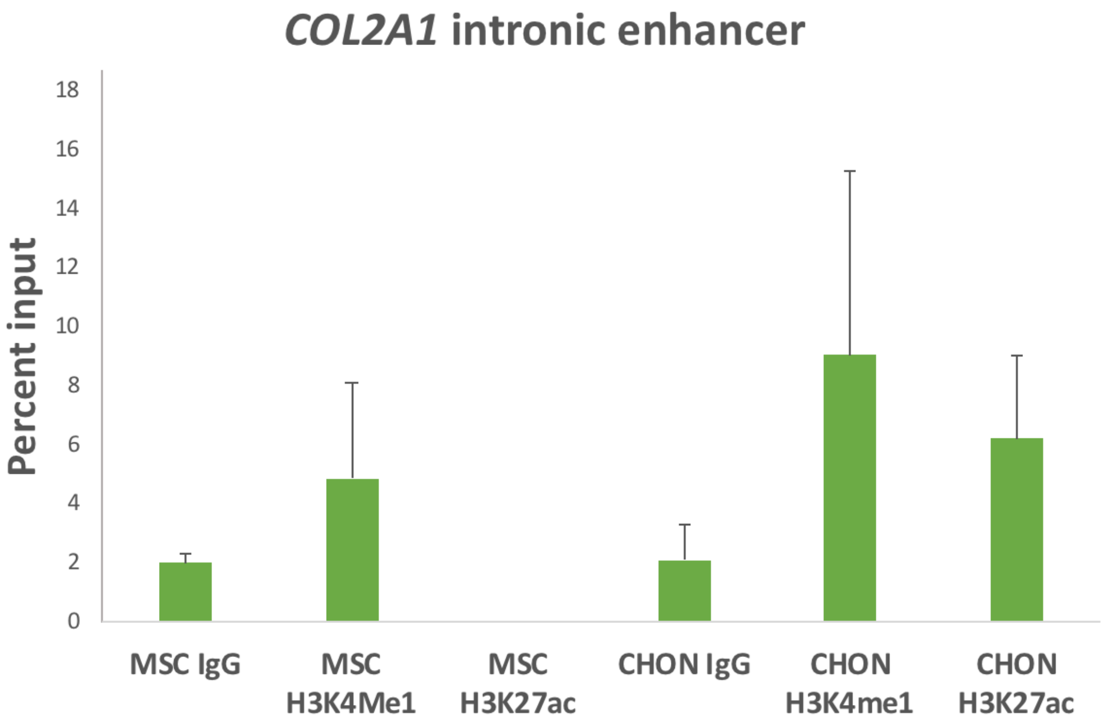


Figure 3.3 – Enrichment of H3K4me1 and H3K27ac histone modifications within the COL2A1 intronic gene enhancer in hMSCs and chondrocytes (CHON) assayed using ChIP-qPCR (n = 3). A positive enrichment was defined as a percent of input greater than 5%, an IgG control was included but not used in the final analysis. Error bars represent the standard deviation.

3.3.3 Read QC and ChIP-seq quality metrics

After optimisation of the ChIP protocol using ChIP-qPCR, ChIP-seq was performed for hMSC and differentiated chondrocytes (n = 2 for each cell type, hMSC donors 071508A and 2454e) using antibodies against H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3. Input controls also were sequenced for each replicate and cell type. The Diagenode iDEAL histone ChIP-seq kit was used. Library preparation and high throughput sequencing was performed externally. ChIP-seq replicate 1 was performed by Diagenode and replicate 2 by the Genomics Core facility, Newcastle University. Replicate experiments were performed one year apart. In both cases, bioanalyzer results showed that DNA fragments were larger than expected and DNA required re-sonicating as fragment sizes after immunoprecipitation were too large for library preparation. ChIP-seq reports from Diagenode and the Genomics Core facility are available at https://github.com/kathleencheung/PhD_Young_lab

Sequencing reads from each ChIP-seq experiment were assessed for quality. ChIP-seq replicate 1 (donor 071508A) samples were sequenced over one lane and generated 14 fastq files (2 of which were IgG controls and were not used). ChIP-seq replicate 2 (donor 2454e) were sequenced over 4 lanes and generated 49 fastq files in total (4 per ChIP sample plus 1 re-sequence) before merging. FastQC was performed on individual read files to assess any differences between them and to determine whether there were any outliers before merging. Summarised MultiQC reports showed that all sample reads from both replicates achieved a minimum quality score above 20 (Appendix i Fig. 1 and Fig.2); therefore, it was not deemed necessary to trim reads. MultiQC also assesses GC content and duplication of reads (Appendix Table 1 and Appendix Table 2). ChIP-seq samples derived from hMSC donor 071508A displayed high read duplication (9/14 samples failed this QC) compared to donor 2454e (no failures). ChIP-seq samples from donor 071508A required additional PCR cycles after DNA library preparation in order to achieve enough DNA for sequencing. The high rate of duplication seen in samples from donor 071508A was likely due to the additional PCR cycles. In total, 1/14 fastq files from donor 071508A and 6/49 read samples from donor 2454e failed GC content QC. This was not considered a problem as GC content failures can arise from enrichment of genomic regions with high GC content in ChIP-seq data. Likewise, duplicated reads were not removed to avoid eliminating biological effects, and so QC metrics can be calculated for the data without biasing the outcomes. Further QC metrics were calculated after merging fastq files from the same sample and aligning reads to hg38 reference genome using the Bowtie2 short read aligner. Reference genome hg38 was the latest human reference genome released at the time of this project. All samples achieved raw read numbers ranging from 31-84 million reads, exceeding recommended minimum read numbers for histone modifications (Chen et al, 2012; Jung et al, 2014). Alignment rates for replicate 1 samples were variable ranging from 38.85% to over 98%. All samples from replicate 2 achieved alignment rates over 97% (Appendix Table 3).

Following alignment of reads to hg38, we assessed the complexity of the data by calculating the PBC. The PBC is a measure of DNA library complexity; values of 0-0.5 indicates severe bottlenecking, 0.5-0.8 is moderate bottlenecking, 0.8-0.9 is mild and 0.9-1 indicates no bottlenecking. All ChIP-seq samples from replicate 1 displayed severe to moderate PCR bottlenecking (PBC values between 0 and 0.8; Table 3.1) reflecting the high levels of duplicate reads uncovered by MultiQC. In contrast, replicate 2 samples displayed mild to no PCR (PBC values > 0.8) bottlenecking suggesting a diverse DNA library. Samples from ChIP-seq replicate 2 did not undergo multiple PCR cycles prior to DNA library preparation. Histone ChIP-seq data from ENCODE displayed a range of PBC values from 0.19 to 0.98 with an average PBC of 0.79. Our ChIP-seq samples displayed a range of PBC values from 0.052 to 0.97 with an average of 0.28 and 0.94 from ChIP-seq replicates 1 and 2 respectively.

To measure enrichment prior to peak calling, the NSC and RSC (measures of enrichment independent of peak calling) values were calculated using ENCODE utility phantompeakqualtools. A range of NSC and RSC values was seen in both replicates. Most NSC and RSC values were low (NSC < 1.1 and RSC < 1) but apart from H3K27me3 and H3K36me3 samples in replicate 1, sample values were higher than their respective input controls. ENCODE sample NSC values ranged from 1.01 to 1.82 and RSC values from 0.09 to 3.47. Our ChIP-seq NSC values ranged from 1 to 1.3 and RSC from 0.17 to 1.1 (see Appendix i Fig. 6-10 for NSC and RSC plots).

After broad peak calling using MACS2, the FRiP was calculated using the Bioconductor DiffBind package. The percentage of reads located in peaks ranged from

46% to 79%, much higher than the recommended ENCODE minimum of 1% (Landt et al, 2012). ENCODE ChIP-seq datasets do not have a FRiP value attached and instead used a metric called signal portion of tags (SPOT) which is analogous to FRiP. ENCODE's SPOT values for histone ChIP-seq datasets range from 7% to over 89%. However, our FRiP values and ENCODE's SPOT are not directly comparable due to the different peak callers and peak calling parameters used.

ENCODE dataset quality metric values were used as a reference. Both ENCODE and our histone ChIP-seq metric values were varied. Quality metrics for ENCODE samples are available on the ENCODE website and have also been uploaded here https://github.com/kathleencheung/PhD_ChIP-seq/

Table 3.1 – Quality metrics PBC, NSC, RSC and FRiP calculated for each histone ChIP-seq sample and input controls.

Replicate no.	hMSC Donor	Sample name	PBC	NSC	RSC	FRiP
1	071508A	MSC_input	0.396324225	1.026624	0.2790935	N/A
1	071508A	MSC_H3K4me3	0.589673045	1.067147	0.6845061	0.67
1	071508A	MSC_H3K4me1	0.15150683	1.032418	0.2886845	0.65
1	071508A	MSC_H3K27ac	0.740986632	1.015744	0.3682942	0.64
1	071508A	MSC_H3K27me3	0.084117163	1.124742	0.2121983	0.46
1	071508A	MSC_H3K36me3	0.052449762	1.173076	0.1646135	0.63
1	071508A	CHON_input	0.261170701	1.033358	0.2738952	N/A
1	071508A	CHON_H3K4me3	0.168367456	1.095136	0.4527105	0.65
1	071508A	CHON_H3K4me1	0.24007756	1.028448	0.376056	0.69
1	071508A	CHON_H3K27ac	0.545050931	1.020625	0.3893155	0.6
1	071508A	CHON_H3K27me3	0.054636704	1.218342	0.1764161	0.51
1	071508A	CHON_H3K36me3	0.069729458	1.128785	0.2053987	0.62
2	2454e	MSC_input	0.957925977	1.007932	0.2389265	N/A
2	2454e	MSC_H3K4me3	0.858682875	1.326625	1.114411	0.79
2	2454e	MSC_H3K4me1	0.964465229	1.031276	0.8018935	0.75
2	2454e	MSC_H3K27ac	0.945940422	1.078115	0.9714882	0.71
2	2454e	MSC_H3K27me3	0.942934438	1.012197	0.3424739	0.58
2	2454e	MSC_H3K36me3	0.935295272	1.013854	0.411228	0.72
2	2454e	CHON_input	0.954169002	1.007501	0.2390396	N/A
2	2454e	CHON_H3K4me3	0.956291968	1.110019	0.9048022	0.65
2	2454e	CHON_H3K4me1	0.942191644	1.022946	0.767487	0.72
2	2454e	CHON_H3K27ac	0.942729131	1.021243	0.7790453	0.66
2	2454e	CHON_H3K27me3	0.94870526	1.013444	0.2955627	0.56
2	2454e	CHON_H3K36me3	0.929758043	1.015522	0.3025558	0.62

Overall mapped read density in the genome for each sample was inspected using the ngs.plot tool (Fig. 3.4 – 3.6; Appendix i Fig. 3-5) before peak calling. Reads for H3K4me3 samples were generally mapped to around the TSS of genes and the highest density of reads was found slightly downstream (within 1000bp) of TSS (Fig. 3.4). This fits in line with expectations as H3K4me3 is an active promoter histone modification. Enhancer and gene body histone modification reads were located further away from the TSS. The highest density of H3K4me1 reads mapped to 2000-3000bp downstream from the TSS with another density peak seen a similar distance upstream (Fig. 3.5). H3K27ac showed the most heterogeneous read profile across samples; H3K27ac reads from hMSC and chondrocyte samples from donor 071508A displayed a read density profile comparable to H3K4me1 whereas 2454e samples were more alike H3K4me3. The highest read densities for gene body marks H3K27me3 and H3K36me3 were seen around 3000-6000bp away from the TSS (Fig. 3.6). Globally, for each histone modification, reads mapped to expected regions of the genome relative to the TSS of genes.

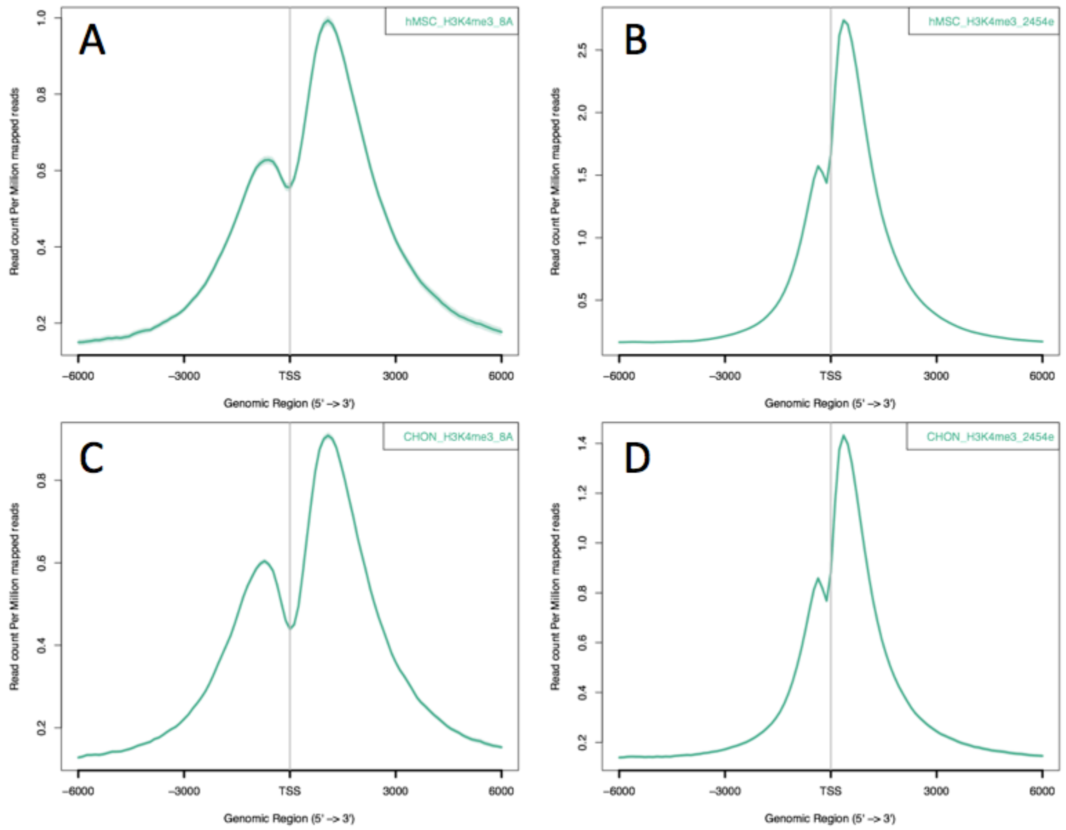


Figure 3.4 –Average read densities for active promoter mark H3K4me3 in both hMSC (A&B) and chondrocyte (C&D) replicates (A&C = donor 017508A; B&D = donor 2454e). Shading represents the 95% confidence intervals.

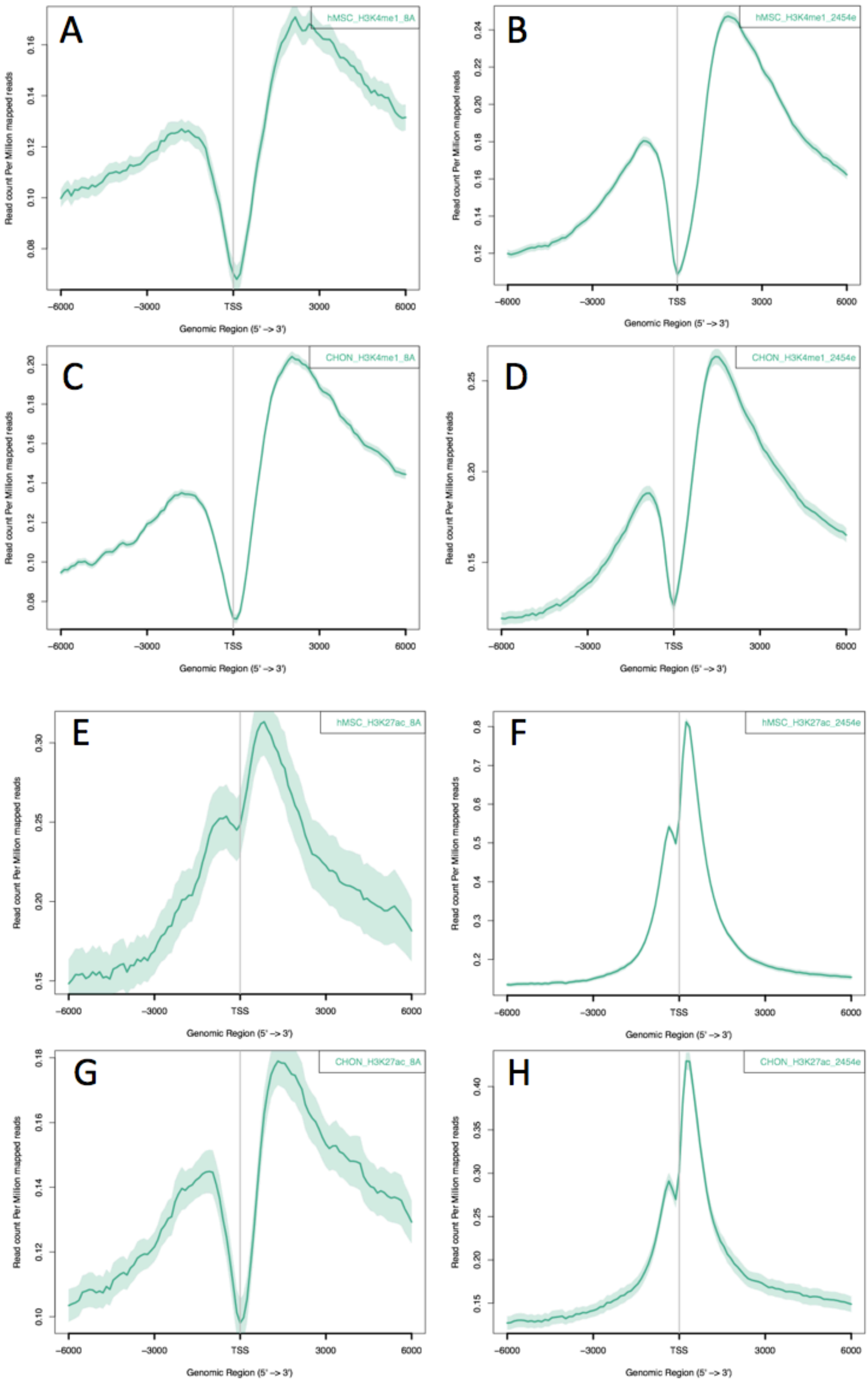


Figure 3.5 – Average read densities for enhancer marks H3K4me1 (A-D) and H3K27ac (E-H) in both hMSC (A&B, E&F) and chondrocyte (C&D, G&H) replicates. Shading represents the 95% confidence intervals.

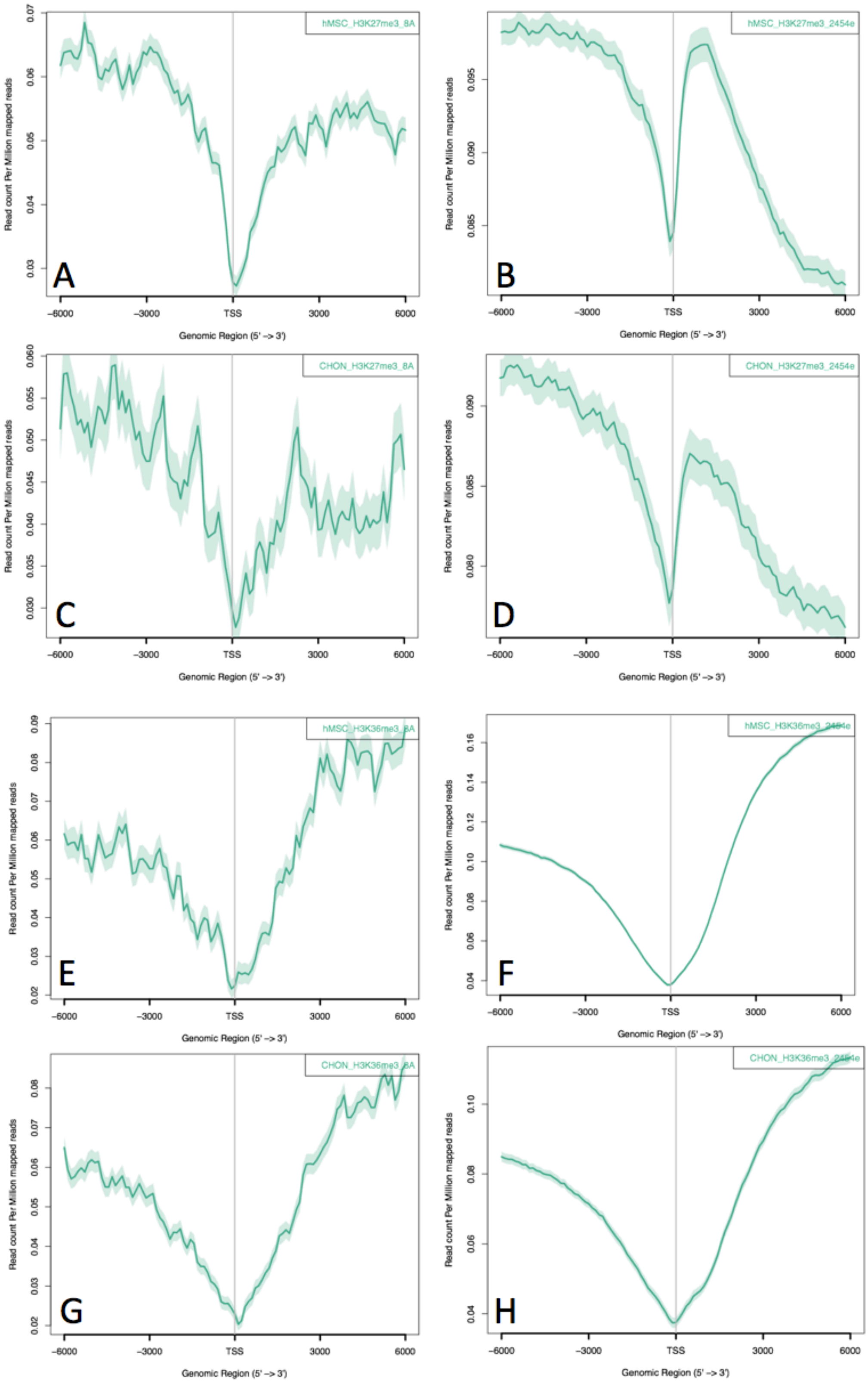


Figure 3.6 – Average read densities for repressive mark H3K27me3 (A-D) and active gene body H3K36me3 (E-H) in both hMSC (A&B, E&F) and chondrocyte (C&D, G&H) replicates. Shading represents the 95% confidence intervals.

In addition to calculating quality metrics recommended by ENCODE, read coverages were also visually inspected using IGV genome browser after normalising by reads per million mapped reads (RPM). As examples of actively expressed genes, ChIP-seq reads around CD44 (a marker of MSCs) in hMSCs (Fig. 3.5) and COL2A1 in chondrocytes are shown (Fig. 3.6). In general, read coverages for each histone modification from both ChIP-seq replicates occur around the same region in the genome. For instance, actively expressed genes tended to exhibit H3K4me3 marks around their TSS and H3K36me3 marks in the gene body whereas non expressed genes did not. Enhancer marks H3K4me1 and H3K27ac reads tended to be seen in the same regions whereas repressive H3K27me3 and active H3K36me3 did not overlap. Although not a quantitative assessment, visualisation on a genome browser proved to be a useful and intuitive method of initially surveying the data before moving on to further analyses.

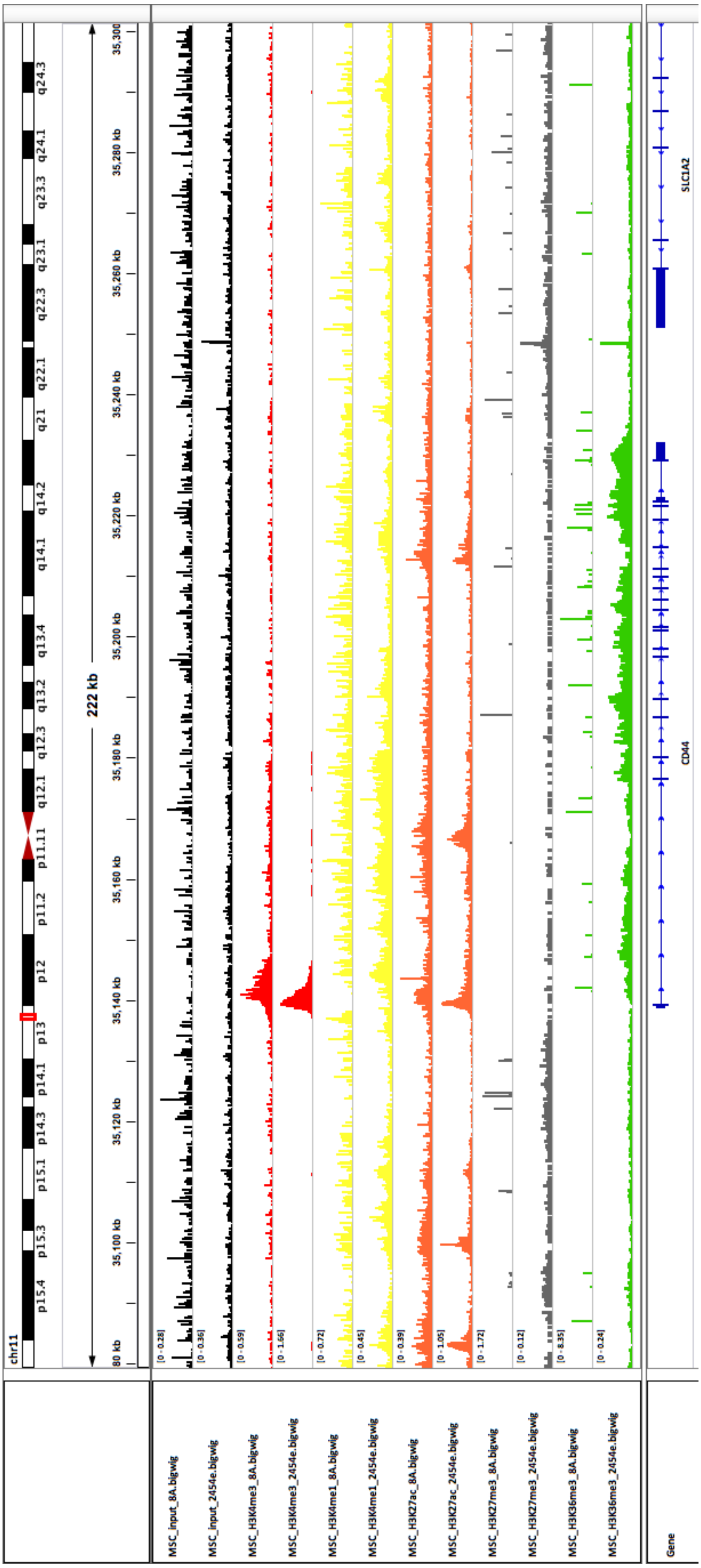


Figure 3.7 – IGV screenshot of read coverages of all hMSC ChIP-seq samples in both replicates around the CD 44 gene.

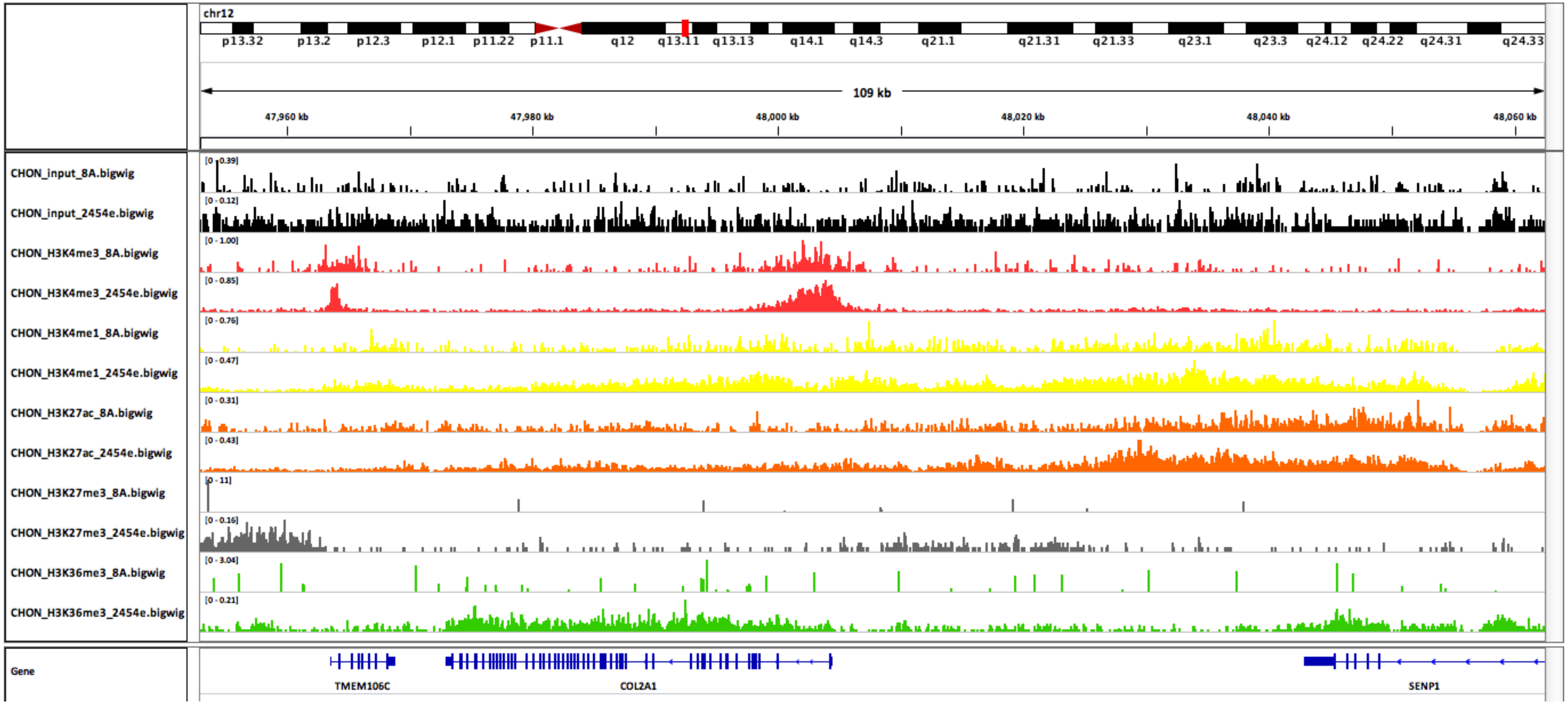


Figure 3.8 – IGV screenshot of read coverages of all chondrocyte ChIP-seq samples in both replicates around the COL2A1 gene.

3.3.4 Peak calling and annotation

Peak calling identifies regions of the genome that are statistically enriched for the protein or histone modification assayed over a background control. The MACS2 peak caller was used for its option to call broad peaks, the type of peaks typically produced by histone modifications. To call peaks, alignment files in BAM format were converted into bedgraph format using the bedtools genomecov tool and the bedgraph files were used as input into MACS2 peak caller. Significant broad peaks were called using MACS2 peak caller (*q* value < 0.05) using the input samples as a control. Peaks were called separately for each ChIP-seq replicate. Numbers of peaks called for each histone mark were variable between cell types and donors (Table 3.2). The number of H3K4me3 peaks in hMSC cells from donor 071508A was roughly half of the other samples (33161 compared to 61348 – 69215). H3K4me1 peak numbers were comparable between cell types from the replicates. There were more H3K27ac peaks called for both hMSCs and chondrocytes from donor 2454e (60465 and 67590) compared to 071508A (147461 and 110775). Numbers of H3K27me3 peaks were variable between all four samples. More peaks were called for H3K36me3 in hMSCs from donor 071508A compared to chondrocytes from the same donor; in contrast, donor 2454e showed the opposite trend.

Table 3.2 – Total number of peaks called by MACS2 (q < 0.05, broad option on, input control) for each ChIP-seq sample.

			Number of peaks called by MACS2				
Replicate no.	**hMSC donor**	**Cell type**	**H3K4me3**	**H3K4me1**	**H3K27ac**	**H3K27me3**	**H3K36me3**
1	071508A	hMSC	33161	163717	60465	110041	138557
1	071508A	CHON	69215	145745	67590	85395	96321
2	2454e	hMSC	61348	173052	147461	151273	95744
2	2454e	CHON	69094	149400	110775	234574	131261

To view genome coverage of called peaks, peak density plots were created (Fig. 3.9). Prior to peak calling, read coverage for each histone mark was plotted and visualised on the IGV genome browser (Fig. 3.3 – 3.7). However, these analyses do not empirically take into account the input controls and may not be a true representation of histone mark enrichment. For example, high sample read density in a region does not necessarily equate to enrichment if the input sample also exhibits many reads in that same region. MACS2 uses a Poisson distribution to compare the local background (input control) enrichment levels with ChIP-seq samples. Consequently, plotting peak density, as opposed to read density, gives a better insight into where specific histone modification enrichments are located in the genome. In a high quality ChIP-seq dataset, profiles of read coverage density and peak density will be very much alike. For our dataset, plots of read coverage and peak density around the TSS of genes (-/+ 6kb) are similar. The highest density of H3K4me3 peaks are located around the TSS (Fig. 3.9A). In comparison to H3K4me3, all other histone mark peaks are at a lower density at gene TSSs. H3K4me1 peaks are at their highest density ~2-3kb downstream of TSS. The highest density of H3K27ac peaks are close to the TSS of genes but they are at a lower density compared to H3K4me3 peaks. In contrast, H3K27ac peaks are at higher densities than H3K4me3 further away from the TSS. Both gene body marks H3K27me3 and H3K36me3 peaks are distal to the TSS and showed a higher density away from the TSS compared to the promoter mark H3K4me3. The signal to noise ratio of reads was also evaluated independently of peak calling (Fig. 3.10), this corroborated the enrichment of peaks around the TSS.

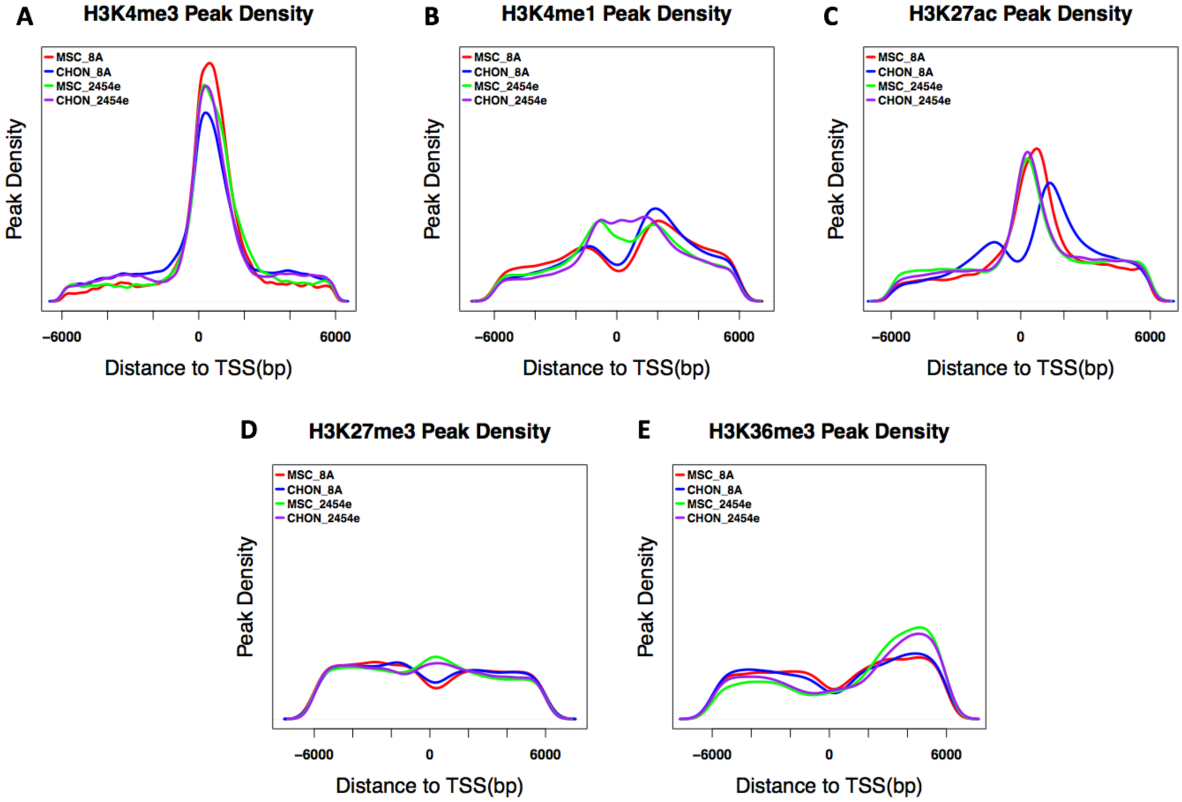


Figure 3.9 – Histone mark peak densities at TSS -/+ 6000bp. Broad peaks were called using MACS2 and density plots created in RStudio. All plots are drawn with the same x and y scales. (A) H3K4me3 peak densities, (B) H3K4me1, (C) H3K27ac, (D) H3K27me3 and (E) H3K36me3.

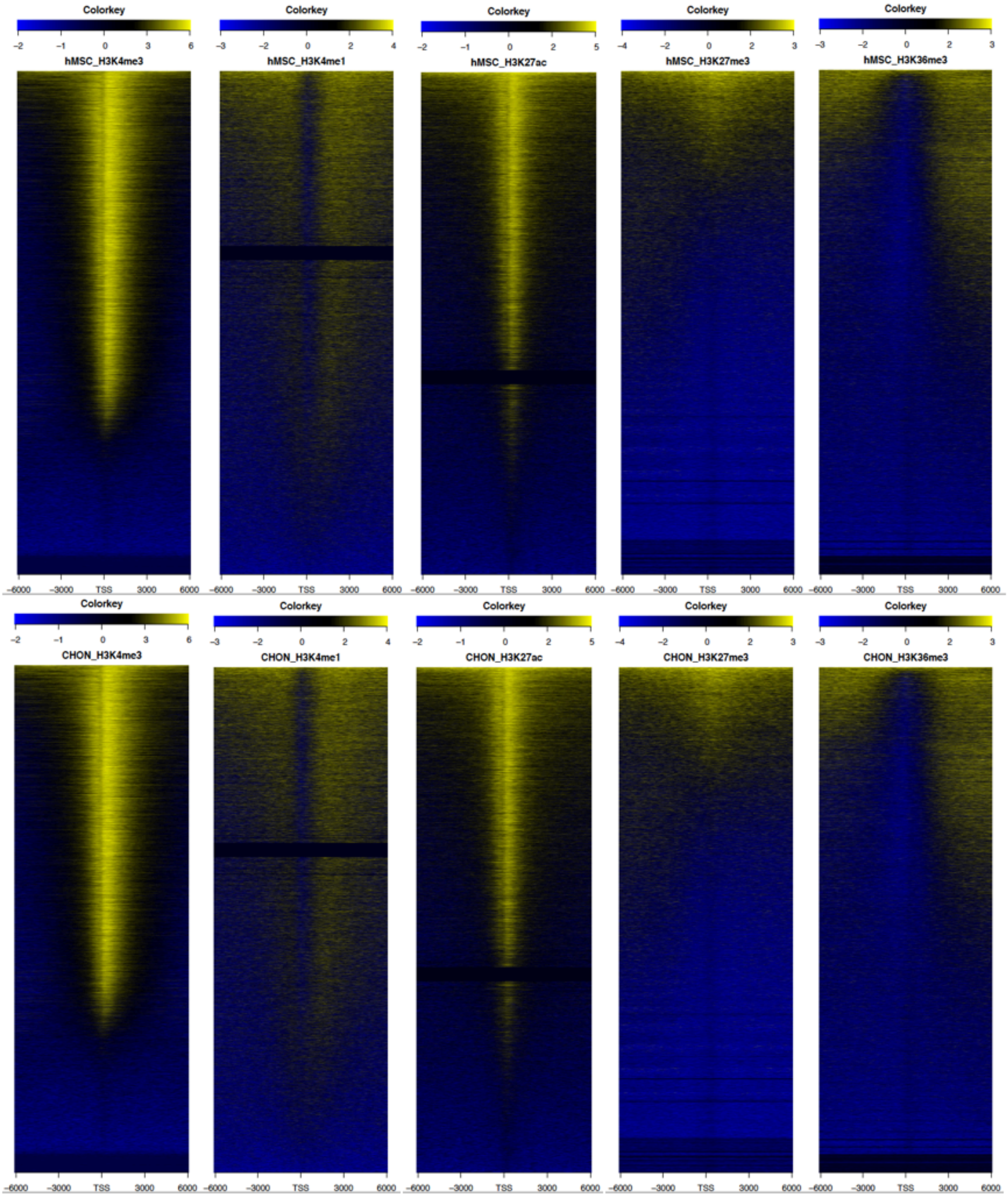


Figure 3.10 – Signal over noise ratio of histone mark reads from combined replicate samples. The ratio of histone mark reads to input controls was assessed independently of peak calling. Alignment files for each histone mark sample were merged and plots were generated using ngs.plot.r

Peaks were annotated by associating to the nearest gene and overlapping genomic feature using HOMER software. Features available are 5' untranslated regions (5'UTR), 3'UTRs, exons, introns, intergenic regions, promoters, transcriptional termination sites (TTS) and non-coding exons. Overlapping genomic features confirms that H3K4me3 peaks are found at gene start sites, with enrichment at 5'UTRs, promoters, introns and intergenic regions (Table 3.3). H3K4me3 peak numbers were relatively low at 5'UTRs (1.11%) and promoters (8.53%) compared to introns (47.1%) and intergenic regions (37%). Although comparatively low, percent of H3K4me3 peaks in these 5'UTR and promoter regions is higher than the other histone marks. The smaller percentage of peaks in these areas may be due to the smaller size of these regions relative to others. 5'UTR regions cover around 100-200bp (Mignone et al, 2002) and HOMER defines promoter regions as -1kb to +100bp from the TSS. Introns and intergenic regions can span much larger areas and therefore it is likely more peaks will fall into these regions. All histone marks showed high percentages of peaks in introns and intergenic regions. Although 47.1% of H3K4me3 are located in introns, a high proportion of these, 51.7%, are found in the first intron. Similarly, 2.71% of H3K4me3 peaks are found in exons with 42.2% of these peaks in the first exons. This adds support that H3K4me3 peaks are mainly found around the 5' end of genes. The other histone mark peaks do not display such a bias towards the first exons and introns. The active gene body modification H3K36me3 had the lowest percentage of peaks in the first intron and exon compared to the other histone marks. Enhancer marks H3K4me1 and H3K27ac showed similar peak percentages in each genomic feature.

Table 3.3 – Average percentages of all peaks from both ChIP-seq replicates for each histone mark in genomic features annotated by HOMER annotatepeaks.pl. Regions classed as N/A have not been annotated as any of the available regions.

Percentage of peaks in genomic features						Percent of total genome
	H3K4me3	**H3K4me1**	**H3K27ac**	**H3K27me3**	**H3K36me3**	
5'UTR	1.11	0.182	0.404	0.064	0.041	0.08
3' UTR	0.947	1.19	1.50	0.363	1.20	0.72
exon	2.71 (of which 42.2% in exon 1)	1.34 (of which 10.4% in exon 1)	2.19 (of which 13.3% in exon 1)	0.697 (of which 15.3% in exon 1)	1.77 (of which 2.92% in exon 1)	1.09
Intron	47.1 (of which 51.7% in intron 1)	54.6 (of which 34.0% in intron 1)	60.8 (of which 36.7% in intron 1)	29.4 (of which 28.1% in intron 1)	51.4 (of which 15.5% in intron 1)	37.9
intergenic	37.0	38.5	29.4	67.14	42.7	58.05
promoter	8.53	2.18	3.46	0.818	0.534	1.02
TTS	1.55	1.45	1.77	0.701	1.36	0.91
non-coding exons	0.856	0.485	0.592	0.289	0.443	0.17
N/A	0.213	0.123	0.035	0.585	0.518	0.06

3.3.5 Correlation of ChIP-seq replicates

ChIP-seq replicates of hMSC and differentiated chondrocytes were derived from two hMSC donors 071508A and 2454e. The replicate donor ChIP experiments and sequencing was performed at different times using different DNA library preparation kits and sequencing platforms. We sought to investigate how reproducible our ChIP-seq experiments were and whether the two replicates were comparable. Using the Bioconductor package DiffBind, a correlation heatmap was generated using read count data (Fig. 3.11). All ChIP-seq samples of the same histone mark clustered together showing that samples from our two replicate experiments are comparable when considering histone mark alone. H3K27me3 samples clustered together on a separate node to the other four histone marks and are negatively correlated to enhancer marks H3K4me1 and H3K27ac. The highest correlation between histone marks was observed between H3K4me1 and H3K27ac. This indicates that H3K4me1 and H3K27ac are likely to co-occur in the genome whereas the presence of H3K27me3 marks are mutually exclusive with H3K4me1 and H3K27ac.

As well as clustering by histone mark, we would expect samples to cluster by cell type and not by replicate/hMSC donor if biological effects are greater than technical effects between the two experiments. This was not the case for H3K27me3 or H3K36me3; samples of both these marks clustered by replicate/hMSC donor before cell type suggesting effects between replicates are greater than the effects between cell type. However, this could be due to technical variation between the experiments or as a result of biological variation between the hMSC donors for these histone modifications. H3K4me3 samples did not cluster by replicate nor cell type. In contrast, both enhancer histone marks H3K4me1 and H3K27ac separated by cell type. This shows that enhancer histone marks between hMSC and chondrocytes are more dissimilar compared to other histone marks in this study, suggesting that the cell type specific effects outweigh donor variation. A PCA plot (Fig. 3.12) showed the same outcome.

The proportion of overlapping peaks for each histone mark in both ChIP-seq replicates was found using the DiffBind package. Peaks were liberally defined as overlapping if they shared at least one base in common, this was the default setting in DiffBind. There was a 37.8% overlap of total H3K4me3 peaks from hMSC replicates and 20.3% of chondrocyte H3K4me3 peaks. H3K4me1 peaks in hMSCs saw an overlap of 20% and 32% overlapped in chondrocytes. For H3K27ac, 20.3% overlapped in hMSCs and 21.7% in chondrocytes. Only 9% of H3K27me3 peak overlapped in hMSCs, similar to chondrocytes which saw an 8.5% peak overlap. Overlap in H3K36me3 peaks between ChIP-seq replicates was also low with 11.9% and 11.8% in hMSCs and chondrocytes respectively (Fig. 3.13).

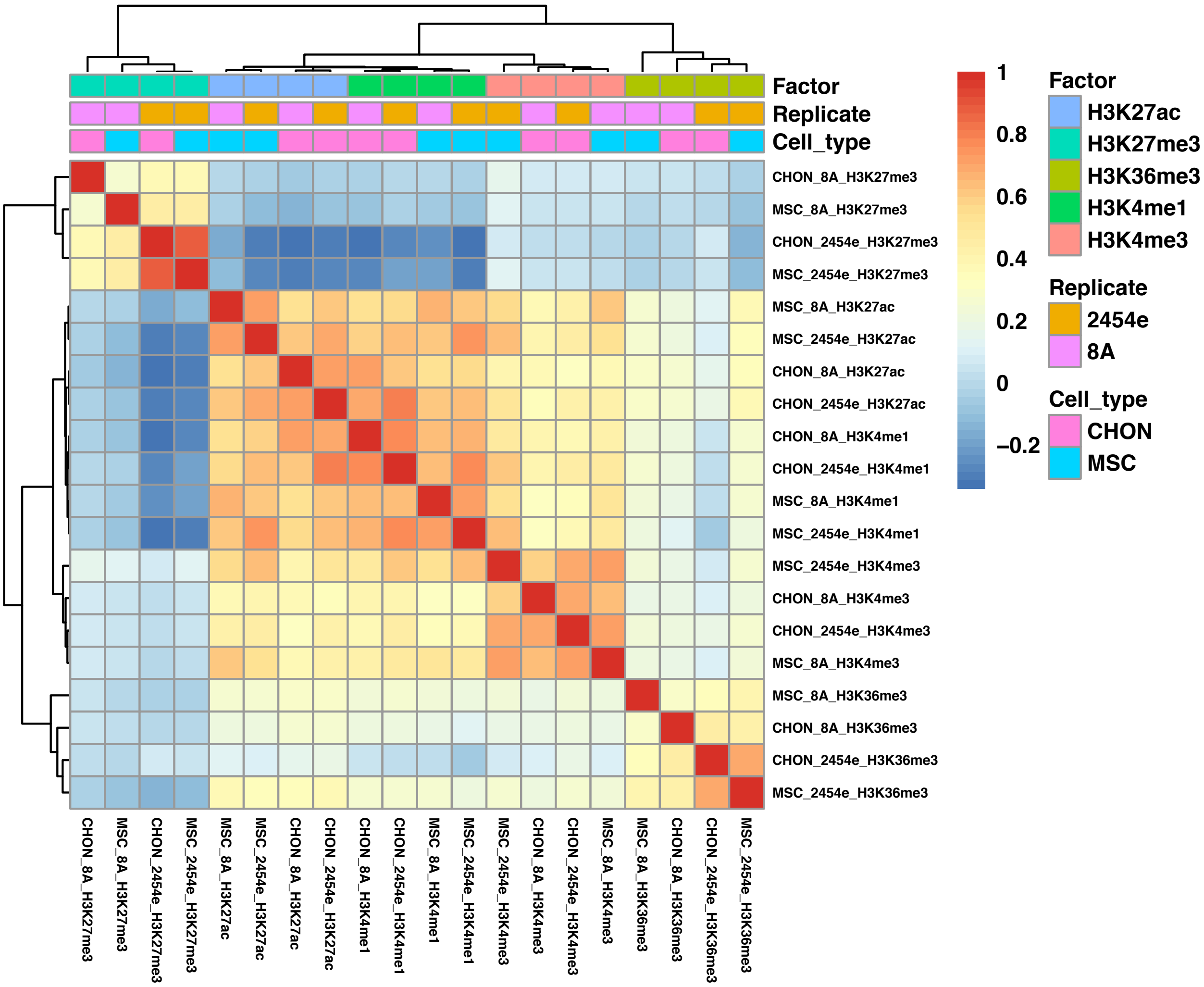


Figure 3.11 – Correlation heatmap of histone modifications across all samples. The correlation matrix was generated using DiffBind with read count data and replotted using the pheatmap() function in RStudio.

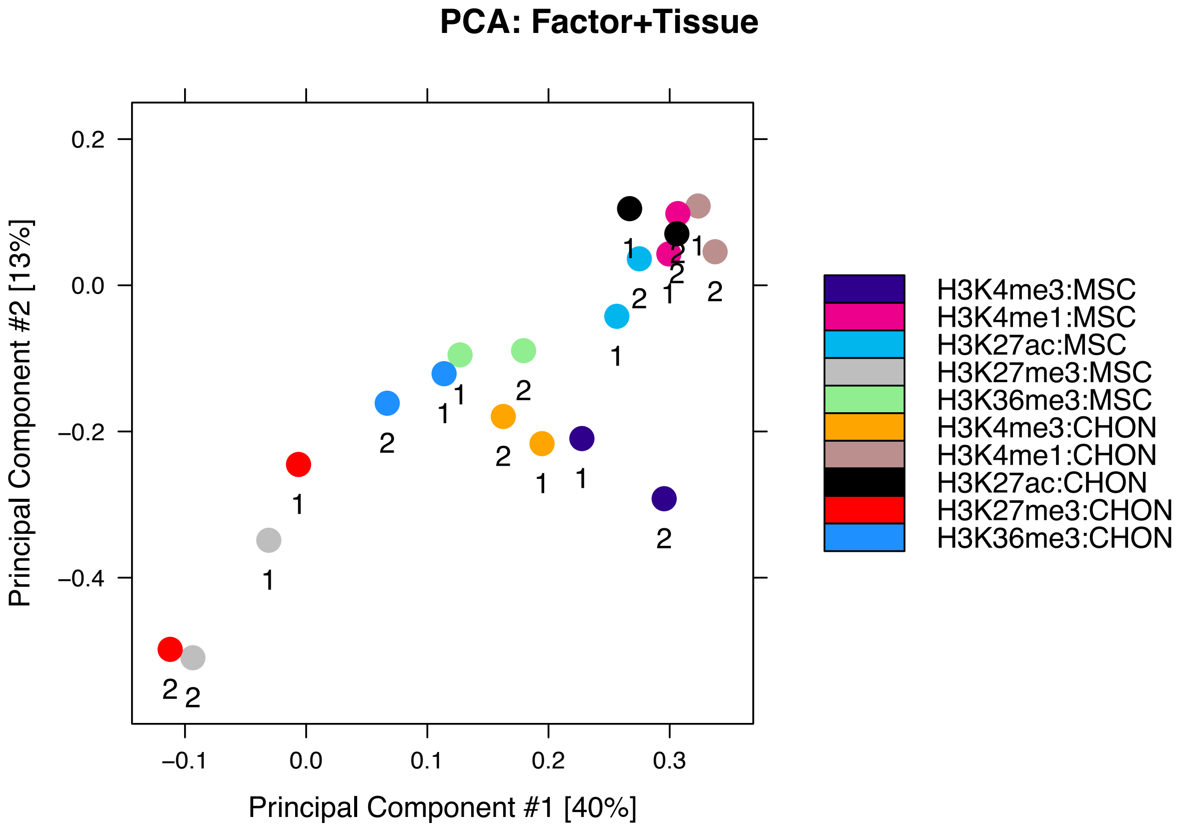


Figure 3.12 – PCA plot of ChIP-seq peak sets. Points labelled 1 are samples from replicate 1 (hMSC donor 071508A) and labelled 2 are from replicate 2 (hMSC donor 2454e).

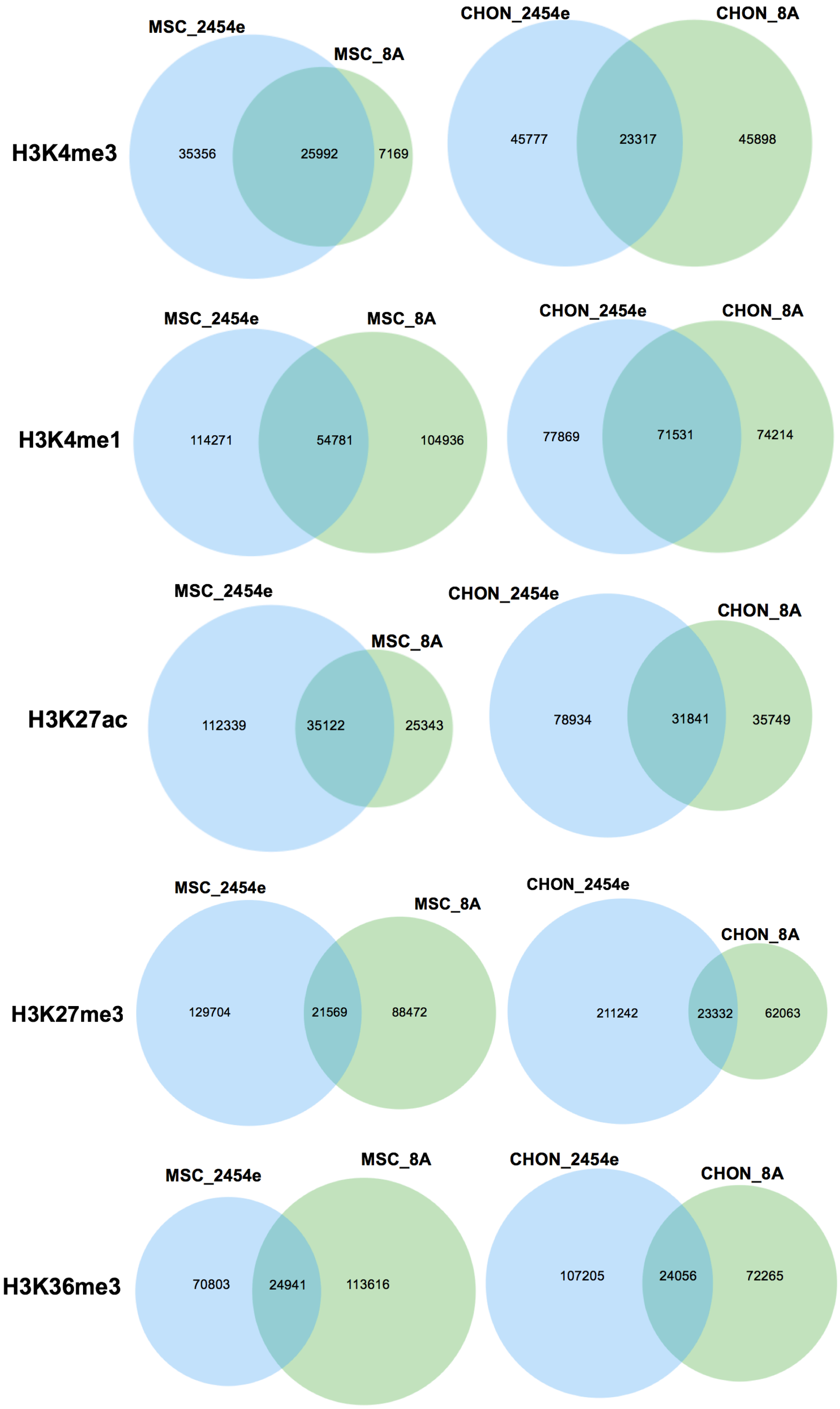


Figure 3.13 – Overlap of histone peaks between the two ChIP-seq biological replicates (hMSC donors). H3K4me3, H3K4me1 and H3K27ac displayed greater overlaps between replicates than H3K27me3 and H3K36me3. Overlapping peaks share at least one base. Overlap analysis was performed using the DiffBind package and Venn diagrams created using the Venneuler package in RStudio.

3.4 Discussion

The initial primary aim of this project was to generate histone ChIP-seq data from the transwell model of chondrogenesis. We developed a ChIP-seq workflow optimised for hMSCs and differentiated chondrocytes. At day 14 in our chondrogenesis model, cells form a cartilage-like disc (Murdoch et al, 2007). We expected the extraction of chromatin from cartilage to be difficult due to the dense extracellular matrix secreted by chondrocytes. In this project, extraction was achieved by first isolating chondrocytes by digesting the cartilage-like disc using hyaluronidase, trypsin and collagenase sequentially before extracting the chromatin. Previous chondrocyte ChIP studies also isolated cells from cartilage prior to chromatin extraction (Otero et al, 2005; Dvir-Ginzberg et al, 2008; Herlofsen et al, 2013). Care must be taken to minimize disruption to cells and to ensure a high yield of chondrocytes from the discs. There is evidence that isolation of chondrocytes from cartilage can alter gene expression. One study found that shorter digestion times of cartilage explants with collagenase led to fewer gene expression changes. The study also concluded that longer digestion yielded more cells (Hayman et al, 2006). Consequently, choosing an appropriate digestion period is a trade-off between harvesting a high yield of cells and minimising gene expression changes. As gene expression and histone modifications are linked and highly dynamic, it would be prudent to reduce any effects which could alter gene expression.

Sonication of chromatin is an important step in generating ChIP-qPCR/ChIP-seq data. Fragment sizes that are too large leads to a poor signal to noise ratio whereas over sonication may result in biases to certain regions in the genome (Diaz et al, 2012) or fragmentation of the protein bound DNA. It is particularly important to consider this when sonicating chromatin for histone ChIP experiments. DNA is wrapped around histone proteins in a unit called a nucleosome. One nucleosome is approximately 146bp (Luger et al, 1997); in order to avoid disrupting nucleosomes, chromatin should be sonicated to no smaller than 150bp. There is no single standardized chromatin fragment size for ChIP-seq and sonication generates a range of fragment sizes. Biotechnology companies providing ChIP products and services recommend 200-1000bp (ABCAM) or 100-1000bp (Diagenode). For ChIP-seq, Diagenode recommend sonicating chromatin to 100-600bp whereas ENCODE guidelines stipulate 100-300bp

(Landt et a, 2012). Smaller sizes are more appropriate for DNA library construction for high throughput sequencing. Although we sonicated chromatin to an average fragment size of ~250bp, all samples required re-sonication prior to DNA library preparation. The larger fragment size seen after immunoprecipitation is indicative of a bias towards pulling down longer fragments during immunoprecipitation. Larger fragments may contain more of the histone modification of interest and therefore be more likely to be bound by an antibody. Furthermore, sonication across the genome is not consistent, for example, heterochromatin is more resistant to sonication compared to euchromatin (Teytelman et al, 2009). Disregarding larger fragments in favour of smaller fragments may lead to a biased representation of the genome. One study found that repeat fragmentation of longer fragments after immunoprecipitation improved resolution of ChIP-seq and that the extra sonication step did not introduce any artefacts (Mokry et al, 2010). Therefore, re-shearing and including the longer immunoprecipitated fragments in the DNA library is the best approach in this scenario.

The sequencing depth of ChIP-seq experiments is another important factor to consider. Histone marks spanning broader genomic regions such as gene body and enhancer marks require more reads to achieve sufficient coverage. The ENCODE project recommends at least 10 million reads for transcription factor ChIP-seq (narrow source peaks) and 20-40+ million reads for histone ChIP-seq (Landt et al, 2012). All our ChIP-seq reads exceeded recommended numbers.

The ENCODE consortium was one of the first large scale projects using ChIP-seq and many of the early bioinformatics tools developed for ChIP-seq analysis was created by or in collaboration with ENCODE. The initial analysis and QC steps illustrated in this chapter largely follows the ENCODE ChIP-seq analysis pipeline. Various metrics attempt to assess the technical quality of ChIP-seq data and whether the data is of high enough quality for downstream analysis. The ENCODE project defined a number of quality metrics for ChIP-seq data and recommends that all ChIP-seq data is assessed using these metrics. ChIP-seq data can be assessed by QC metrics such as the PBC, NSC, RSC and FRiP. The PBC is a measure of DNA library complexity and ranges from 0 to 1, with increasing values indicating higher complexity. Diverse DNA libraries will have a high PBC score whereas lower scores represent PCR bias and read duplication. However, ChIP-seq is inherently biased as antibodies are used specifically to immunoprecipitate genomic regions of interest. Therefore, low PBC values do not necessarily mean the ChIP-seq data is of poor quality. In fact, this can be a sign that certain genomic regions are overrepresented in the DNA library which is expected in ChIP-seq. Very high scores can indicate no enrichment of specific genomic regions. For example, H3K4me3 marks are located around active gene promoters which may share common motifs such as transcription factor binding sites in a specific cell type. Therefore, duplicate reads could occur if these motifs are highly enriched in the DNA library. Removing duplicate reads in this case may remove true biological information. However, very low PBC scores are undesirable and could be a symptom of PCR amplification bias. ChIP-seq replicate 1 (donor 071508A) exhibited moderate to severe PCR bottlenecking, most likely due to excessive PCR cycles.

Cross strand correlation analysis assesses enrichment of the immunoprecipitated protein independent of peak calling using read density on forward and reverse strands. It is expected that in a good quality ChIP-seq dataset, there will be high densities of read counts on both the forward and reverse strands centred around the binding site of the protein. The distance between the read density peaks on the forward and reverse strands should reflect the size of the predominant sonicated DNA fragment. A cross correlation metric can be calculated by shifting the two strands by fixed, incremental distances in both directions and finding the Pearson's correlation at each shift. The NSC is the maximum correlation value, occurring at a shift equal to the DNA fragment size, divided by the minimum correlation value at strand shift (the background cross correlation). The minimum NSC value is 1; values lower than 1.1 are considered low and values above 1.1 indicates higher enrichment of the bound protein. The correlation at the fragment length shift minus the background cross correlation shows a tag peak when plotted against all other distance shifts. In short read data (read length smaller than 100bp), which represent the majority of next generation sequencing datasets, of complex mammalian genomes such as human and mouse, a peak is also seen at a strand shift equal to the read length. This is known as a phantom peak. The ratio of the fragment length peak compared to the phantom peaks gives the RSC value. A bigger fragment length peak relative to the read length peak demonstrates more enrichment of the immunoprecipitated protein. The RSC value ranges from 0 to above 1 with 0 indicating no enrichment and above 1 indicating high enrichment. Similar to PBC, the cross strand correlation metric values provide an indication of histone enrichment and signal to noise ratio but taken alone, they do not necessarily determine whether ChIP-seq data is biologically meaningful. They are useful for gauging whether ChIP-seq data is technically sound but further analysis is needed before drawing conclusions. Furthermore, they are designed on the basis that enrichment of a DNA binding protein gives a sharp peak signal. This occurs for transcription factor ChIP-seq but not for proteins or histone marks that give both narrow and broad peaks such as RNA polymerase II and H3K27ac (Furey, 2012; Wang et al, 2016). Histones produce broad peaks as they tend to be enriched over a larger area compared to transcription factors. This enrichment profile can result in lower NSC and RSC scores. For all our samples, FRiP far exceeded the minimum threshold of 1% stipulated by ENCODE.

Enrichment for specific histone marks was seen at expected genomic features, for example, H3K4me3 at gene promoters. The histone marks we assayed are very well characterised so this information is not novel. Nonetheless, confirming the enrichment profiles of individual histone marks in the genome is a fundamental initial step in the ChIP-seq analysis pipeline.

ENCODE recommends a minimum of two ChIP-seq replicates (Landt et al, 2012) and therefore we generated histone ChIP-seq data for two chondrogenesis experiments using different hMSC donors. However, they were performed separately. For this reason, we assessed the similarity between the replicates. For measuring reproducibility of transcription factor ChIP-seq experiments, ENCODE recommends calculating the irreproducible discovery rate (IDR). IDR scores peaks on how reproducible they are between two replicates, peaks with low scores can be excluded from the data (Li et al, 2011). Unfortunately, this method was not designed to be used with histone ChIP-seq. This further highlights the limited bioinformatics tools available at this time to interrogate histone ChIP-seq. We examined our replicates using a correlation heatmap, PCA analysis and by determining overlapping peaks. Comparison of the two biological ChIP-seq replicates showed that whilst samples for the same histone marks clustered together, peak overlap between the replicates was generally poor. This is not surprising considering the two ChIP-seq experiments were performed and sequenced at different times. Other studies also found poor peak overlap between replicates, with high PBC values exacerbating the difference (Yang et al, 2014). H3K4me3, H3K4me1 and H3K27ac showed more parity between the two replicates compared to H3K27me3 and H3K36me3. Although the ChIP-seq replicates derived from the same *in vitro* chondrogenesis model, the experiments were carried out at different times and therefore are not true biological replicates. Furthermore, the hMSC donors were different and therefore any the differences identified could reflect natural biological variation. Surprisingly, there is no consensus for combining ChIP-seq replicates from the same experiment or otherwise. Published ChIP-seq studies have either pooled all replicates, used ENCODE's IDR method, selected the best replicate or, for experiments with n > 3, used a majority rule approach (Yang et al, 2014). Each method has its own pros and cons; ChIP-seq analysis pipelines must be tailored to the project. Our peak sets were kept separate as it was unnecessary to combine them and peaks were not used in further analysis. With this in mind, we considered our ChIP-seq replicates to be sufficiently consistent for the next step in the data analysis.

In summary, an optimised workflow for extracting chromatin from hMSCs and chondrocytes from a cartilage-like disc was developed. We generated a histone (H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3) ChIP-seq dataset for hMSCs and differentiated chondrocytes. Initial QC of the data showed that histone read enrichments were located at expected genomic regions. Further analysis is required to elucidate the histone modification changes during chondrogenesis.

3.5 Conclusions

- We generated a histone ChIP-seq dataset of the *in vitro* transwell model of chondrogenesis using antibodies against H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3. An input control was included.

- QC of our dataset established that our samples were of varying quality, with samples from replicate 2 showing higher quality metrics than replicate 1.

- Enrichment of histone marks were seen at expected genomic features, with a high density of H3K4me3 present at promoters, enhancer marks H2K27ac and H3K4me1 were found further away from promoters and gene body marks H3K27me3 and H3K36me3 further still.

- Analysis of peaks revealed that the same histone modifications clustered together and enhancer marks also showed clustering of the same cell types from both replicates. H3K4me3, H3K4me1 and H3K27ac peak sets showed more overlapping peaks between replicates compared to H3K27me3 and H3K36me3.

Chapter 4. Integration of chondrogenesis histone ChIP-seq to RNA-seq data

4.1 Introduction

We generated a histone ChIP-seq dataset for an *in vitro* model of chondrogenesis, described in the previous chapter. Analysis of the data showed that histone mark enrichments were found at expected locations in the genome.

Regulation of gene transcription is partly regulated though epigenetic mechanisms during chondrogenesis (Furumatsu and Asahara, 2010) and is also mediated by transcription factors such as SOX9. SOX9 is widely considered the master transcription factor driving chondrogenesis and is essential for chondrogenesis (Akiyama et al, 2004). Histone modifying enzymes have been found to drive SOX9 induced gene expression during chondrogenesis (Hata et al, 2013) and play a role in chondrocyte maintenance (Huh et al, 2007). Histone modifications act along with other factors to regulate gene expression.

RNA-seq is a high throughput method utilising next generation sequencing to quantify transcripts in a population of cells. This method offers many advantages over conventional gene expression assays such as RT-qPCR and microarrays. Gene microarrays and RNA-seq are both considered genome wide methods unlike RT-qPCR. However, RNA-seq has a greater dynamic range compared to microarrays at quantifying both low and highly expressed genes. Furthermore, unlike microarrays, RNA-seq does not rely on transcript specific probes and can aid in the detection of novel genes and transcripts (Zhao et al, 2014). We generated RNA-seq data for the same chondrogenesis model used for histone ChIP-seq. RNA-seq was performed at n = 1 (hMSC donor 017508A), therefore statistical testing for differential expression was not possible and consequently, these results are exploratory. However, a DNA microarray was previously performed (n = 3) for the same chondrogenesis model and this was correlated with RNA-seq data to assess the similarity between the two datasets.

Both RNA-seq and microarray data were generated prior to the conception of this project and a full in-depth analysis of these data is beyond the scope of this PhD. In this chapter, the connection between histone mark enrichment and gene expression during the *in vitro* transwell model of chondrogenesis is investigated.

4.2 Aims

- Associate histone marks (H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3) assayed using ChIP-seq to gene expression quantified by RNA-seq (correlate RNA-seq to microarray data).

- Explore the relationship between histone modifications and gene expression during chondrogenesis.

4.3 Results

4.3.1 Chondrogenesis RNA-seq analysis

Gene expression in hMSCs and differentiated chondrocytes was determined using RNA-seq. RNA-seq reads were assessed using FastQC and MultiQC (Appendix ii, Fig. 1 and Table 1). Transcript isoforms were quantified using Salmon in quasi-mapping mode and summarised to gene level in RStudio using the tximport package. Differential expression tests were not possible due to insufficient sample numbers (n = 1 for both hMSC and differentiated chondrocytes), therefore genes were quantified in TPM and up- or down-regulated genes were considered as such by their log2 fold change in TPM. In the absence of statistical analysis, genes with a log2 fold change (TPM + 1) > 1.5 were arbitrarily considered to be differentially expressed in our analysis. Using this fold change cut off, 447 genes were upregulated and 2771 genes were downregulated (Fig. 4.1).

RNA-seq data was correlated to microarray data (n = 3, hMSC donor 017508A) of the same chondrogenesis model to confirm gene expression changes. Concordance between the two technologies was good and there was a significant correlation between gene expression changes in the RNA-seq and microarray datasets (Fig. 4.2). Although variation between individual genes cannot be assessed, the correlation between RNA-seq and microarray shows that overall, the RNA-seq dataset is similar. More genes were downregulated compared to upregulated in our RNA-seq data (2771 and 447) but it is not known whether these changes are statistically significant. Our chondrogenesis microarray data showed that 598 genes were significantly (q < 0.05) upregulated with a log2 fold change > 1.5 and 562 were significantly downregulated by a log2 fold change < -1.5. A separate cDNA microarray study found more genes were downregulated during chondrogenesis (Yoo et al, 2011).

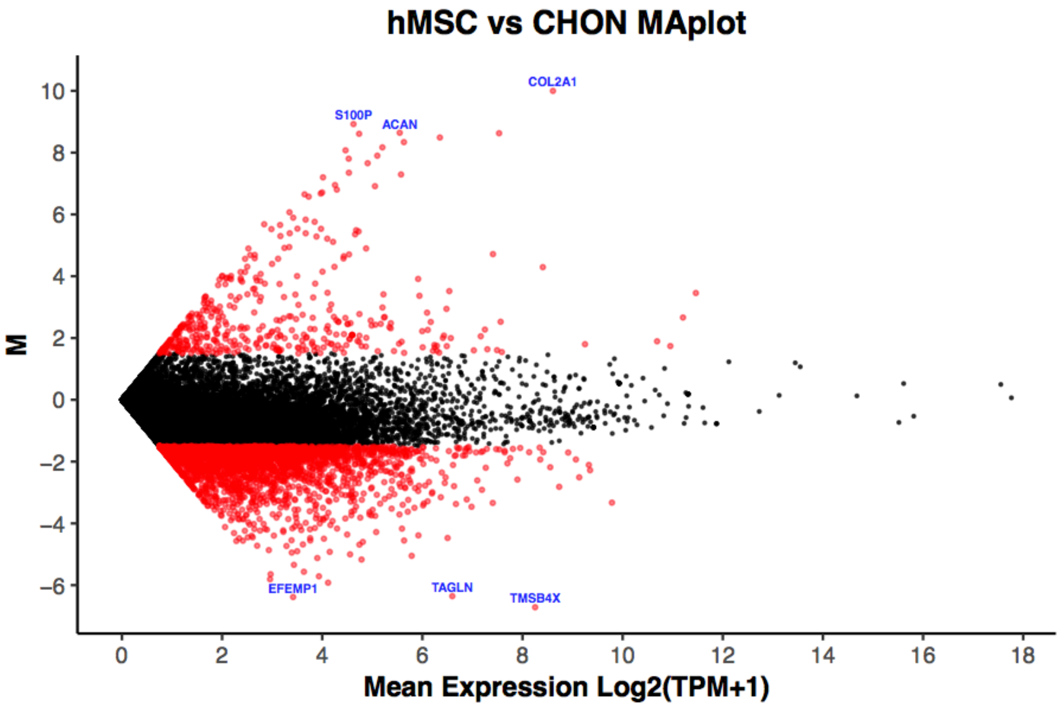


Figure 4.1 – MA (log ratio vs average expression of all samples) plot of hMSC vs differentiated chondrocyte RNA-seq. Genes with a log2 FC (TPM + 1) > 1.5 are shown in red. The top 3 upregulated genes COL2A1, S100P and ACAN are labelled as are the top 3 downregulated genes TMSB4X, EFEMP1 and TAGLN. Plot generated in RStudio using the ggplot2 package

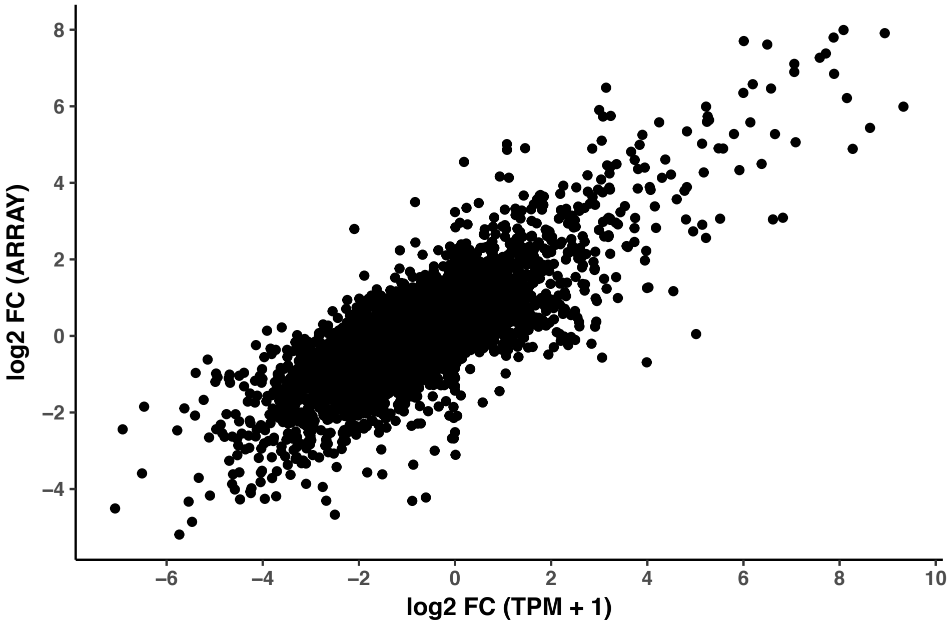


Figure 4.2 – Correlation of gene expression changes quantified by RNA-seq (x axis) and microarray (y axis). A Pearson's correlation test showed there was a significant correlation between RNA-seq and microarray data (r = 0.75, p < 2.2x10^{-16}). Plot generated in RStudio using the ggplot2 package.

RNA-seq was used to correlate histone modifications to gene expression due to it having a larger dynamic range than microarrays. RNA-seq determined that the most upregulated gene during chondrogenesis is *COL2A1* with a log2 fold change of 10, followed by *S100P* with a log2 fold change of 8.9 and *ACAN* with 8.64. The *COL2A1* gene encodes for the alpha I chain of type II collagen, the major collagen found in cartilage. *S100P* encodes for calcium binding protein P; S100 protein family members are found in articular cartilage and are involved in cell proliferation and survival (Yammani, 2012). They are also found to have an important role in both osteoarthritis and inflammatory arthritis (Bertheloot and Latz, 2017). Aggrecan, encoded by the *ACAN* gene, is the main proteoglycan found in articular cartilage and was the third most upregulated gene in our analysis. A network of proteoglycans and collagens form the extracellular matrix present in cartilage (Kiani et al, 2002). The most downregulated gene during chondrogenesis was *TMSB4X* (Thymosin Beta 4, X linked); this gene is highly expressed in BM-MSCs and plays a role in cell migration (Huang et al, 2015). The second most downregulated gene was *EFEMP1* which encodes for the Fibulin-3 protein, an extracellular matrix glycoprotein. Overexpression of this gene was found to negatively regulate chondrogenesis (Wakabayashi et al, 2010). The third most downregulated gene was *TAGLN* (encodes for the Transgelin protein) which is highly expressed in hMSCs (Silva et al, 2003). Knockout of TAGLN promoted chondrogenesis in vascular smooth muscle cells (Shen et al, 2011) although it is upregulated during hMSC differentiation into adipocytes and osteoblasts *in vitro* (Elsafadi et al, 2016). This suggests downregulation of *TAGLN* is required for MSCs to enter the chondrogenic lineage as opposed to the adipogenic or osteoblastogenic lineages.

We also assessed up- and downregulated genes using the DAVID GO term analysis tool. Upregulated genes displayed GO terms relating to chondrogenesis and cartilage development (Table 4.1). In contrast, downregulated genes were associated with more heterogeneous GO terms, including those related to epigenetic regulation of gene expression and cell adhesion (Table 4.2). Both up- and downregulated genes shared the GO term *osteoblast differentiation*. Osteoblasts are one of the three main cell lineages MSCs can differentiate into and chondrocytes can transdifferentiate into osteoblasts during the terminal hypertrophic stage of chondrogenesis (Zhou et al, 2014; Park et al, 2015). The mutual GO term indicates that some genes related to osteoblast differentiation are upregulated and others are downregulated. The same is also true for the GO term *cell adhesion*. This illustrates the precise regulation of gene expression during the differentiation process.

To summarise, RNA-seq of an *in vitro* model of chondrogenesis revealed that, by our definition, 447 genes are upregulated (log2 fold change > 1.5) with GO terms relating to chondrogenesis and 2771 genes were downregulated which had GO terms not related to chondrogenesis. This validates that gene expression changes in our *in vitro* model quantified by RNA-seq mimics those of developing chondrocytes *in vivo*.

Table 4.1 – All significant (Benjamini-Hochberg FDR < 0.05) biological process GO terms for 447 genes upregulated (log2 fold change > 1.5) during chondrogenesis. GO terms found using DAVID GO analysis tool.

TERM	PVALUE	BENJAMINI
GO:0001501~SKELETAL SYSTEM DEVELOPMENT	4.40E-20	7.29E-17
GO:0030198~EXTRACELLULAR MATRIX ORGANIZATION	4.25E-16	3.68E-13
GO:0001503~OSSIFICATION	7.27E-11	4.02E-08
GO:0030199~COLLAGEN FIBRIL ORGANIZATION	7.47E-10	3.10E-07
GO:0051216~CARTILAGE DEVELOPMENT	4.06E-09	1.35E-06
GO:0001958~ENDOCHONDRAL OSSIFICATION	7.90E-09	2.18E-06
GO:0030574~COLLAGEN CATABOLIC PROCESS	1.00E-08	2.37E-06
GO:0001649~OSTEOBLAST DIFFERENTIATION	2.11E-07	4.37E-05
GO:0018146~KERATAN SULFATE BIOSYNTHETIC PROCESS	3.45E-07	6.35E-05
GO:0030206~CHONDROITIN SULFATE BIOSYNTHETIC PROCESS	3.19E-06	5.28E-04
GO:0042340~KERATAN SULFATE CATABOLIC PROCESS	3.71E-05	0.005580342
GO:0002062~CHONDROCYTE DIFFERENTIATION	4.80E-05	0.006614541
GO:0061621~CANONICAL GLYCOLYSIS	6.90E-05	0.008763528
GO:0006094~GLUCONEOGENESIS	9.67E-05	0.011392251
GO:0007155~CELL ADHESION	1.30E-04	0.014249007
GO:0048706~EMBRYONIC SKELETAL SYSTEM DEVELOPMENT	1.41E-04	0.014544664
GO:0006029~PROTEOGLYCAN METABOLIC PROCESS	1.64E-04	0.015902774
GO:0001837~EPITHELIAL TO MESENCHYMAL TRANSITION	2.61E-04	0.023763274
GO:0022617~EXTRACELLULAR MATRIX DISASSEMBLY	3.16E-04	0.027223699
GO:0001502~CARTILAGE CONDENSATION	4.79E-04	0.038973293
GO:0050679~POSITIVE REGULATION OF EPITHELIAL CELL PROLIFERATION	5.46E-04	0.042206073

Table 4.2 – Top 35 significant (Benjamini-Hochberg FDR < 0.05) biological process GO terms for 2771 genes downregulated (log2 fold change < -1.5) during chondrogenesis. GO terms found using DAVID GO analysis tool.

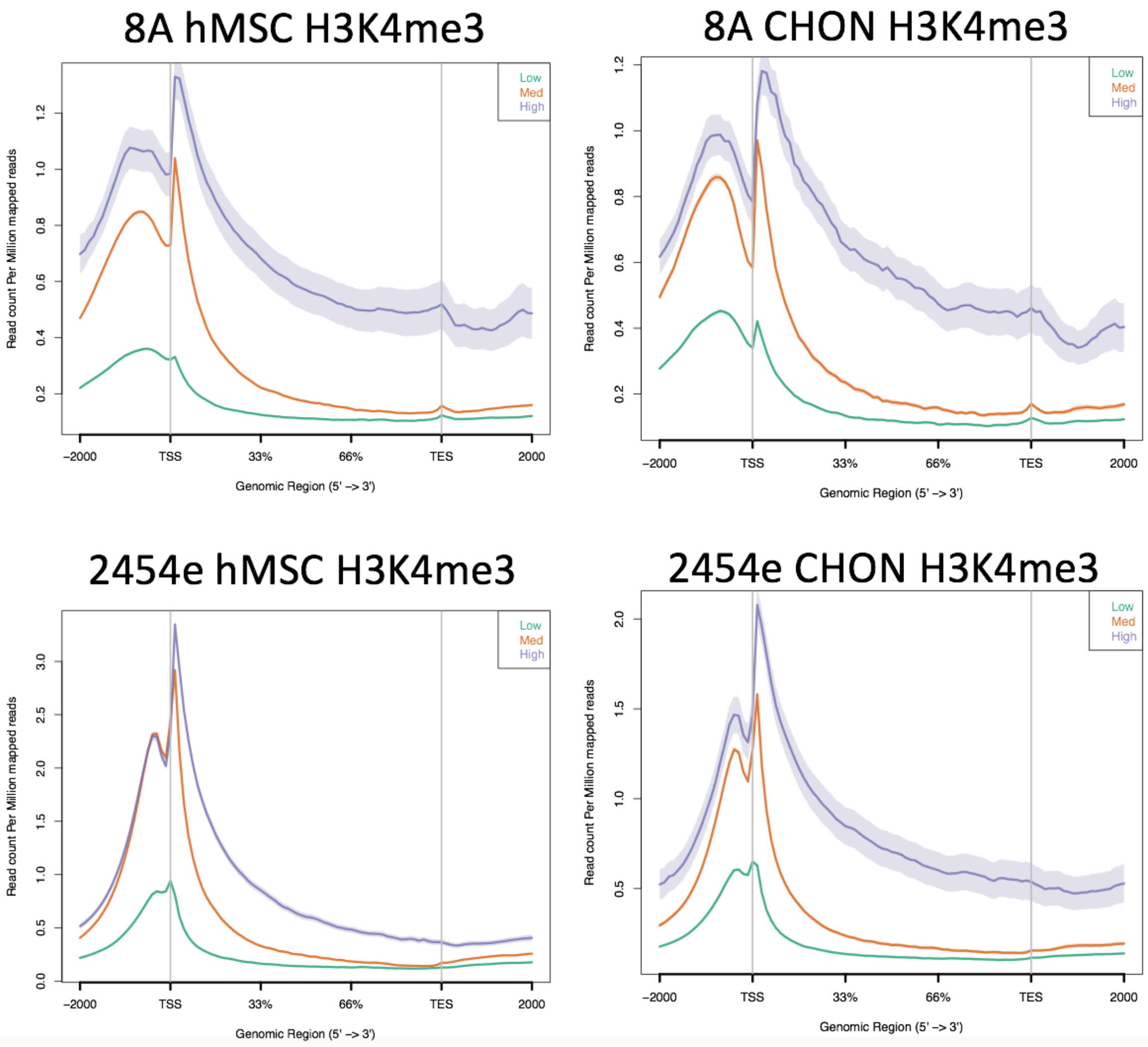
TERM	PVALUE	BENJAMINI
GO:0098609~CELL-CELL ADHESION	3.20E-13	1.77E-09
GO:0006334~NUCLEOSOME ASSEMBLY	1.00E-10	2.77E-07
GO:0007155~CELL ADHESION	1.91E-09	3.53E-06
GO:0045815~POSITIVE REGULATION OF GENE EXPRESSION, EPIGENETIC	1.17E-08	1.62E-05
GO:0000183~CHROMATIN SILENCING AT RDNA	3.96E-08	4.38E-05
GO:0006342~CHROMATIN SILENCING	4.80E-08	4.42E-05
GO:0006335~DNA REPLICATION-DEPENDENT NUCLEOSOME ASSEMBLY	1.39E-07	1.10E-04
GO:0032200~TELOMERE ORGANIZATION	4.80E-07	2.65E-04
GO:0051290~PROTEIN HETEROTETRAMERIZATION	4.40E-07	2.70E-04
GO:0045814~NEGATIVE REGULATION OF GENE EXPRESSION, EPIGENETIC	3.95E-07	2.73E-04
GO:0008283~CELL PROLIFERATION	7.10E-07	3.27E-04
GO:0006915~APOPTOTIC PROCESS	6.63E-07	3.33E-04
GO:0000086~G2/M TRANSITION OF MITOTIC CELL CYCLE	1.03E-06	4.37E-04
GO:0048146~POSITIVE REGULATION OF FIBROBLAST PROLIFERATION	7.33E-06	0.002890489
GO:0008285~NEGATIVE REGULATION OF CELL PROLIFERATION	8.02E-06	0.002951469
GO:0048661~POSITIVE REGULATION OF SMOOTH MUSCLE CELL PROLIFERATION	1.08E-05	0.003734314
GO:0045669~POSITIVE REGULATION OF OSTEOBLAST DIFFERENTIATION	1.08E-05	0.003734314
GO:0060071~WNT SIGNALING PATHWAY, PLANAR CELL POLARITY PATHWAY	1.67E-05	0.004608263
GO:0051436~NEGATIVE REGULATION OF UBIQUITIN-PROTEIN LIGASE ACTIVITY INVOLVED IN MITOTIC CELL CYCLE	1.51E-05	0.00461467
GO:0030036~ACTIN CYTOSKELETON ORGANIZATION	1.61E-05	0.004673126
GO:0043488~REGULATION OF MRNA STABILITY	1.79E-05	0.004690179
GO:0051437~POSITIVE REGULATION OF UBIQUITIN-PROTEIN LIGASE ACTIVITY INVOLVED IN REGULATION OF MITOTIC CELL CYCLE TRANSITION	1.50E-05	0.00485641
GO:0051603~PROTEOLYSIS INVOLVED IN CELLULAR PROTEIN CATABOLIC PROCESS	1.99E-05	0.004998955
GO:0031145~ANAPHASE-PROMOTING COMPLEX-DEPENDENT CATABOLIC PROCESS	3.00E-05	0.006604989
GO:0016032~VIRAL PROCESS	2.99E-05	0.006857345
GO:0007067~MITOTIC NUCLEAR DIVISION	2.93E-05	0.007025921
GO:0090263~POSITIVE REGULATION OF CANONICAL WNT SIGNALING PATHWAY	5.23E-05	0.009585958
GO:0000226~MICROTUBULE CYTOSKELETON ORGANIZATION	5.07E-05	0.009622283
GO:0007179~TRANSFORMING GROWTH FACTOR BETA RECEPTOR SIGNALING PATHWAY	4.97E-05	0.009756755
GO:0006521~REGULATION OF CELLULAR AMINO ACID METABOLIC PROCESS	4.89E-05	0.009968827
GO:0016477~CELL MIGRATION	4.76E-05	0.01007307
GO:0001649~OSTEOBLAST DIFFERENTIATION	6.05E-05	0.010724088
GO:0006979~RESPONSE TO OXIDATIVE STRESS	6.51E-05	0.011183195
GO:0031047~GENE SILENCING BY RNA	7.73E-05	0.012869512

4.3.2 Histone modification enrichment is associated with gene expression levels

Following analysis of RNA-seq data, gene expression levels were correlated to histone modifications. Quantified transcripts were grouped into low, medium or high expression genes and the profile of histone mark enrichment across the gene was plotted. Gene expression was measured in TPM and density plots (Appendix ii, Fig. 2) were generated for genes at day 0 (hMSCs) and day 14 (differentiated chondrocytes). Genes with a TPM < 2 were defined as low; density plots showed that the majority of genes had a TPM below this. The mean TPM in both hMSCs and differentiated chondrocytes was 16.7 and genes with a medium level of expression were defined as having a TPM between 2 and 16.7. Genes with a TPM > 16.7 were defined as highly expressed. These thresholds are somewhat arbitrary as is no consensus method of categorising gene expression. Using this method, 50089 and 53999 transcripts were defined as low in hMSCs and chondrocytes respectively. The number of medium genes in hMSCs was 9161 and 6581 in chondrocytes. There were 2982 genes defined as highly expressed in hMSCs and 1652 in chondrocytes.

Normalised (RPM) read counts for all ChIP-seq samples were mapped to low, medium and highly expressed gene groups quantified by RNA-seq. There was good concordance between both ChIP-seq replicates although replicate 2454e displayed more variance. Highly expressed genes showed greater H3K4me3 read counts close to the gene TSS, followed by medium and low expressed genes (Fig. 4.2). The greatest density of H3K4me3 reads mapped to slightly downstream of the TSS, before dropping along the gene body. Histone marks H3K4me1 (Fig. 4.3), H3K27ac (Fig 4.4) and H3K36me3 (Fig. 4.5) displayed a similar pattern with respect to expression levels although these histone marks show higher enrichment further along the gene body compared to H3K4me3. H3K4me1 reads displayed a dip in enrichment at the TSS but rises along the gene body. H3K27ac displayed higher enrichment at the TSS and also remained high within the gene body. H3K36me3 reads are low at the TSS, gradually rising until the TES before decreasing again. In contrast, increased read enrichment of H3K27me3 is associated with genes with a low level of expression (Fig. 4.6). Interestingly, for H3K27me3, highly expressed genes showed greater enrichment compared to genes with a medium level of expression.

When read counts from replicate samples were merged and further normalised to the input control, read profiles remained the same for all marks apart from H3K27me3 which showed that genes with a low level of expression had the highest H3K27me3 enrichment followed by medium and highly expressed genes (Fig. 4.8). Overall, active histone marks in our analysis are associated with higher levels of gene expression during chondrogenesis. Additionally, the profile of histone mark reads in the genome relative to the gene TSS corroborates with the analysis performed in the previous chapter (Results Chapter 3). This analysis confirms expected outcomes and verifies that it is possible to integrate ChIP-seq and RNA-seq datasets to link gene expression to histone modification enrichments.


8A hMSC H3K4me3

8A CHON H3K4me3

2454e hMSC H3K4me3

2454e CHON H3K4me3

Figure 4.3 – H3K4me3 read enrichment in relation to gene TSS in all ChIP-seq samples for high, medium and low expressed genes in chondrogenesis. Shading around the line represents the standard error from the mean. Plots were generated using the ngs.plot tool.

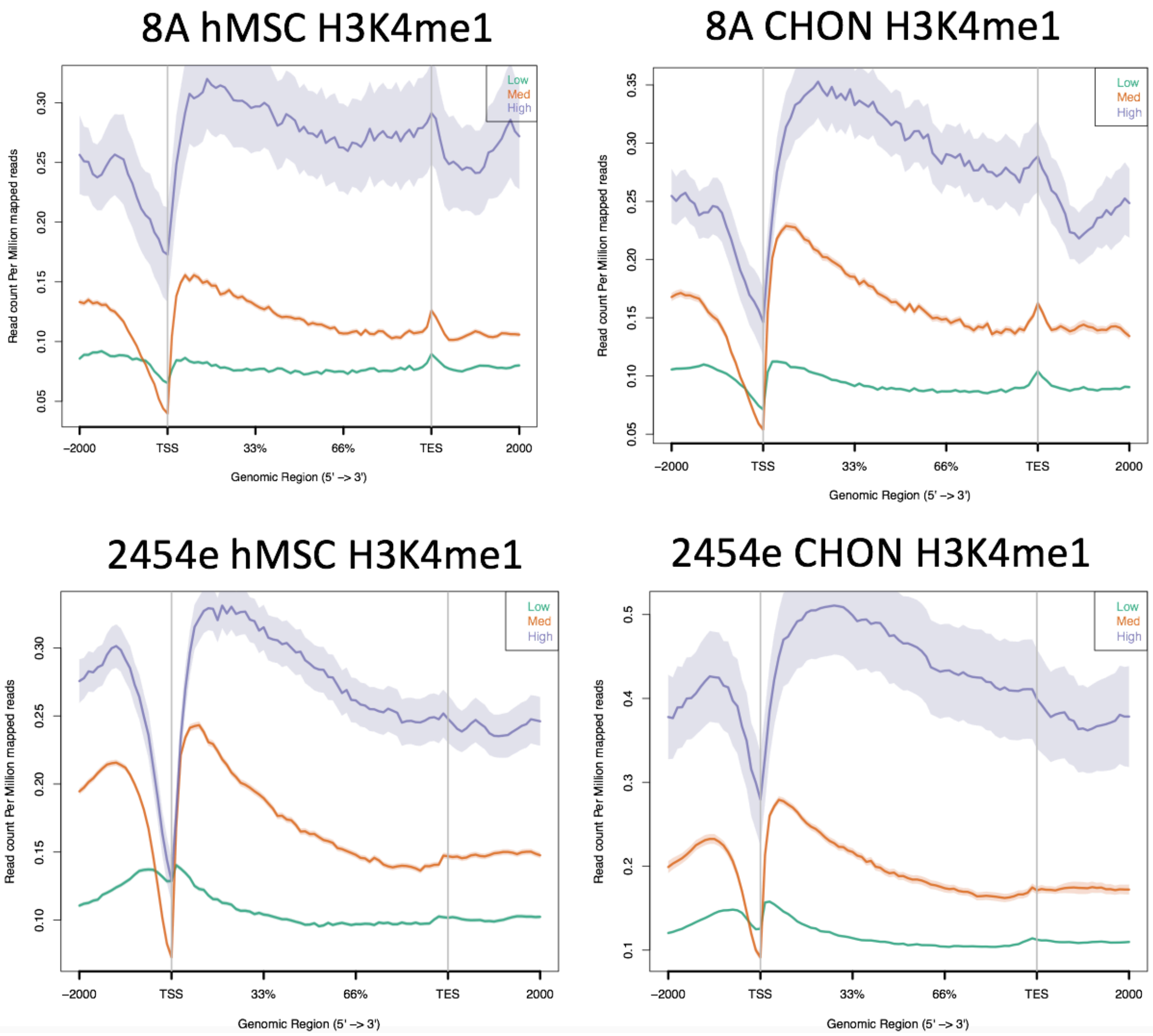


Figure 4.4 – H3K4me1 read enrichment in relation to gene TSS in all ChIP-seq samples for high, medium and low expressed genes in chondrogenesis. Shading around the line represents the standard error from the mean. Plots were generated using the ngs.plot tool.

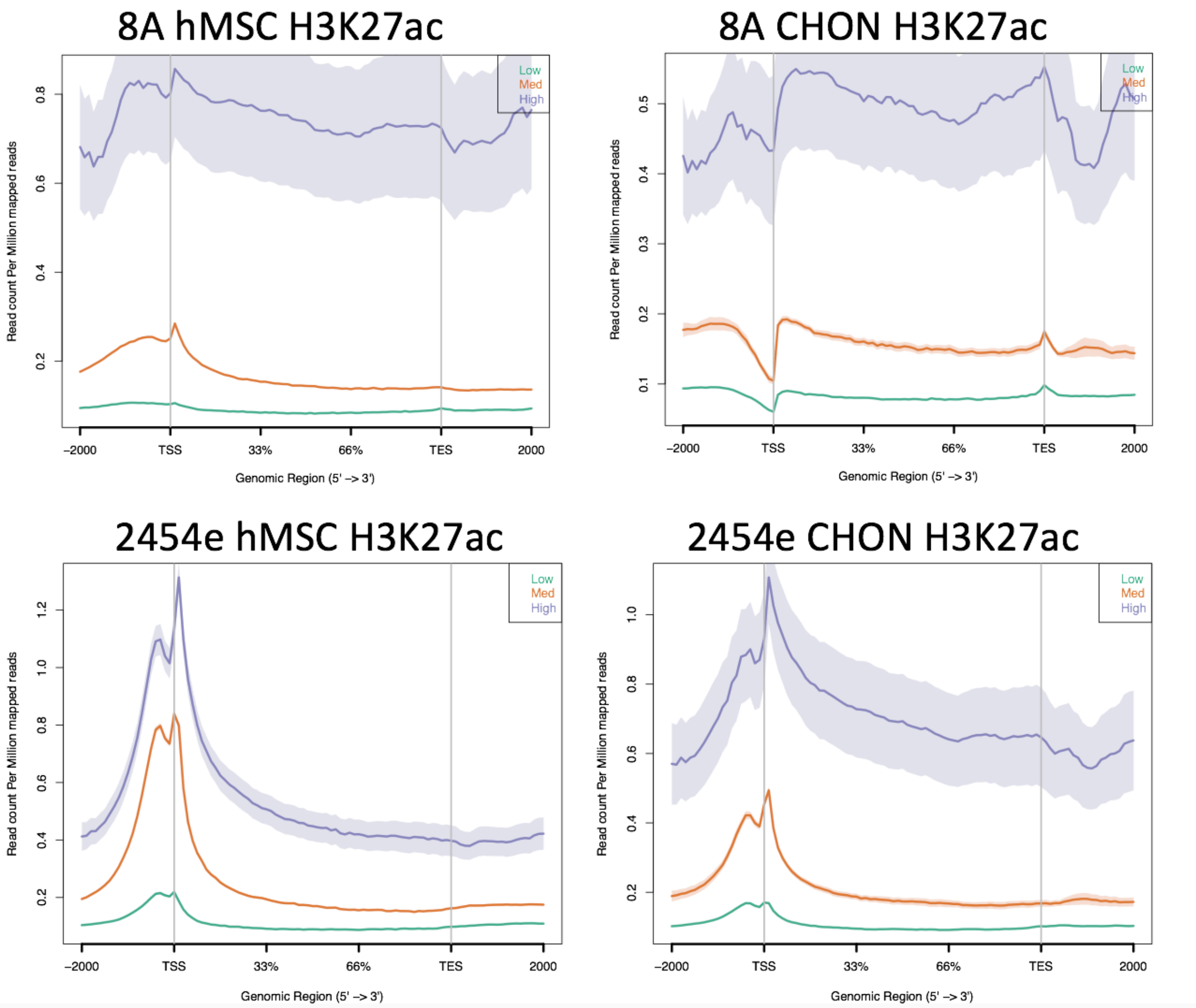


Figure 4.5 - H3K27ac read enrichment in relation to gene TSS in all ChIP-seq samples for high, medium and low expressed genes in chondrogenesis. Shading around the line represents the standard error from the mean. Plots were generated using the ngs.plot tool.

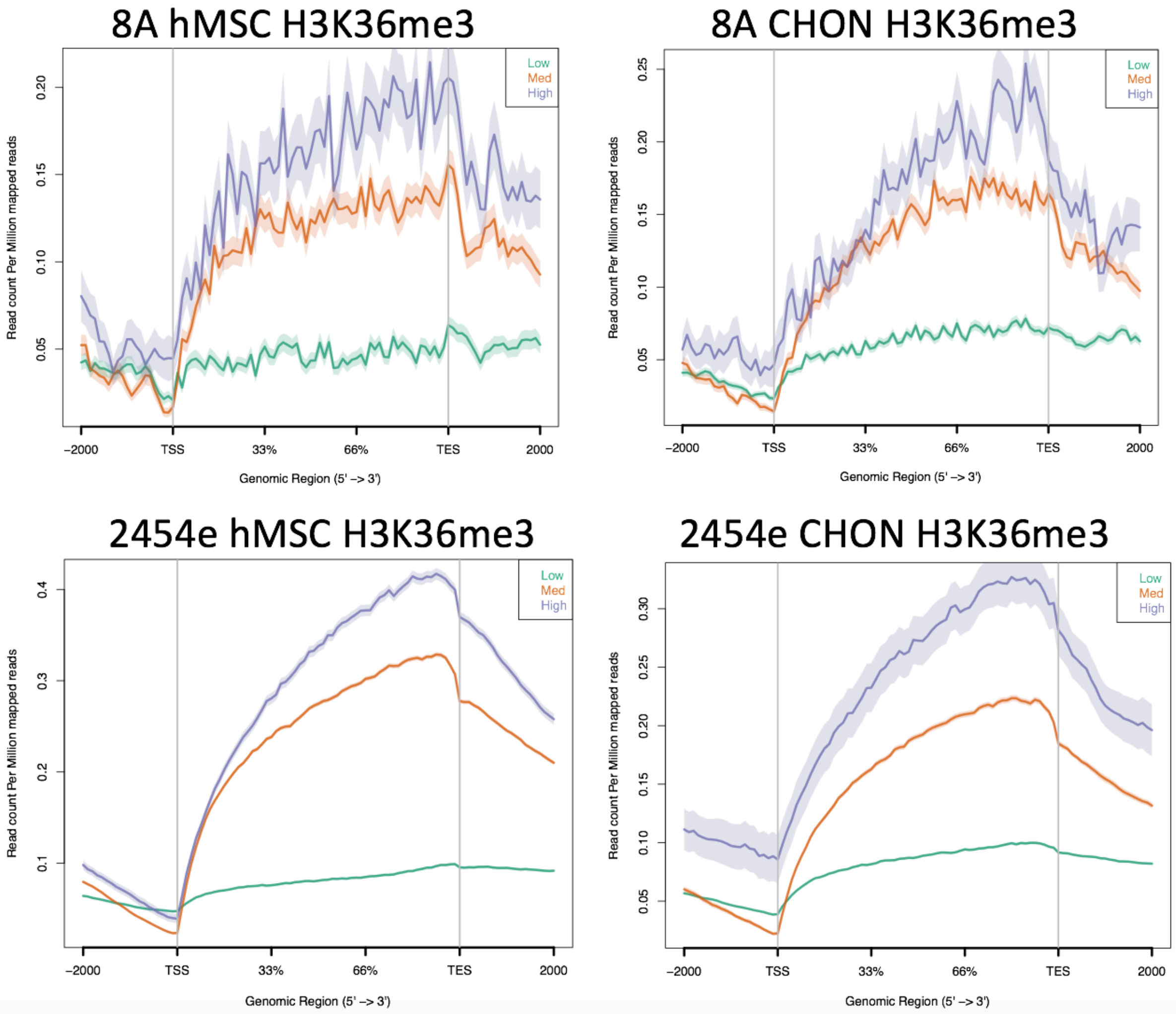


Figure 4.6 – H3K36me3 read enrichment in relation to gene TSS in all ChIP-seq samples for high, medium and low expressed genes in chondrogenesis. Shading around the line represents the standard error from the mean. Plots were generated using the ngs.plot tool.

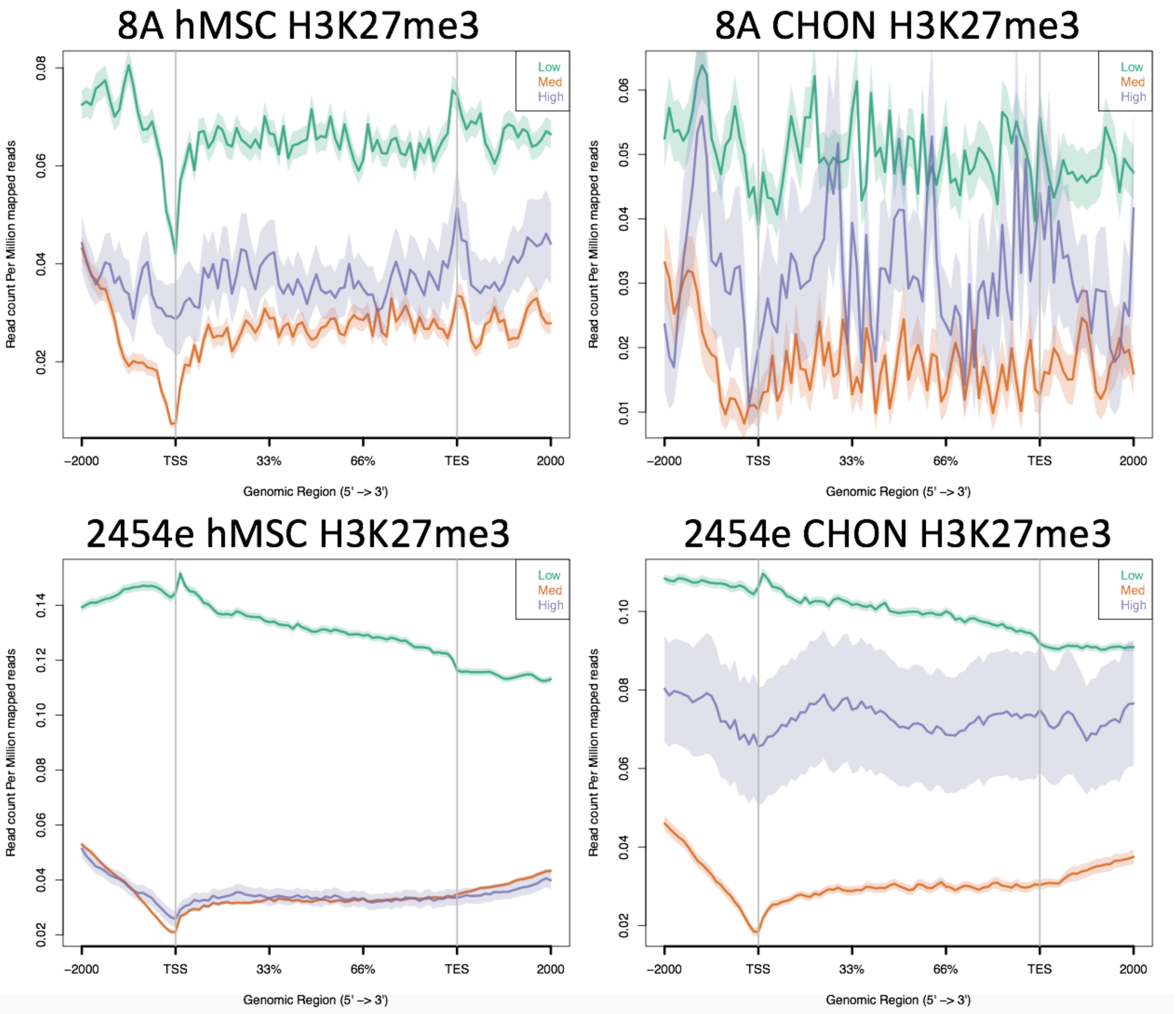


Figure 4.7 - H3K27me3 read enrichment in relation to gene TSS in all ChIP-seq samples for high, medium and low expressed genes in chondrogenesis. Shading around the line represents the standard error from the mean. Plots were generated using the ngs.plot tool.

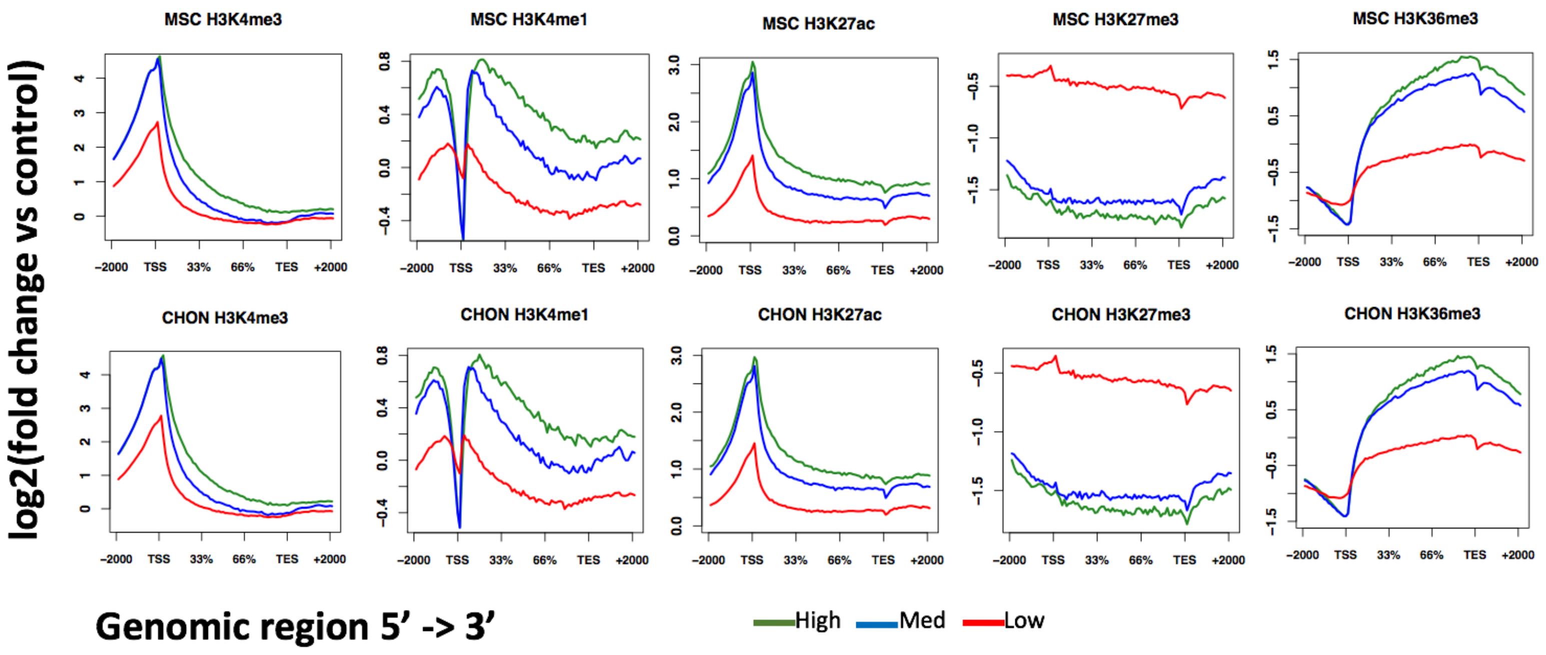


Figure4.8 – Signal to noise ratio of histone modifications in low, medium and highly expressed genes. Profile plots were created in ngs.plot.r to calculate the log2 fold change of histone marks vs input control and re-plotted in RStudio

4.3.3 Correlation of changes in histone mark enrichment to changes in gene expression

We observed that levels of specific histone marks correlate with gene expression levels during chondrogenesis. We next sought to determine whether a change in histone mark enrichment corresponds to a change in gene expression. In the previous chapter, we saw that H3K4me1 and H3K27ac histone marks clustered by cell type rather than ChIP-seq replicate suggesting that biological differences between hMSCs and differentiated chondrocytes were greater than the differences between replicates or other technical effects. Therefore, only these two histone marks were chosen for differential peak enrichment analysis after combining hMSC and chondrocyte replicates. Significant differentially enriched peaks between hMSCs and differentiated chondrocytes were found for H3K4me1 and H3K27ac samples using the Bioconductor package DiffBind. There were 3537 significantly differentially enriched (FDR < 0.05) H3K4me1 peaks between hMSCs and differentiated chondrocytes (Fig. 4.9A), of which 2939 increased in enrichment and 598 decreased in enrichment. For H3K27ac, there were 7004 significantly differentially enriched peaks (Fig. 4.9B), of which 6298 showed an increase in enrichment and 706 decreased in enrichment during chondrogenesis.

All peaks were associated to the nearest gene using HOMER annotatepeaks.pl and the change in gene expression was plotted alongside the change in histone mark enrichment (Fig. 4.10). Using this method, genes may be associated with more than one peak. A clear correlation between a change in H3K4me1 (Fig. 4.8A) or H3K27ac (Fig. 4.10B) enrichment and a change in gene expression was not apparent except for genes that were highly upregulated (log2 fold change > 1.5). A Pearson's correlation test gave an overall significant correlation of 0.016 (*p* = 6.109X10^{-14}) for H3K27ac and 0.07 (*p* < 2.2X10$^{-16)}$ for H3K4me1. However, at such a large sample size, tests are more likely to yield significant *p* values and the effect size must be given more importance (Cohen, 1992). With such small effect sizes, it is difficult to conclude whether these correlations are of biological relevance. The correlation effect sizes improve when only significant differentially enriched peaks are taken into account, with *r* = 0.12 (*p* < 2.2X10^{-16}) for H3K27ac and *r* = 0.24 (*p* < 2.2X10^{-16}) for H3K4me1. Although small, the improvement in correlation indicates there may be a true correlation between change in histone mark enrichment and a change in gene expression. There were many genes that were differentially expressed between hMSCs and differentiated chondrocytes that did not show any changes in H3K4me1 or H3K27ac enrichment. These genes may be regulated by other histone modifications or by other epigenetic mechanisms.

Combinations of histone marks are a better indicator of gene expression, therefore we sought to determine whether genes associated with both H3K4me1 and H3K27ac enrichment increases were more likely to be upregulated. To view the association between changes in histone mark enrichment and gene expression with more clarity, only genes with a log2 fold change > 1.5 that are also associated with a significant differentially enriched H3K4me1 and/or H3K27ac peak were considered (Fig. 4.11). Only the enrichment of the closest H3K27ac/H3K4me1 peak was considered. In this analysis, genes may be associated with a combination of both increased H3K27ac and H3K4me1 enrichment, increased H3K27ac and decrease in H3K4me1 enrichment or vice versa. Alternatively, they may be associated with a decrease in enrichment of both histone marks. Using these criteria, upregulated genes tended to exhibit increased enrichment of both H3K4me1 and H3K27ac marks during chondrogenesis. However, the enrichment of histone peaks associated with downregulated genes were more variable. Downregulated genes were found to be associated with all four combinations of H3K27ac/H3K4me1 enrichment. This suggests that other mechanisms may also be involved or association of histone marks to genes using the nearest gene approach may not be optimal.

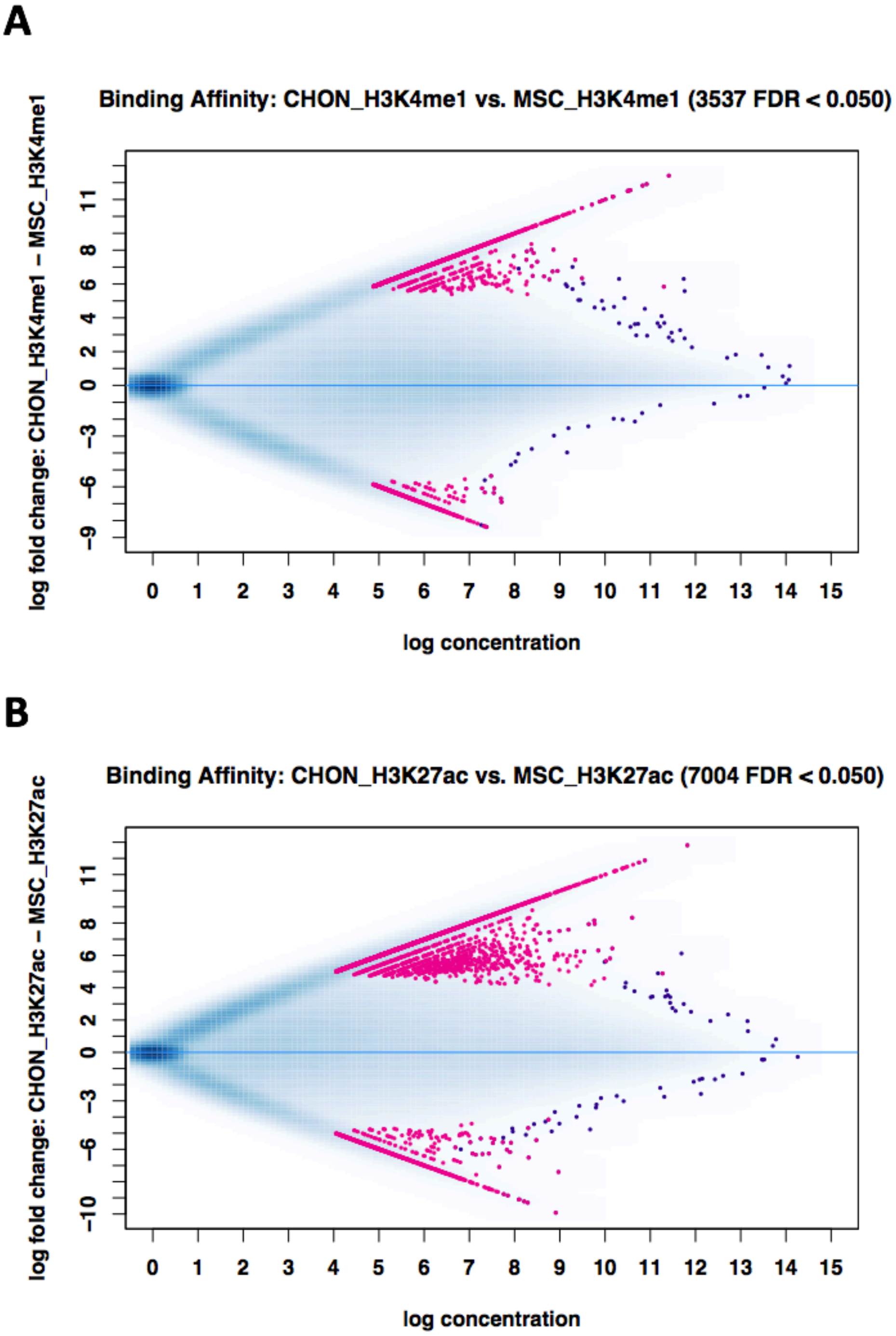


Figure 4.9 – (A) significant (FDR < 0.05) differentially enriched H3K4me1 peaks and (B) H3K27ac peaks between hMSCs and differentiated chondrocytes. Read counts (log concentration) are log2 transformed. Significantly enriched peaks are coloured in pink. Differential binding analysis and MA plot generated using the DiffBind package in RStudio.

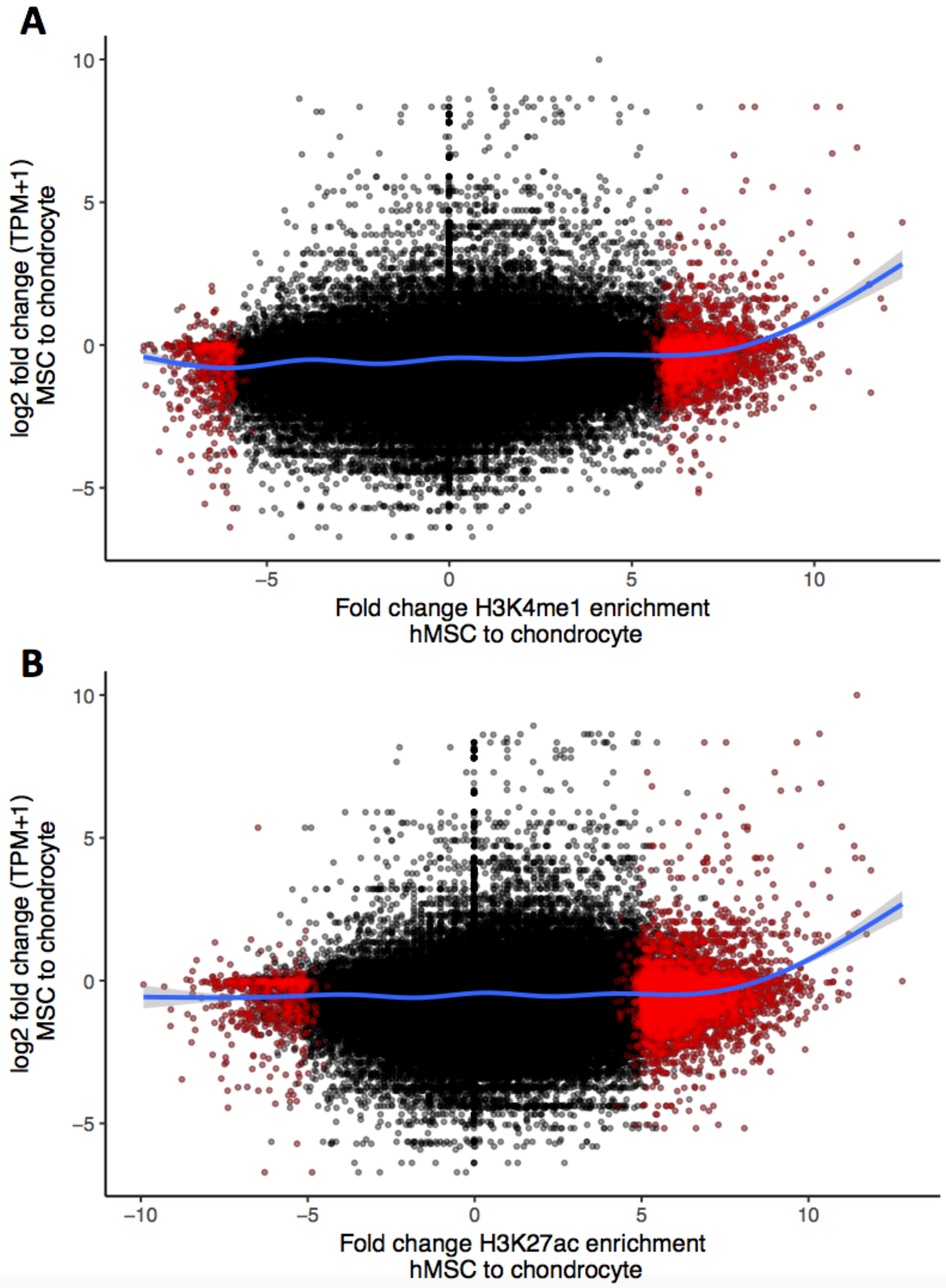


Figure 4.10 – (A) log2 fold change in H3K4me1 enrichment and (B) log2 fold change in H3K27ac enrichment alongside change in gene expression during chondrogenesis of the nearest gene. Significantly enriched histone peaks are coloured in red. A generalised additive model line was fitted to the data. Plots were built using the ggplot2 package in RStudio.

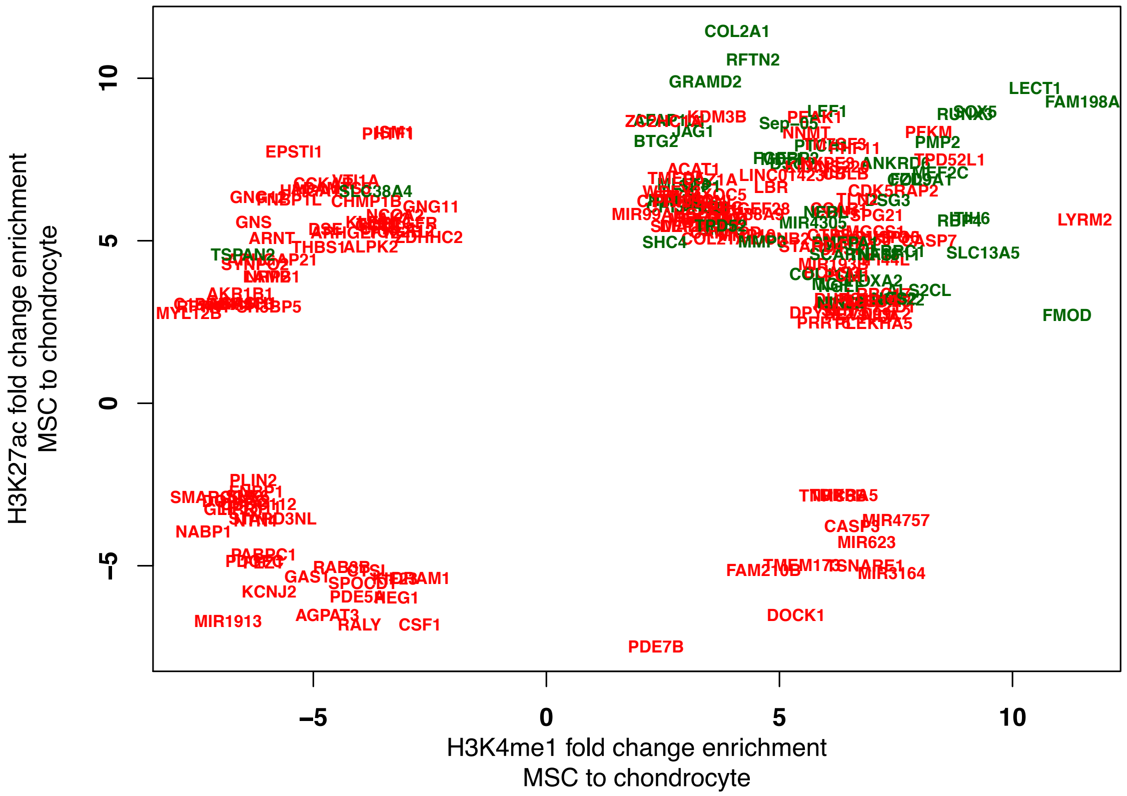


Figure 4.11 – Change in enrichment of H3K4me1 and H3K27ac peaks during chondrogenesis. Peaks were associated with the nearest gene using HOMER annotatepeaks.pl and only the closest peak to the gene was considered. Genes with a log2 fold change > 1.5 and a significant differentially enriched H3K4me1 and/or H3K27ac peak are shown. Upregulated genes are coloured in green and downregulated genes are coloured red. Plot was built using the ggplot2 package in RStudio.

4.4 Discussion

In this chapter, gene expression measured by RNA-seq during chondrogenesis was correlated to histone modification enrichment. RNA-seq of hMSCs and differentiated chondrocytes revealed 3218 genes had a log2 fold change > 1.5. This cut off value is somewhat arbitrary and there is no consensus on fold change thresholds. This threshold was used to define differentially expression genes as statistical analysis of this RNA-seq dataset was not possible due to a sample number of 1 for both hMSC and differentiated chondrocytes. Consequently, there is no way of identifying or taking into account potential technical or hMSC donor effects; a gene with a high log2 fold change in our analysis may not be showing a true biological difference between hMSC and differentiated chondrocytes. However, there was a significant correlation of RNA-seq gene expression changes to microarray data of the same chondrogenesis model with replicates and this and acted as an extra quality control step for the RNA-seq data in absence of replicates. GO terms related to chondrogenesis were detected for upregulated genes with a log2 fold change > 1.5. Furthermore, *ACAN* and *COL2A1*, the two major components of articular cartilage, were highly upregulated.

Integration of expression levels to histone ChIP-seq samples yielded expected outcomes. Association of RNA-seq data to histone ChIP-seq data revealed a correlation between active histone marks and high gene expression during chondrogenesis. Conversely, genes with a low level of expression (TPM < 2) were enriched in repressive H3K27me3 marks. However, genes with a high level of expression exhibited greater enrichment of H3K27me3 marks compared to the medium expression geneset. This may be due to bivalent promoters of genes involved in the differentiation process. However, when reads from the two replicates were merged and normalised to the input control, highly expressed genes showed the least amount of enrichment for H3K27me3. Previous studies have found that highly expressed genes are likely to exhibit both active marks such as H3K4me3 as well as repressive mark H3K27me3. Genes with bivalent marks were more likely to change during differentiation (Shah et al, 2014). Bivalency is thought to allow prompt activation of genes upon an appropriate signal whilst maintaining the gene in an inactive state in the absence of signals (Voigt et al, 2013). The presence of bivalent regulatory states is important for embryonic stem cells where precise timely regulation of gene transcription is crucial to control both maintenance of the stem cell state and differentiation into other cells (Bernstein et al, 2006; Vastenhouw and Schier, 2012). H3K27me3 marks are removed prior to gene activation when differentiating to a specific lineage. On the other hand, H3K4me3 marks may be removed instead to repress gene activation. Alternatively, the bivalent state may be maintained even post-differentiation (Gobbi et al, 2011). ChIP-seq is commonly performed on a population of cells, including our chondrogenesis ChIP-seq experiments, and therefore it is impossible to establish whether H3K4me3 and H3K27me3 are located in the same genomic region of the same cell, or whether the marks occur heterogeneously between cells in the same population. It is possible to use single cell ChIP-seq to assay co-occurrence of histone marks in genomic regions of the same cell but single cell ChIP is challenging due to the low amount of starting material which increases the level of background noise (Clark et al, 2016). Another method to assay histone modifications at the same genomic loci is to use a second ChIP assay using a different antibody to pull down regions of co-localisation (Furlan-Magaril, 2009).

In the previous chapter, correlation of ChIP-seq replicate samples showed that only H3K27ac and H3K4me1 samples clustered by cell type and not by ChIP-seq replicate/hMSC donor. We observed that upregulated genes (log2 fold change > 1.5), tended to be associated with increased enrichment of both H3K27ac and H3K4me1 marks although there were also downregulated genes with increased H3K27ac and H3K4me1 marks. In our analysis H3K4me1 and H3K27ac peaks were associated to their nearest gene. H3K4me1 marks gene enhancers whereas H3K27ac marks both promoters and enhancers. Enhancers can be located large distances from their target gene, do not necessarily target the nearest gene and not all genes have enhancers (Marsman and Horsfield, 2010). Therefore, linking H3K4me1 and H3K27ac peaks to the nearest gene is not optimal and without further data such as chromatin conformation assays, it is difficult to predict which genes H3K4me1 and H3K27ac peaks are associated with. This could explain why we did not observe a strong correlation between the change in enrichment of these histone marks and a change in gene expression except for in highly upregulated genes. This also offers an explanation as to why some downregulated genes appeared to also be associated with increased enrichment of H3K4me1 and H3K27ac marks.

We observed that more H3K4me1 and H3K27ac histone marks show a significant increase in enrichment compared to decreased enrichment during chondrogenesis. As these two marks are associated with active gene enhancers, an increase in enrichment suggests an increase of enhancer activity during chondrogenesis. Gene enhancers are instrumental in regulating differentiation of stem cells (Zhou et al, 2014; Cico et al, 2016; Wu et al, 2017) including chondrogenesis (Lui and Lefebvre, 2014). A previous study integrating histone modifications to gene expression in a different *in vitro* chondrogenesis model showed the same outcomes as this analysis (Herlofsen et al, 2013). Regulatory states including enhancers and the combination of histone marks that define them will be explored in later chapters.

To summarise this chapter, changes in single histone modifications, specifically H3K4me1 and H3K27ac, are a poor indicator of gene expression changes in our data. Whilst individual histone marks can infer specific gene regulatory meanings, combinations of marks are often a more powerful indicator of regulatory state. There was a link between an increase of H3K4me1 and H3K27ac enrichment and increased gene expression in highly upregulated genes. Correlating histone ChIP-seq and gene expression in this way comes with a number of challenges and caveats, besides the challenges of associating histone mark peaks to genes mentioned earlier. Traditional ChIP-seq methods are not strictly quantitative between cell types without additional external controls such as the addition of an exogenous reference epigenome (Orlando et al, 2014). Additionally, it is unknown how enriched a ChIP-seq peak needs to be in order to influence gene transcriptions or whether this is consistent across cell types or biological processes.

Integration of data from different technologies can be an arduous task; lack of benchmarks and consensus methodologies means analyses and interpretation of results can be subjective or inconsistent between researchers. Despite this, integration of RNA-seq and ChIP-seq data offers a high throughput, genome wide approach for associating gene transcription to the epigenome. With the rise of affordable ‘omics technologies and a growing amount of publically available data generated by large consortia, it has been acknowledged that methods for analysing and integrating data need to be improved (Gomez-Cabrero et al, 2014). There have been many studies involving the integration of ChIP-seq and gene expression data (Dudziac et al, 2012;

Angelini and Costa, 2014; Ayyapan et al, 2015), including during chondrogenesis (Herlofsen et al, 2013). Herlofsen et al found a high correlation between H3K4me3 and gene expression, as well as an enrichment of enhancer marks with chondrogenic genes in agreement with our study. ChIP-seq data can be used to predict gene expression (McLeay et al, 2012) and integration of epigenomic and gene expression data can be used to infer gene regulatory networks (Angelini and Costa, 2014). With increasing interest in the epigenome and the regulation of gene transcription using genome wide methods, integration of heterogeneous datasets will become more commonplace.

4.5 Conclusion

- RNA-seq determined that 447 genes were upregulated in chondrogenesis and 2771 genes were downregulated by log2 fold change > 1.5. Upregulated genes displayed GO terms related to chondrogenesis and cartilage development.

- H3K4me3, H3K4me1, H3K27ac and H3K27me3 ChIP-seq read enrichments are higher in highly expressed genes compared to genes with a lower level of expression.

- Repressive mark H3K27me3 read enrichment is higher in low expressed genes than highly expressed genes.

- There is a correlation between a change in histone mark enrichment and a change in gene expression in chondrogenesis. Multiple histone marks are better indicators of gene expression changes compared to single marks.

Chapter 5: Correlation of DNA methylation to chromatin states

5.1 Introduction

Histone marks around genes can influence gene transcription. The histone code is the theory that gene expression can be predicted simply by studying the histone modifications around a gene (Jenuwein and Allis, 2001). Specific marks located near or on genes can influence whether that gene is actively transcribed or repressed Previous chapters have explained that H3K4me3 marks active promoters, H3K4me1 and H3K27ac mark gene enhancers, H3K27me3 marks repressed genes and H3K36me3 mark transcriptionally active gene bodies. Histone modifications attract other proteins including chromatin remodellers that can alter the structure of chromatin to make the DNA either more or less accessible to the transcription machinery. Active histone modifications lead to chromatin being remodelled into an open configuration, allowing genes within the open chromatin to be transcribed (Yan and Boyd, 2006). Repressive marks attract protein complexes that further compact the chromatin rendering genes inaccessible to transcription factors (Kadoch et al, 2016). The histone code is more complex than originally thought and it is now known that single histone marks do not often exist in isolation. In the previous chapter, we explored how single histone modifications are a poor predictor of gene expression. Combinations of different histone modifications can define gene *cis*-regulatory elements which give a better indication of whether a gene is transcribed compared to investigating individual histone modifications.

Multiple histone ChIP-seq datasets may be integrated to define chromatin states. These are distinct regions of chromatin that are marked by varying levels of different histone modifications. Chromatin states can reflect *cis*-regulatory elements such as gene promoters and enhancers. Defining and characterising chromatin states forms the basis of the Epigenomics Roadmap project, a large scale project which aims to elucidate the epigenome of multiple cell types – similar to the premise of the ENCODE project. Using ChIP-seq data, the Epigenomics Roadmap project has created chromatin states for 127 human cell types (Kundaje et al, 2015; Romanoski et al, 2015). The Epigenomics Roadmap project generated a 15 state model using five histone marks (H3K4me3, H3K4me1, H3K9me3, H3K27me3 and H3K36me3) and an extended 18 state model using six histone marks for cell types with an extra histone modification (H3K27ac) available. Our chondrogenesis histone ChIP-seq experiment included all the same marks apart from H3K9me3. H3K9me3 is a marker of constitutively repressed heterochromatin (Saksouk et al, 2015). The Roadmap project used the ChromHMM software (Ernst and Kellis, 2012) to learn chromatin states from their data. This was also used to generate chromatin states in our chondrogenesis ChIP-seq data.

Other epigenetic mechanisms also exist alongside histone mark defined chromatin states and these also contribute to the regulation of gene transcription. DNA methylation occurs at CpG sites within the genome, mediated by DNA methyltransferases. Specifically, the cytosine within a CpG site is methylated to 5-methylcytosine. CpG sites are typically located in CpG islands near the start site of genes (Lim and Maher, 2010). Although DNA methylation is generally a repressive mark, DNA methylation of CpG sites within gene bodies is associated with active transcription (Yang et al, 2014). During differentiation of stem cells, large scale changes in DNA methylation are observed (Sheaffer et al, 2014; Altun et al, 2010). Both histone modifications and DNA methylation contribute to control of gene transcription by attracting chromatin remodellers to alter the structure of chromatin (Geiman and Robertson, 2002). The two mechanisms influence each other (Cedar and Bergman, 2009), therefore integrating information from DNA methylation and histone modification data could lead to new insights into gene regulation.

One of the aims of this project was to generate chromatin states using our chondrogenesis histone ChIP-seq dataset. Our lab has also previously performed a DNA 450K methylation array using the same *in vitro* model of chondrogenesis. We hypothesised that by integrating DNA methylation and histone modification data, we would be able to identify regulatory elements important for controlling gene transcription in chondrogenesis.

5.2 Aims

- Characterise chromatin states for hMSC and differentiated chondrocyte ChIP-seq dataset using the ChromHMM software.

- Compare hMSC and chondrocyte derived chromatin states to Roadmap Epigenomics chromatin states.

- Analyse DNA 450K methylation array of hMSCs vs chondrocytes.

- Correlate DNA methylation to chromatin states and identify chromatin states where DNA methylation is prevalent.

- Carry out pathway analysis of methylated CpG sites and chromatin states.

5.3 Results

5.3.1 ChromHMM defined 16 distinct chromatin states

Chromatin states for hMSCs and differentiated chondrocytes were computed using the ChromHMM software in order to investigate all histone marks collectively and define regulatory elements in the genome. All aligned ChIP-seq reads were included in chromatin state learning with all input samples used as control. ChromHMM learned 16 chromatin states from our ChIP-seq data (Fig. 5.1). The 16 state model was arrived at by executing ChromHMM with different numbers of states until a good separation of distinct states was seen. This was also the method used by the Epigenomics Roadmap project. Furthermore, this number also corroborates with the 18 state extended model from Roadmap; their extended 18 state model included one extra histone mark, H3K9me3. This histone mark displayed noticeable emission probabilities in only 2 states in their 18 state model. This histone mark was not included in our experiment, therefore without the 2 extra states present in the 18 state Epigenomics Roadmap model we decided 16 would be a reasonable number for chromatin state learning. Chromatin states are computed as emission probabilities which indicate the likelihood that a specific histone mark will be present in a chromatin state and represent the enrichment of specific histone marks in that state.

Chromatin states were annotated and named using information from published literature including the ENCODE and Epigenomics Roadmap projects. Active promoter/TSS states (1_TssA and 2_TssS) displayed high enrichment of H3K4me3 and H3K27ac marks with low levels of other histone marks. Flanking TSS states (3_TssFlnk, 4_TssFlnkU, 5_TssFlnkD) also included H3K4me1 and H3K36me3 marks. The bivalent TSS state 6_TssBiv is characterised with high amounts of active H3K4me3 mark as well as the repressive mark H3K27me3. Weak and strong transcription states (7_TxWk and 8_TxS) displayed low and high enrichment of H3K36me3 marks respectively. Flanking active transcription (9_TxFlnk) showed equal levels of H3K4me3, H3K4me1 and H3K27ac marks. Genic enhancers (10_EnhG1 and 11_EnhG2) displayed enhancer histone marks as well as H3K36me3. In contrast, only enhancer histone marks were present in enhancer states (12_EnhA and 13_EnhS). The poised enhancer state (14_EnhP) was characterised by low levels of H3K27ac marks and high levels of H3K4me1. Repressed state (15_Repr) only displayed H3K27me3 histone marks. The final state, 16_Quies, defines regions of the genome not marked by any of the five histone marks included in our ChIP-seq experiment. Emission parameter values are available in Appendix iii, Table 3.

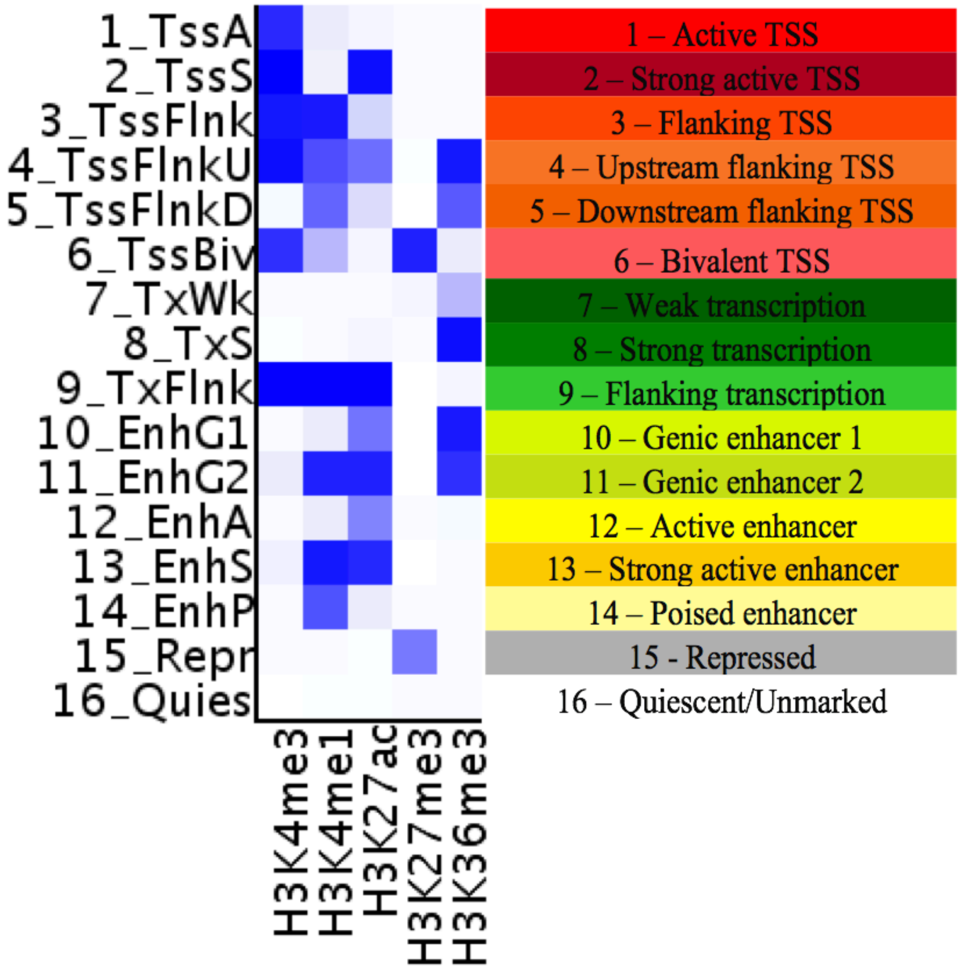


Figure 5.1 - State emission parameters output by ChromHMM from combined hMSC and differentiated chondrocyte alignments. Chromatin state learning using ChromHMM found 16 distinct states marked by varying levels of histone modifications. Input samples were used as controls. All histone marks are represented in at least one state.

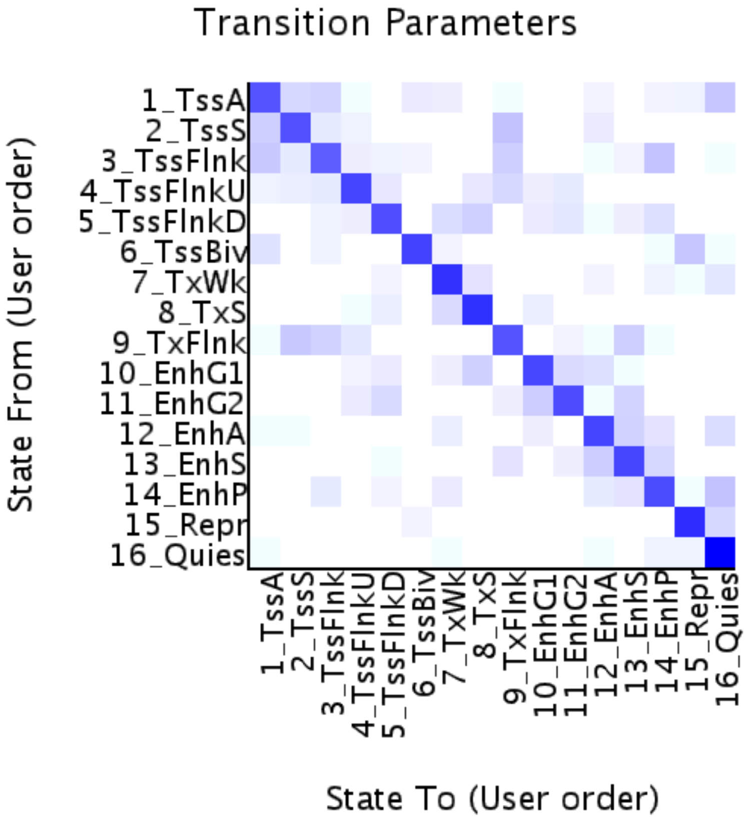


Figure 5.2 - State transition probabilities output by ChromHMM from combined hMSC and differentiated chondrocyte alignments. Transition probabilities shows the likelihood that states in 200bp bins are adjacent to each other in the genome.

ChromHMM also outputs transition probabilities between states (Fig. 5.2). State transition probabilities indicate how likely chromatin states are found to be next to each other in the genome. The spatial relationship between chromatin states in the genome also reflect the spatial changes in histone modifications and transformations of contiguous genomic features such as from promoters to exons and then introns, representing the typical structure of a gene (Ernst and Kellis, 2010). The quiescent state has higher transition probabilities to chromatin states typically located at the peripheral of genes, such as promoters and enhancers rather than states representing gene bodies. Therefore, it could be deduced that quiescent chromatin states may be more likely to fall within intergenic regions. Promoter states (1_TssA, 2_TssS and 3_TssFlnk) are likely to be neighbours; promoter states are also likely to be next to 9_TxFlnk (transcription flanking). The 6_TssBiv state is more likely to be found adjacent to the 15_Repr state. Overall, chromatin states with similar histone marks are more likely to have higher transition probabilities. ChromHMM splits the genome into 200bp bins to learn chromatin states; each chromatin state is made up of these 200bp sections. Identical chromatin state bins are most likely to be adjacent to each other, showing that all chromatin states are on average, longer than 200bp in size.

ChIP-seq sample alignments from both replicates of hMSC and differentiated chondrocytes were used to train the chromatin state model and hence state definitions are the same across the two cell types. ChromHMM also overlaps defined states and genomic features. Overlapping states with genomic features revealed that the majority of the hMSC and differentiated chondrocyte genome is quiescent or unmarked (Fig. 5.3). In hMSCs, 60.9% of the genome is quiescent whereas 65.1% of the genome is quiescent in chondrocytes (Appendix iii, Table 2-3). As a percentage of the genome, most of the other states were within 1% between hMSCs and chondrocytes. Exceptions were seen for 8_TxS; this state represented 7.1% of the hMSC genome and dropped to 2.7% in differentiated chondrocytes, suggesting that either fewer genes exhibited the strong transcription state or that this state is smaller in size in chondrocytes. Furthermore, chondrocytes displayed a slightly higher proportion of enhancer states (12_EnhA, 13_EnhS and 14_EnhP) compared to hMSCs.

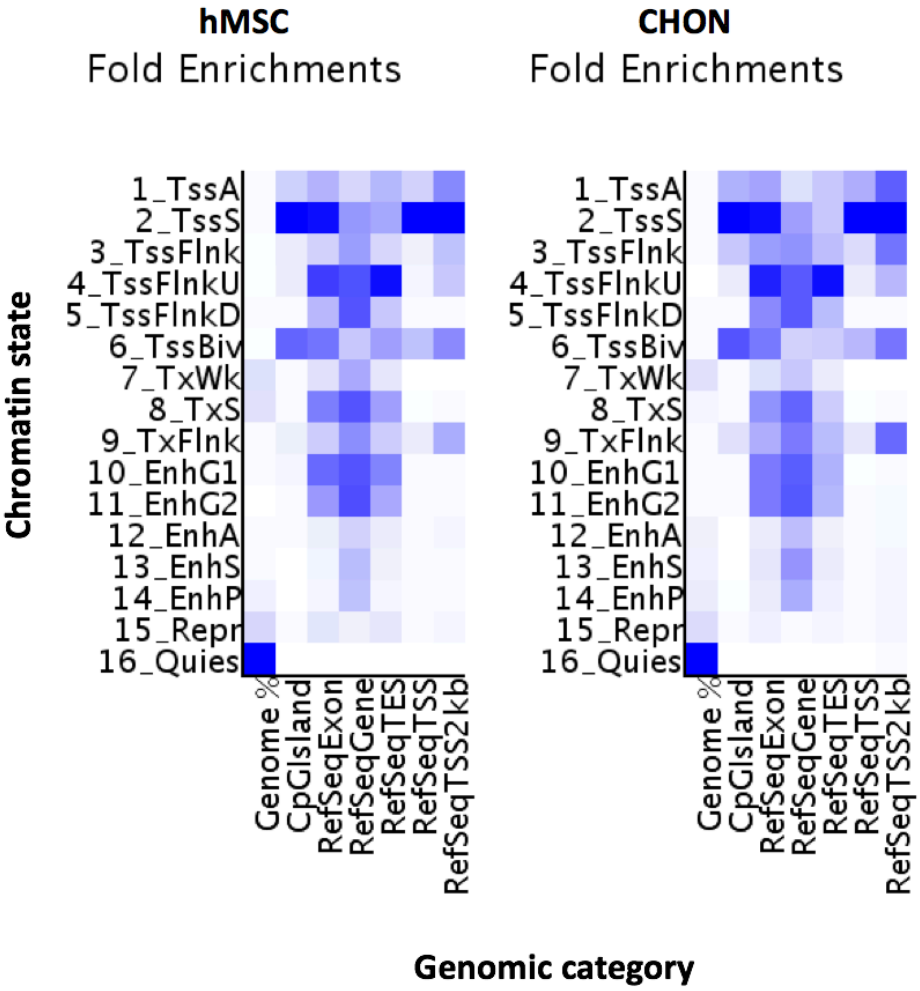


Figure 5.3 - Chromatin state enrichment overlapping genomic categories in hMSCs and chondrocytes determined by ChromHMM. These categories were genome percentage, CpG islands, exons, genes, transcriptional end sites (TES), TSS and the region +/- 2kb from TSS.

Active promoter states generally overlap CpG islands and gene promoters, and at higher proportions than other chromatin states (Fig. 5.3). The 2_TssS state is more likely to be found within 1kb of gene TSSs (Fig. 5.4) compared to other states. Promoter flanking and transcription flanking states are likely to be outside of the 1kb window from the start of the TSS. In general, chromatin states exhibiting high H3K4me3 enrichment are seen closer to the TSS. The pattern of states with a 2kb window from the transcriptional termination site (TTS) is less clear; the probability of finding the 4_TssFlnkU state in this region is greater than the other states, this could occur if genes are situated close together in the genome so that the TTS of one gene is near the TSS of a neighbour gene. Overall, the locations of chromatin states relative to the TSS are in accord with the enrichment peak profiles of the histone marks comprising the states (Chapter 3).

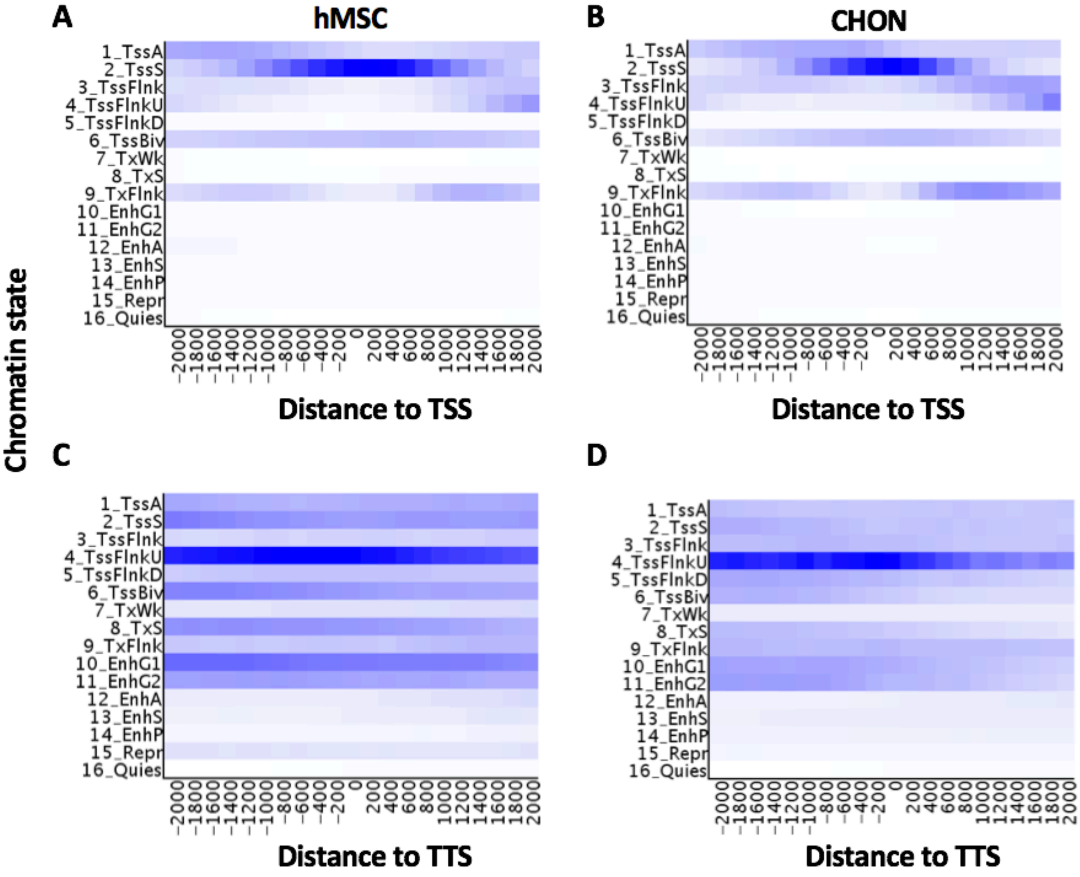


Figure 5.4 – Chromatin state enrichments within 2kb of the TSS and TTS in hMSCs (A and C) and chondrocytes (B and D) determined by the ChromHMM software.

5.3.2 Chromatin state changes

Chromatin state changes between hMSC and differentiated chondrocytes were investigated. Initially, genome browser tracks of chromatin states around specific genes were visualised in IGV. For example, the *COL2A1* gene is highly upregulated during chondrogenesis (RNA-seq measured expression of 17.2 TPM in hMSCs; 16984 TPM in chondrocytes) and the chromatin states prior and post-differentiation reflects its gene expression change (Fig. 5.5). In hMSCs, the *COL2A1* gene promoter is marked by a bivalent TSS state and other regions within the gene body also display this state. There are large stretches of repressed states both up- and downstream of the gene as well as within the gene body. In contrast, in chondrocytes the *COL2A1* is marked by active TSS states at its promoter and flanking active transcription states downstream of the promoter. Poised and active enhancer states are seen upstream of the *COL2A1* gene in chondrocytes and genic enhancer states are present within the gene body.

The *TMEM016C* gene is situated approximately 4kb downstream of the COL2A1 gene. In hMSCs, there are active promoter and transcription states within this gene suggesting that *TMEM106C* (encodes for transmembrane Protein 106C) is transcriptionally active. In chondrocytes, the active promoter state is still present although the transcription states have switched into enhancer states. Gene expression measured by RNA-seq shows that *TMEM106C* is lowly expressed in hMSCs and is downregulated in chondrogenesis although expression is very low in both cell types (0.25 TPM in hMSCs; 0.03 TPM in chondrocytes). The two genes are separated by a comparatively short intergenic region; despite this, there is a quiescent chromatin state between them which separates the gene body chromatin states of the two genes. This demonstrates a segregation of histone modifications between genes in order to prevent the regulation of one gene being affected by a neighbouring gene. Overall, this example illustrates the pronounced histone modification changes surrounding genes that correspond with gene transcription changes.

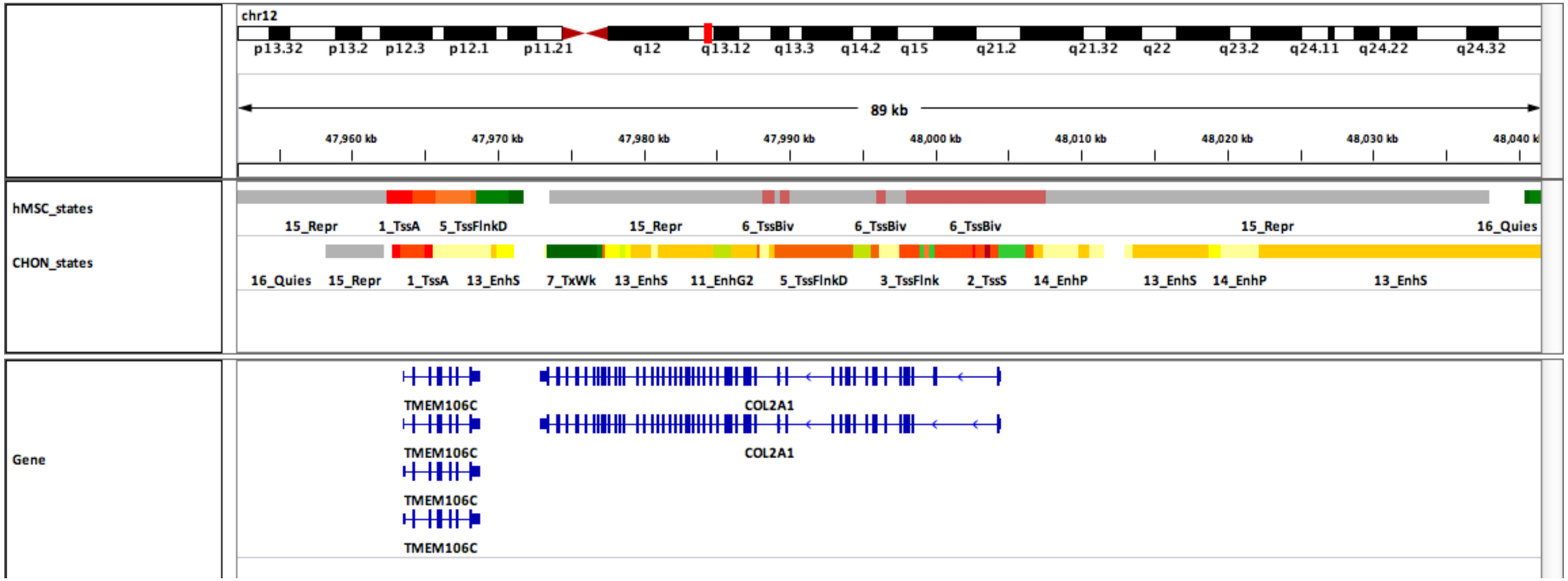


Figure 5.5 – Screenshot of IGV genome browser showing hMSC and chondrocyte chromatin states around the COL2A1 gene. The TMEM106C gene, located downstream of COL2A1 is also displayed.

The hMSC state origins of chromatin states in differentiated chondrocytes were determined to investigate global chromatin state changes. The frequency of which hMSC state each chondrocyte chromatin state originated from was calculated (Fig. 5.6; Fig.5.7). This shows whether chromatin states in chondrocytes were derived from a different state in hMSC or whether the chromatin state remained stable during chondrogenesis. For example, the hMSC origins of the strong enhancer state (13_EnhS) in differentiated chondrocytes were 1% 7_TxWk, 18% 12_EnhA, 42% 13_EnhS, 24% 14_EnhP and 14% 16_Quies. Only 1% of 1_TssA states derived from 6_TssBiv, showing that bivalent promoters constitute a small amount of overall promoter activation. However, 36% of chondrocyte 6_TssBiv originated from the 15_Repr state, demonstrating that promoters become bivalent during chondrogenesis. Many active chromatin states derived from 16_Quies in hMSCs; 2-48% of each active chondrocyte state derived from this state. These changes indicate an activation of genes during chondrogenesis that were previously not active in hMSCs.

The large scale chromatin state changes observed between hMSCs and differentiated chondrocytes implies that the chromatin is remodelled extensively during chondrogenesis to alter gene transcription. Whilst we can view chromatin states around single genes of interest using a genome browser, it is not feasible to do this for all genes in the genome. Therefore, pathway analysis using GO terms for gene lists offers a way of associating chromatin states to groups of genes.

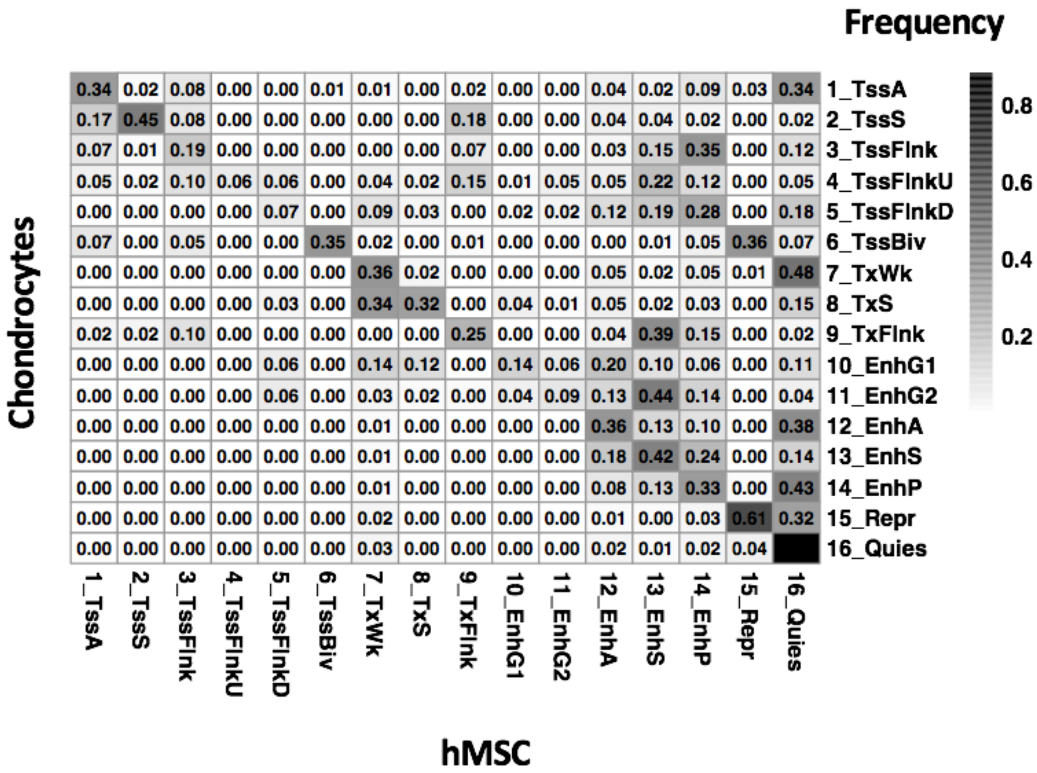


Figure 5.6 – Frequency of chromatin state changes from hMSCs to differentiated chondrocytes. Genome co-ordinates were split into 200bp bins and the chromatin states in hMSC and differentiated chondrocytes were noted at each bin. Plot generated using ggplot2 in RStudio.

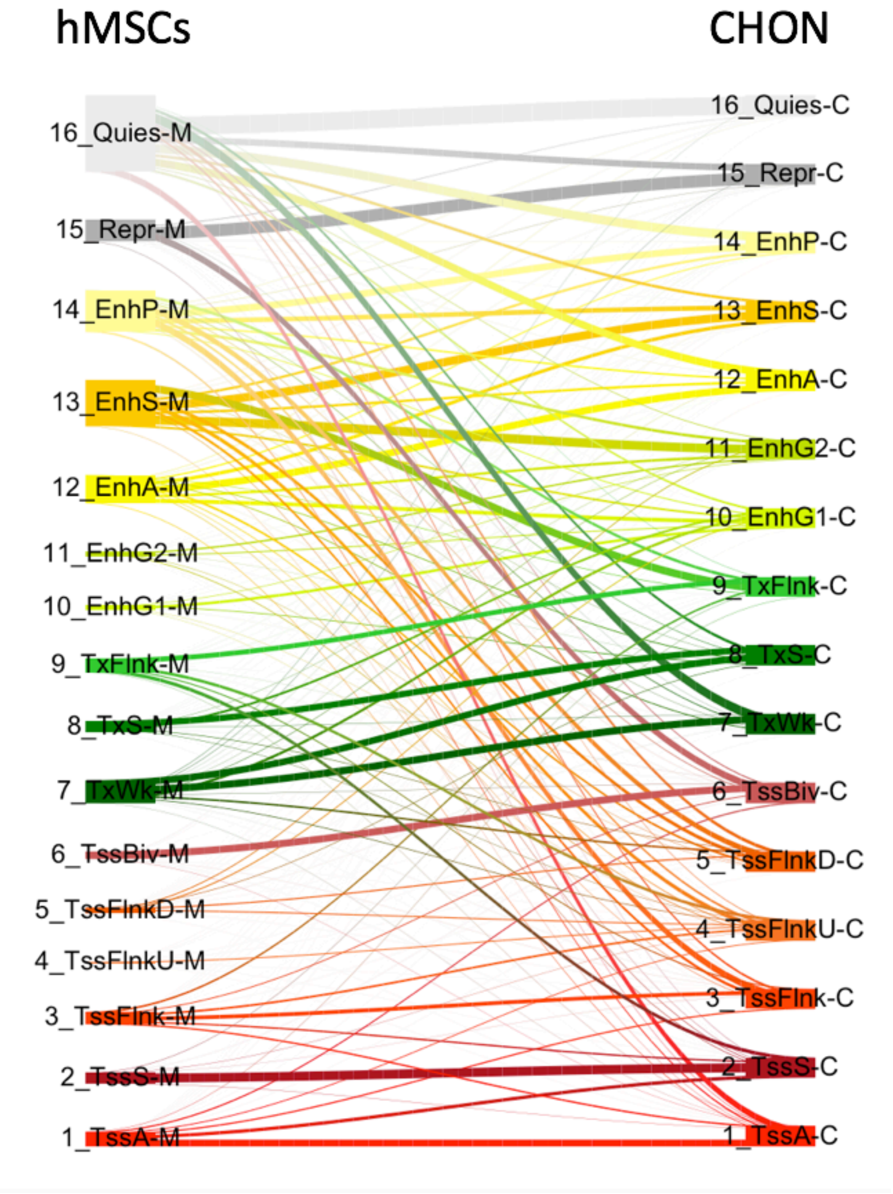


Figure 5.7 – Sankey plot representing chromatin state changes from hMSCs to differentiated chondrocytes. Each node is a chromatin state and the edges show the transition of states from hMSCs during chondrogenesis. Thicker edge lines indicate a higher observed frequency of change. Plot generated using the riverplot package in RStudio.

5.3.3 GO term analysis of chromatin states

Gene ontology analysis was used to characterise chromatin states and connect them to genes and functions. The GREAT ontology tool was used to associate chondrogenesis chromatin states to nearby genes and retrieve GO terms to associated genes. For enhancer states (10_EnhG1, 11_EnhG2, 12_EnhA, 13_EnhS and 14_EnhP), the default GREAT basal plus extension associated method was used. Gene enhancers can be located long distances away from their target gene so it is essential that a larger region is considered. For all other chromatin states, the single nearest gene method was used (up to a maximum distance of 1Mb). The GREAT GO tool was designed to be used with *cis*-regulatory features such as enhancers (McLean et al, 2010). Using the basal plus extension method, enhancers were likely to be associated with multiple genes within the association area compared to the single nearest gene method used for other chromatin states. This method was preferred for enhancers because without other information, it is difficult to confirm which gene is the genuine target of an enhancer and this approach considers all genes in the vicinity. The other chromatin states were associated to genes using the single nearest gene method as they are more likely to be closer to or lie within their target gene.

GO term analysis is a valuable method for grouping genes with similar functions. However, GO term analysis is not suitable or effective when numerous heterogeneous features comprise a dataset. For example, promoter related chromatin states (1_TssA, 2_TssS, 3_TssFlnk, 4_TssFlnkU and 5_TssFlnkD) yielded GO terms largely related to the general upkeep and maintenance of the cell rather than cell type specific terms (Appendix iii Fig. 1-10). These chromatin states cover a broad range of genes and the list is too ambiguous to be able to select GO terms specific to cell type; any cell specific genes are diluted by those related to general function. In contrast, chromatin states with fewer features were associated with more cell type specific GO terms. hMSC bivalent promoter state yielded GO terms associated with differentiation (Appendix iii Fig. 1), indicating that bivalent promoters are linked to genes that are currently non-transcribed but primed to become active depending on the cell lineage the hMSC enters. Interestingly, for the differentiated chondrocyte bivalent promoter state, GO terms were also those related to differentiation, albeit different terms with the exception of one. Both hMSCs and differentiated chondrocytes had a GO term of *dorsal spinal* *cord development*, demonstrating that during differentiation of hMSC some bivalent promoters become active and others remain bivalent after chondrogenesis (Appendix iii Fig. 5.11). Similar to promoter related chromatin states, active transcription states yielded GO terms related to general cell function rather than cell specific functions (Fig. Appendix iii Fig. 12-16). An exception is the hMSC transcription flanking state which returned three GO terms all relating to TGFβ regulation (Appendix iii Fig. 17); TGFβ family members mediate differentiation of hMSCs down different lineages (Roelen and Dijke, 2003). All genic enhancer states (except for chondrocyte genic enhancer 2 (12_EnhG2) had general cell function GO terms. Chondrocyte 11_EnhG2 showed some GO terms related to cartilage function. hMSC active enhancers (12_EnhA) did not yield any obvious stem cell related GO terms although strong enhancers had a GO term of *negative regulation of transforming growth factor receptor signalling pathway*. Similar to chondrocyte 11_EnhG2, chondrocyte enhancer chromatin states returned GO terms related to chondrogenesis, cartilage development and MSC differentiation. Poised chromatin states for hMSC and chondrocytes had GO terms linked to development and differentiation. GO terms for repressed chromatin states were also non-specific, although there were some GO terms associated with development in both hMSCs and differentiated chondrocytes. Quiescent chromatin states encompass the majority of the genome and thus finding distinct genesets and GO terms is more challenging. See Appendix iii for figures.

The locations of chromatin states in relation to the TSS determined by the GREAT tool corroborates with those determined by ChromHMM. Overall, GO term analysis found that enhancer chromatin states are more informative for inferring cell specific processes (Fig. 5.8 – 5.9).

hMSC 13_EnhS

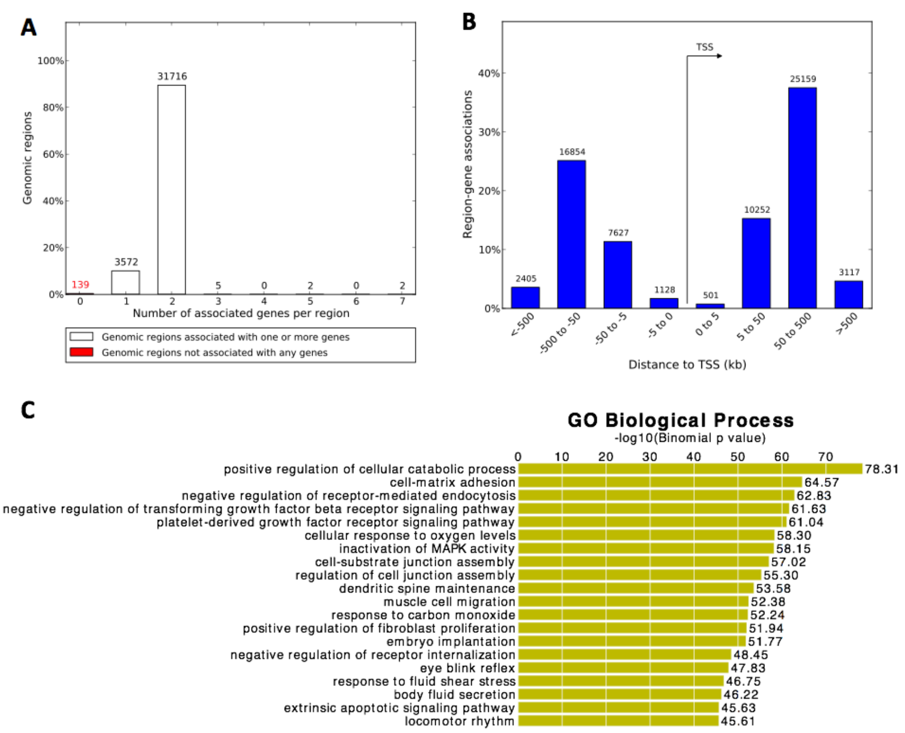


Figure 5.8 - GREAT gene ontology analysis for hMSC 13_EnhS chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 13_EnhS

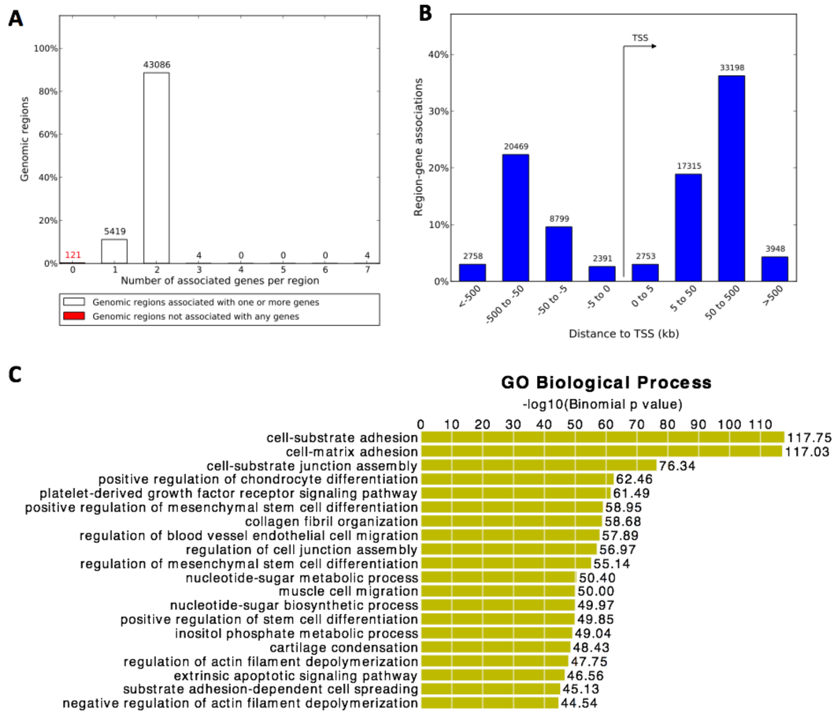


Figure 5.9 - GREAT gene ontology analysis for CHON 13_EnhS chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

5.3.4 Comparison of hMSC and chondrocyte chromatin states with Roadmap Epigenomics chromatin states

A 16 state model was generated from our chondrogenesis histone ChIP-seq data using ChromHMM and chromatin states were annotated using information from the literature, including the Roadmap Epigenomics project. The Roadmap Epigenomics project generated a core 15 state model for 127 human epigenomes using histone modifications H3K4me3, H3K4me1, H3K9me3, H3K27me3 and H3K36me3. An additional histone mark, H3K27ac, was available for 98 epigenomes and an extended 18 state model was trained using this extra mark. Whilst we did not assay H3K9me3, we did include H3K27ac in our ChIP-seq experiment and therefore we sought to compare our chromatin states to the 98 epigenomes with the extended 18 state model. The Roadmap Epigenomics project comprises a wide repertoire of cell types from various cell origins, including primary cells and cell lines. The full list of cells is available in Appendix iii, Table 2. The aim of these comparisons was to ascertain whether similar cell types can be classified as such by their regulatory features in the epigenome. Of particular interest was to compare our hMSCs and chondrocytes to those present in the Roadmap project to observe how alike, or otherwise, the cell types were. The chondrogenesis model included in the Roadmap project also differentiated bone marrow derived hMSC into chondrocytes, albeit using a different *in vitro* model (Herlofsen et al, 2013). We hypothesise that the epigenome of our hMSCs and differentiated chondrocytes would be more similar to the Roadmap hMSCs and chondrocytes respectively relative to the other cell types.

Roadmap and our chondrogenesis chromatin state models were trained on different ChIP-seq samples and Roadmap chromatin states contained an extra histone mark. Therefore, we only compared states that appeared to have the same or similar histone mark emission parameters between the chromatin state models. We compared all equivalent states, of which there were eight (Fig. 5.10).

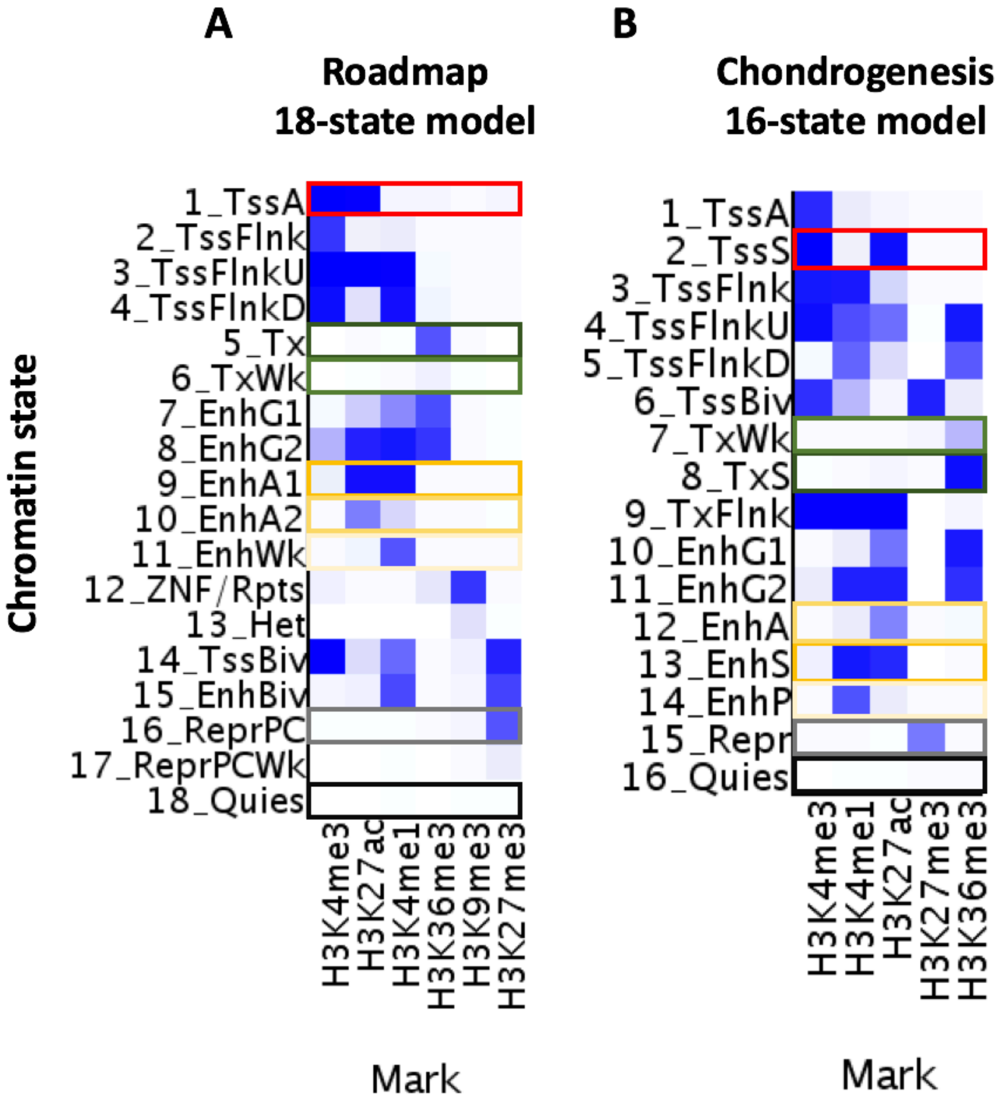


Figure 5.10 - ChromHMM emission parameters for Roadmap's 18 state model (A) and (B) chondrogenesis 16 state model. Roadmap's 1_TssA state and the chondrogenesis 2_TssS state comprises equal probabilities of H3K4me3 and H3K27ac histone marks. Likewise, Roadmap's 9_EnhA1 and chondrogenesis 13_EnhS have similar levels of H3K4me1 and H3K27ac.Other states considered comparable were Roadmap's 5_Tx with chondrogenesis 8_TxS, Roadmap's 6_TxWk with chondrogenesis 7_TxWk, Roadmap's 11_EnhWk with chondrogenesis 14_EnhP, Roadmap's 16_ReprPC with chondrogenesis 15_Repr and Roadmap's 18_Quies with chondrogenesis 16_Quies.

To compare between equivalent Roadmap and chondrogenesis chromatin states, we measured the similarity by calculating the Jaccard coefficient for comparable states across all 98 Roadmap epigenomes, hMSCs and differentiated chondrocytes. Principal component analysis was performed for each comparison. PCA plots provide an effective method of correlating large numbers of samples and visualising related groups or samples. However, with many samples it can be difficult to view individual samples due to overlapping labels in a two dimensional plot. Therefore, we generated heatmaps incorporating hierarchical clustering dendrograms as well as PCA plots.

When promoters were correlated, cells did not separate out by cell origin with the exception of primary cells. ESC derived, cell lines, primary tissue and primary culture cells overlapped forming one large cluster. hMSCs and differentiated chondrocytes are found within this large cluster. Hierarchical clustering shows that hMSCs shared the nearest common node with Roadmap cells E114 (lung carcinoma cell line), E118 (hepatocellular carcinoma cell line), E116 (lymphoblastoid primary culture cells) and E123 (leukaemia primary culture cells). These cells bear no apparent close relation with hMSC respective to other cell types. However, the heatmap of Jaccard values for promoters showed that overall, all cell types are relatively similar. Comparison of Roadmap weak enhancers (11_EnhWk) and chondrogenesis poised enhancers (14_EnhP) displayed greater separation of cells by cell origin with high variation between blood cells. The closest cell type that clusters with our differentiated chondrocytes are Roadmap's chondrocytes (E049). hMSCs form a cluster with Roadmap's BM-MSCs (E026) and osteoblasts (E129; primary culture). Altogether, these cells form a larger subcluster consisting of all the cells of mesenchymal origin in the analysis.

When the similarity of strong transcription state was investigated there appeared to be no apparent clustering of cells by type or cell origin. Hierarchical clustering shows that hMSCs were most similar to some blood (E029 – primary monocytes), skin (E061 – primary cultured melanocytes, E058 – primary cultured keratinocytes) and hESC derived cells when comparing strong transcription (E011 - hESC derived cultured endoderm cells, E013 – hESC derived cultured mesoderm). Chondrocytes were closest to blood (E038 – primary naïve T cells, E123 – primary cultured leukaemia cells) and HeLa cell line (E117). Interestingly, cells clustered by cell origin more by the weak transcription state rather than strong transcription. Furthermore, when considering weak transcription states, hMSCs and differentiated chondrocytes formed their own cluster away from the other cell types. Jaccard similarity values for pairwise comparisons using weak transcriptions were lower than those of strong transcription. See Appendix iii for figures.

A large variation between primary tissue cells was seen for active enhancer comparisons, with brain cells forming their own cluster away from all other cells (Fig. 5.11A) Hierarchical clustering shows that hMSCs and differentiated chondrocytes formed their own cluster away from other cells (Fig. 5.12). Correlation of strong enhancers showed a more distinct separation of cells by type and origin (Fig. 5.11B). The most similar cell type to our differentiated chondrocytes based on strong enhancers were hMSCs and Roadmap's differentiated chondrocytes (E049; Fig. 5.13).

When quiescent states were compared across all epigenomes, there was some clustering of cells by cell origin although there was a large central overlap. Correlation of equivalent repressed states showed no obvious clustering of cell types (Fig. 5.50B). In both quiescent and repressed states, hMSCs and differentiated chondrocytes formed their own cluster (Appendix iii).

Overall, with the exception of the repressed state, more closely related cell types and particularly those from the same cell origin tended to cluster together, albeit loosely with considerable overlaps in some state comparisons. Encouragingly, blood cells from different cell origins were inclined to cluster together away from other cell types in almost all chromatin state comparisons. This demonstrates that cell types can indeed display distinct epigenomes. With respect to chromatin states, poised enhancer (or Roadmap weak enhancers) and strong enhancer states displayed a greater degree of clustering by cell type and origin suggesting these chromatin states are more cell specific compared to others. Furthermore, enhancer states (equivalent poised/weak, active and strong active) displayed lower similarity values across all comparisons demonstrating more cell type specificity. In contrast, repressed and strong transcription states displayed higher similarities compared to other chromatin states.

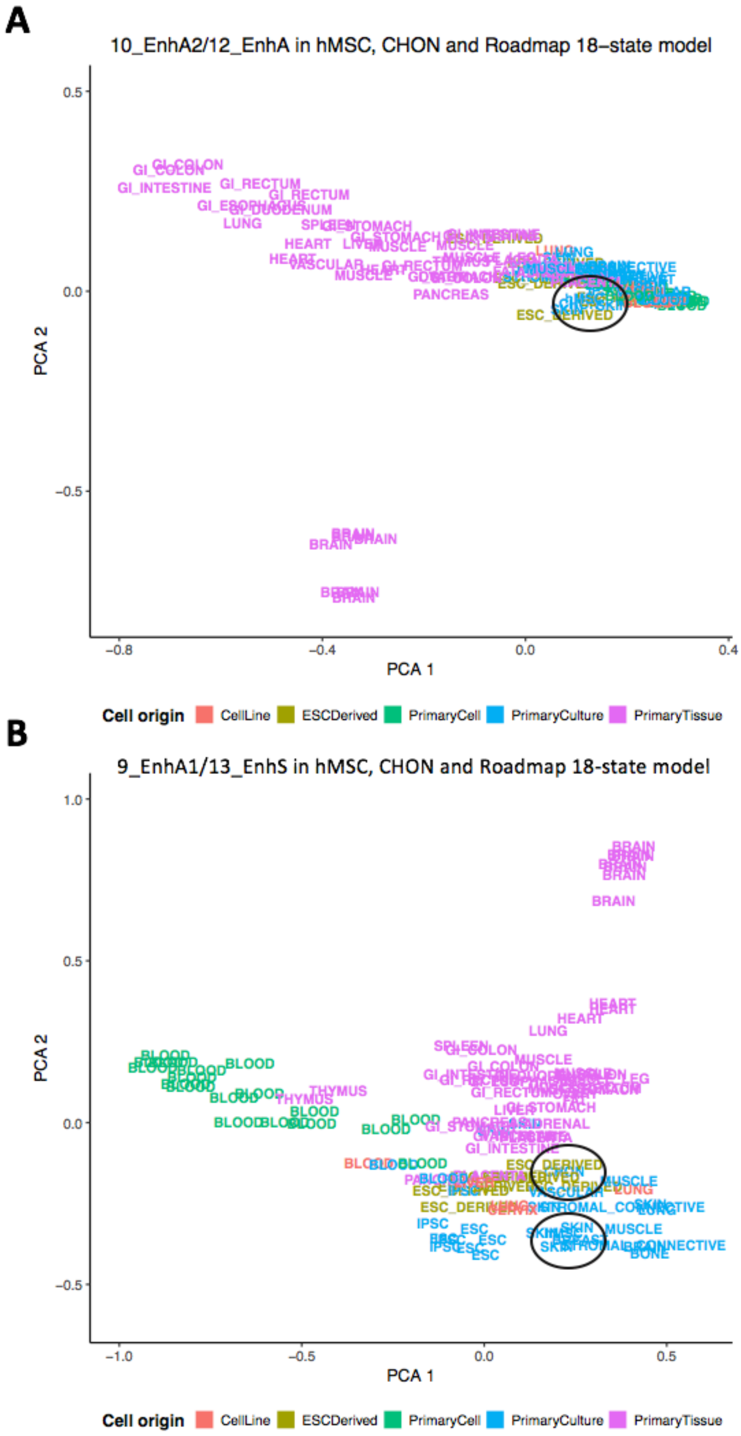


Figure 5.11 - Principal component analysis of equivalent active enhancer states (A) and strong enhancers (B) between Roadmap 18 state model and chondrogenesis 16 state model. hMSC and differentiated chondrocytes are circled on the plot. PCA plot was generated using ggplot2 in RStudio.

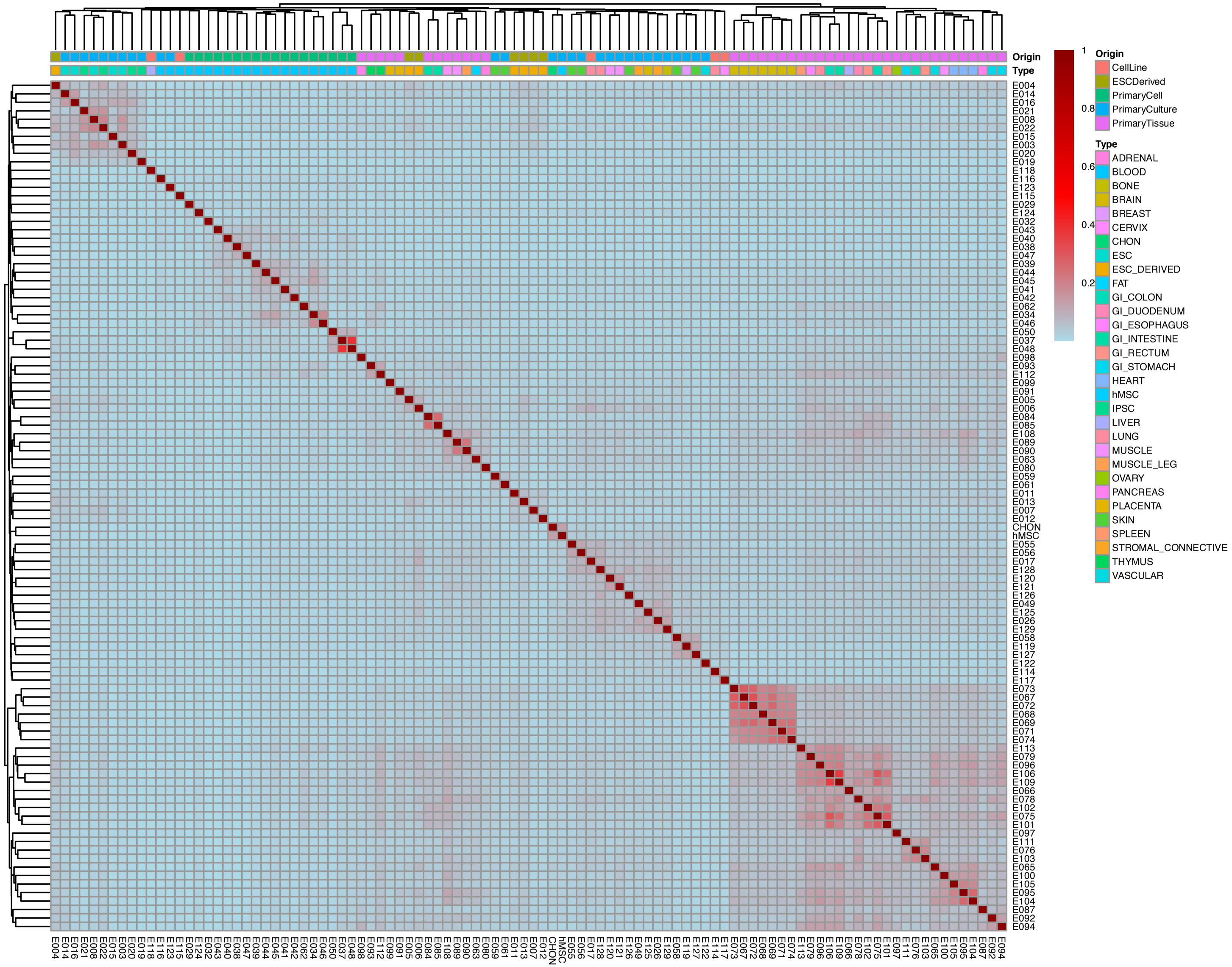


Figure 5.12 - Heatmap and dendrogram of Jaccard index values for Roadmap 10_EnhA2 compared to chondrogenesis 12_EnhA across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.

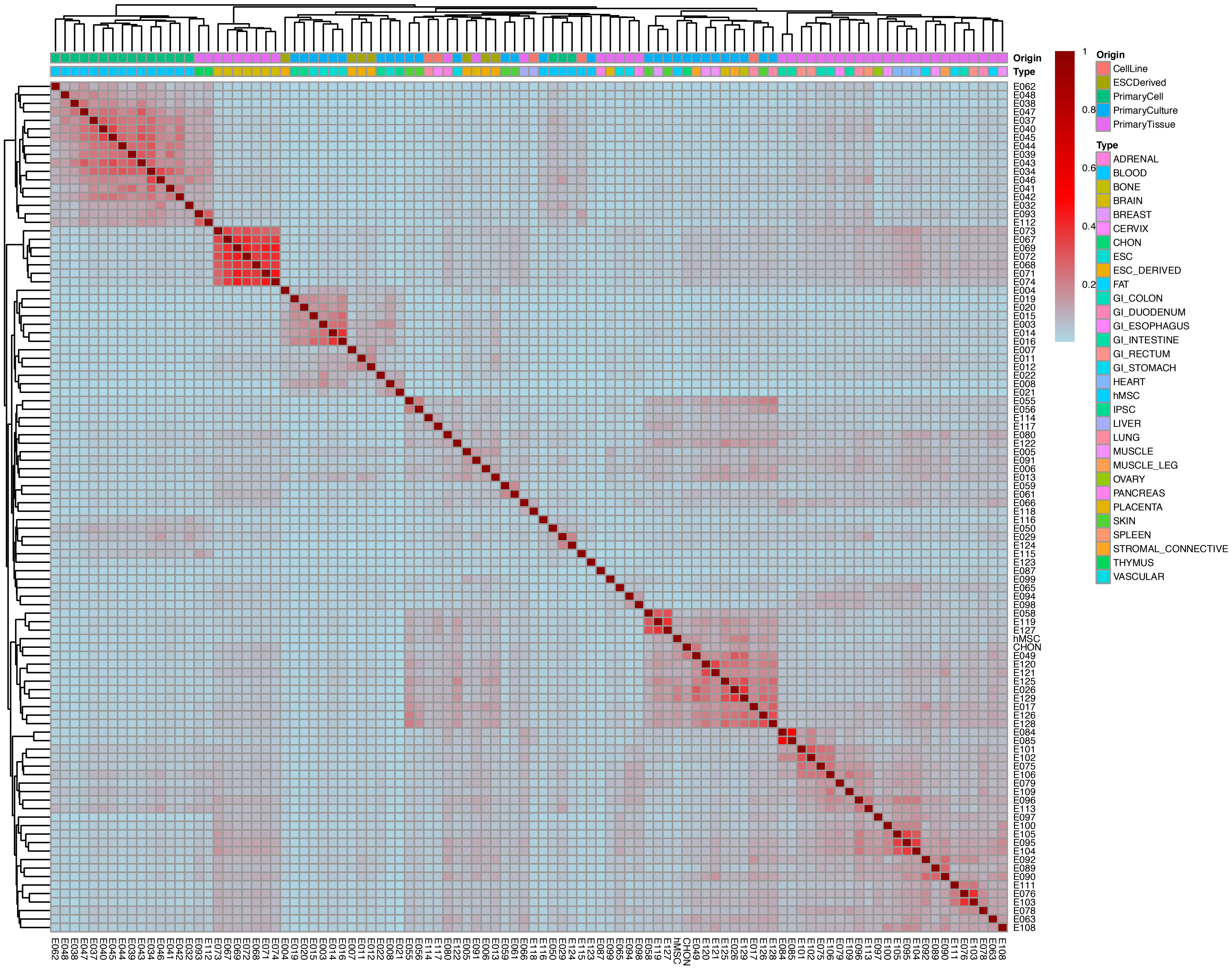


Figure 5.13 - Heatmap and dendrogram of Jaccard index values for Roadmap 9_EnhA1 compared to chondrogenesis 13_EnhS across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.

5.3.5 DNA is hypomethylated during chondrogenesis

DNA methylation of hMSCs (n = 7) and differentiated chondrocytes (n = 5) at CpG sites was quantified using a 450K methylation array. Raw data was kindly provided by Dr. Matt Barter (Newcastle University) who also performed the data collection with Dr. Catherine Bui (Newcastle University).

Differential DNA methylation between hMSCs and differentiated chondrocytes was determined using the Bioconductor package limma after pre-processing and normalising the data using minfi. Normalisation of samples was performed using the functional normalisation method (Hansen et al, 2014). Methylation plots showing samples before and after normalisation are available in Appendix iii, Fig 1-2. Differential DNA methylation tests revealed that 8837 CpG sites were significantly differentially methylated ($q < 0.05$) between hMSCs and differentiated chondrocytes. The vast majority of CpG sites became hypomethylated during chondrogenesis – 94% (8310 CpG sites). After a 10% methylation change filter was applied, the number of significant CpGs dropped to 6601, of which 97.7% were hypomethylated.

DNA methylation is usually linked to transcriptional repression; the hypomethylation of DNA during chondrogenesis indicates a large scale transcriptional activation of genes during this process. GO term analysis using missMethyl of genes linked to significantly differentially methylated CpG sites showed that these genes were associated with extracellular matrix organisation and extracellular structure organisation (Table 5.1). In a 450k array, some genes have more than one probe associated with them and therefore may be overrepresented. The GO term analysis tool from the missMethyl Bioconductor package was preferred because it normalises for multiple CpG probes linked to the same gene (Phipson et al, 2016).

Table 5.1 - GO term analysis of genes associated with the top 500 most hypomethylated CpGs (q < 0.05). GO terms were found using the missMethyl Bioconductor package.

GO term	**GO term name**	**FDR**	**Number of genes**
GO:0030198	Extracellular matrix organisation	0.0067	325
GO:0043062	Extracellular structure organisation	0.0067	326

We next sought to determine the methylation level of chromatin states during chondrogenesis and where significantly hypomethylated CpGs are located.

5.3.6 Correlation of DNA methylation to chromatin states

Chromatin states for hMSCs and differentiated chondrocytes were generated previously using ChromHMM (Fig. 5.1). The genomic co-ordinates of CpG dinucleotides were intersected with chondrogenesis chromatin states to determine where DNA methylation occurred in the genome. CpGs sites assayed using 450K array in chondrocytes were intersected with chromatin states in chondrocytes and hMSCs (Fig. 5.14; Fig. 5.15). Promoter and promoter flanking chromatin states displayed a lower overall methylation level than transcription, enhancer, repressed and quiescent states. To directly compare methylation levels, the methylation levels of the same CpGs in chondrocytes prior to differentiation were plotted (Fig. 5.14). Likewise, methylation of CpGs in hMSC chromatin states was plotted and for a direct comparison, the methylation levels of the same CpGs in chondrocytes were investigated (Fig. 15).

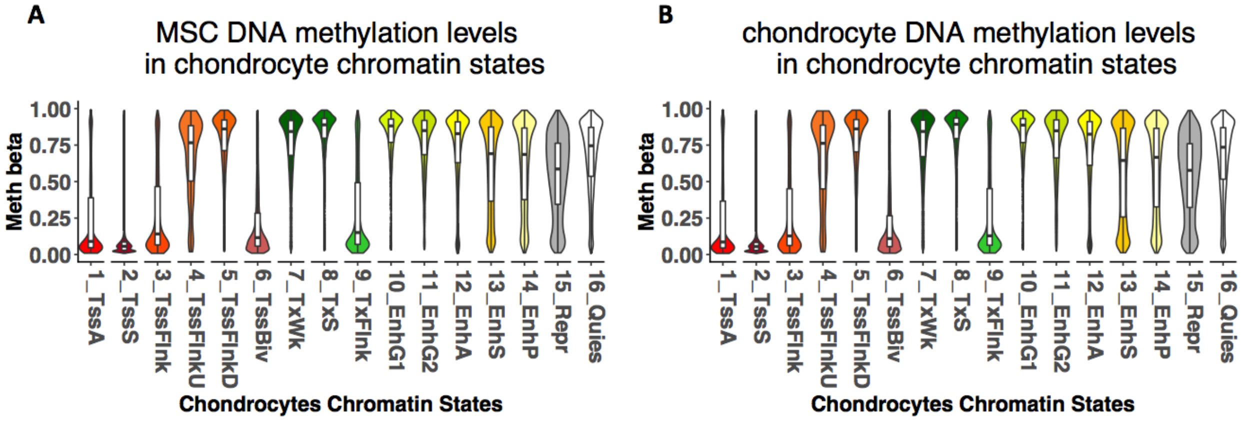


Figure 5.14 – Violin plots showing the percentage of methylation (beta values) of CpGs differentiated chondrocyte chromatin states and prior to differentiation. DNA methylation levels of all CpGs located in hMSCs (B) and the same CpGs in differentiated chondrocytes (A). BEDTools intersect was used to find chromatin states of CpG sites. Violin plots generated using ggplot2 in RStudio. The width of each violin is the density of values and the box inside the violin is the interquartile range, the horizontal line within the box is the median value and the vertical line through the violin is the 95% confidence interval.

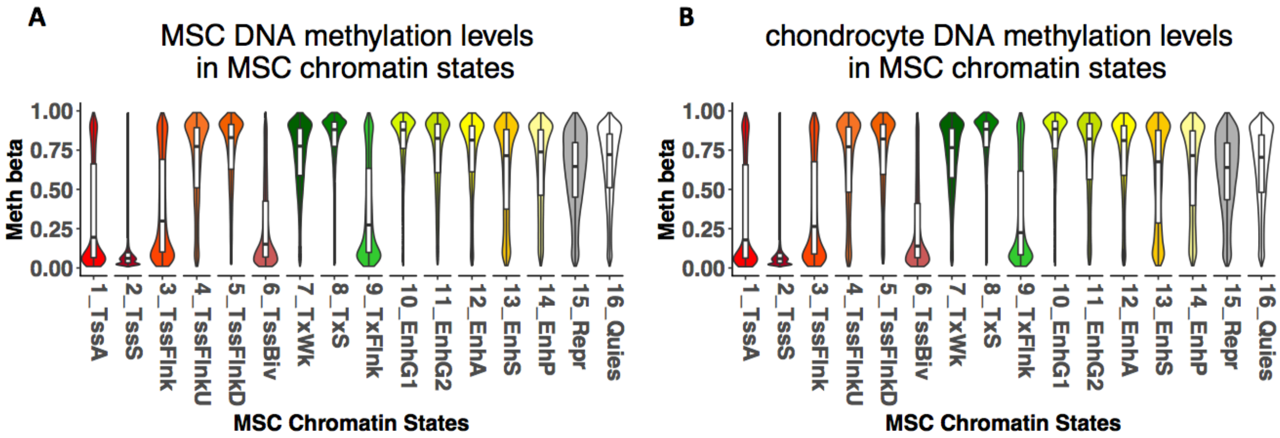


Figure 5.15 - DNA methylation levels of CpGs in hMSC chromatin states (A) and the same CpGs in differentiated chondrocytes (B).

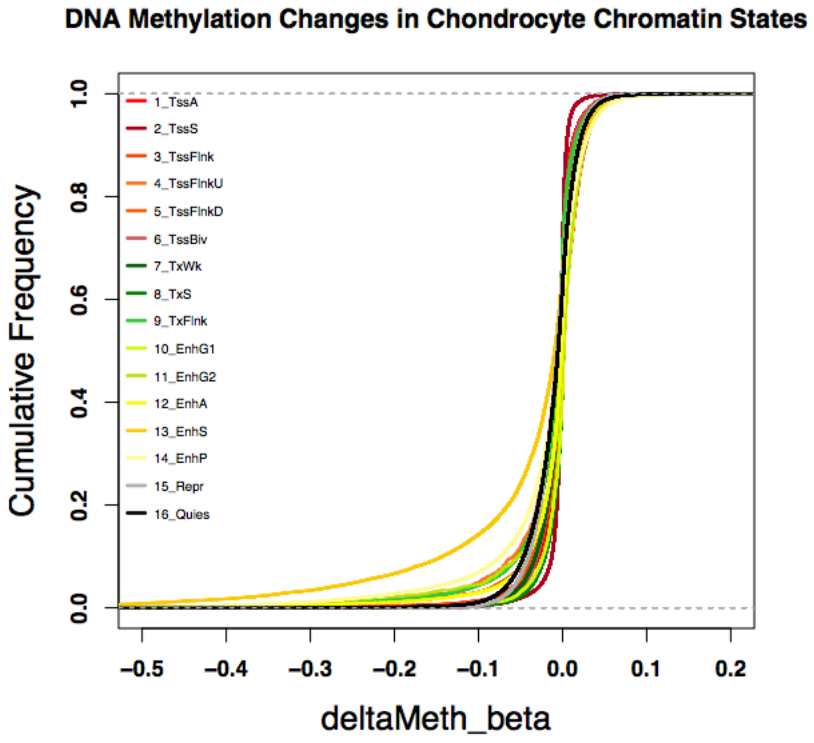


Figure 5.16 – Empirical cumulative distribution frequency plot of all DNA methylation changes in chondrocyte chromatin states. All CpG sites in the 450k array were intersected with chondrocyte chromatin states. The changes in methylation (∆β) of CpG sites within each state was plotted as an empirical cumulative frequency curve using ggplot2 in RStudio.

There was no obvious difference between methylation in chromatin states during chondrogenesis except the strong enhancer state (13_EnhS), where a tendency towards hypomethylation in chondrocytes was observed. An empirical cumulative distribution frequency plot showed that state with the greatest change in hypomethylation (mean -0.04) in differentiated chondrocytes was 13_EnhS (Fig. 5.16; Appendix iii, Table 5). The difference becomes more evident when CpGs in chondrocyte 13_EnhS are plotted side by side with their methylation levels in hMSC (Fig. 5.56A) and when only significant (*q* < 0.05) CpGs are considered (Fig. 5.17B). Methylation levels of significantly methylated CpGs in the chondrocyte strong enhancer state is ~25%, the methylation level of the same CpGs in hMSCs is over 50%.

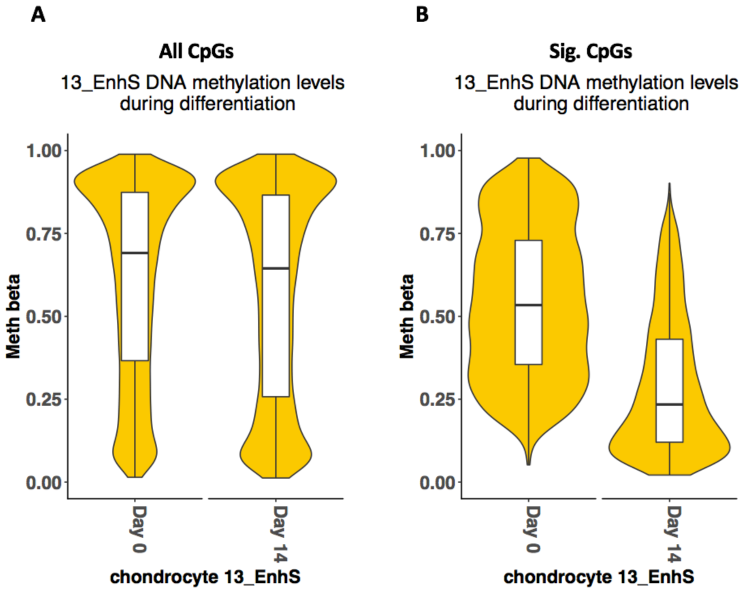


Figure 5.17 – DNA CpG methylation changes intersected with the day 14 chondrocyte strong enhancer state (13_EnhS). Violin plot showing the percentage methylation (β value) of CpGs at day 0 (hMSCs) and day 14 (chondrocytes). (A) All CpGs and (B) significant (q < 0.05) differentially methylated CpGs only.

5.3.7 Strong enhancers are hypomethylated during chondrogenesis

We observed that CpGs located in the chondrocyte strong enhancer state (13_EnhS) saw a greater change in hypomethylation compared to other chromatin states during chondrogenesis. We next sought to determine whether hypomethylation of the strong enhancer state could be explained by the number of CpG probes corresponding to this state by investigating the locations of CpG probes on the 450K array. In chondrocytes, most CpG probes on the array were located in repressed, quiescent and active promoter states (Fig. 5.18B). Only 4.96% of total CpG probes were located in the chondrocyte strong enhancer chromatin state yet 41.8% of significant hypomethylated CpGs were found in this state (Fig. 5.18A). This demonstrates that significant hypomethylated CpGs were indeed disproportionately located in the strong enhancer state.

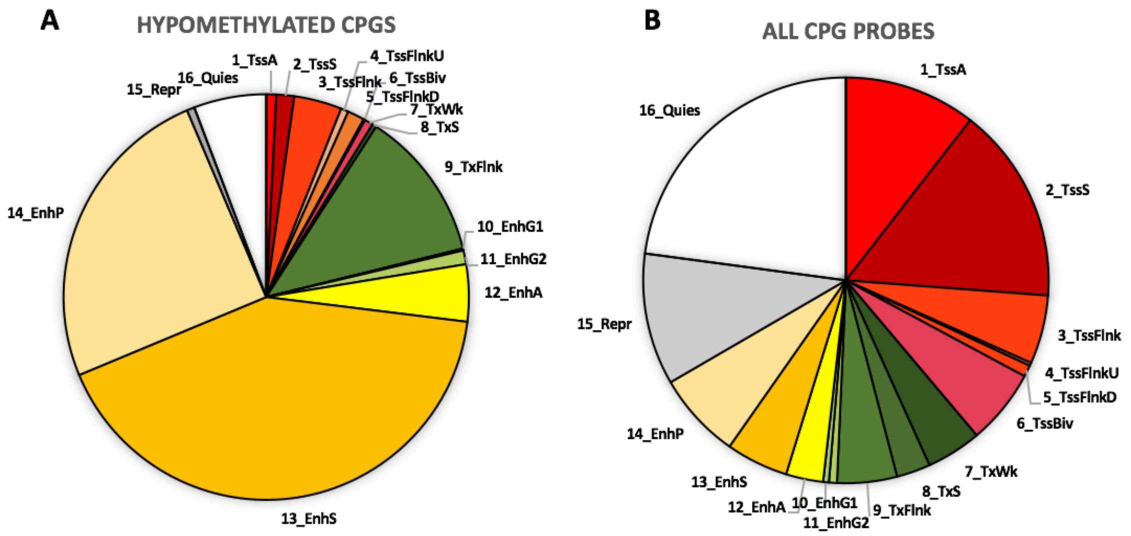
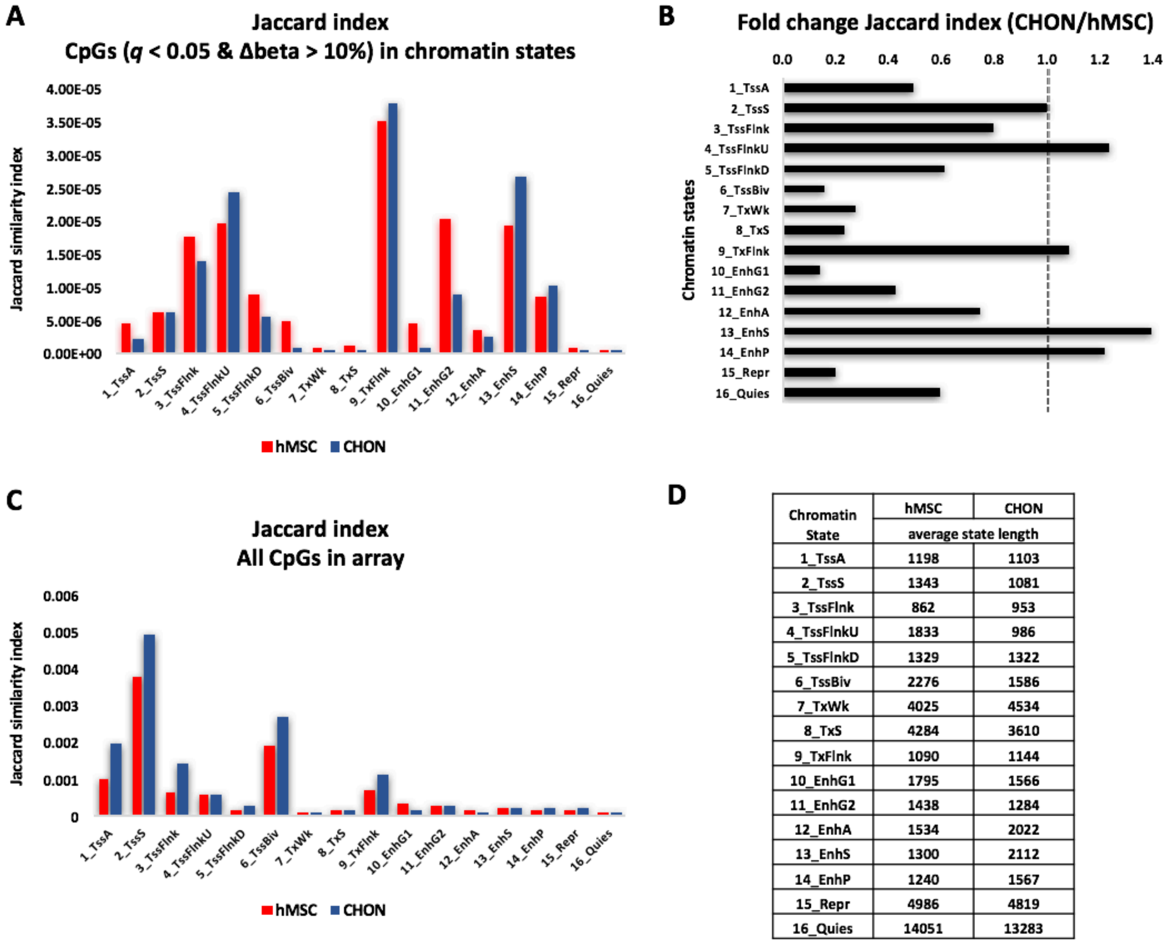


Figure 5.18 - Locations of significant hypomethylated CpGs (q < 0.05 & > 10% methylation change) in (A) chondrocyte chromatin states and (B) all CpGs. CpG co-ordinates were intersected with chondrocyte chromatin states using BEDTools intersect and pie charts were plotted using Microsoft Excel.

Although the analysis above shows that strong enhancers contain more hypomethylated CpGs relative to total CpG probes on the array, this method does not take into account the size of chromatin state. Assuming CpG probes are randomly placed across the genome, more total probes on the array and significant differentially methylated CpGs sites would fall into longer chromatin states by chance. Therefore, the Jaccard similarity index was also employed to calculate which chromatin state showed the most hypomethylation. This considers the size of the union between chromatin states and CpG sites and thus the size of the chromatin state. Higher Jaccard values indicate greater numbers of CpG sites in a chromatin state, taking into account the chromatin state size.



Chromatin State	hMSC	CHON
	average state length	
1_TssA	1198	1103
2_TssS	1343	1081
3_TssFlnk	862	953
4_TssFlnkU	1833	986
5_TssFlnkD	1329	1322
6_TssBiv	2276	1586
7_TxWk	4025	4534
8_TxS	4284	3610
9_TxFlnk	1090	1144
10_EnhG1	1795	1566
11_EnhG2	1438	1284
12_EnhA	1534	2022
13_EnhS	1300	2112
14_EnhP	1240	1567
15_Repr	4986	4819
16_Quies	14051	13283

Figure 5.19 - Jaccard similarity index was used to determine relative methylation levels of hypomethylated CpGs in each chondrogenesis chromatin state. (A) Significantly hypomethylated CpGs intersected with hMSC and differentiated chondrocyte chromatin states, (B) change in Jaccard index of significant CpGs, (C) Jaccard values of all CpG sites intersected with hMSC and chondrocyte chromatin states and (D) average size of chromatin states.

Overall Jaccard index values were small; this is due to the 2bp genome co-ordinates of CpG dinucleotides leading to small intersection values. Promoter flanking, transcription flanking and strong enhancer states displayed higher Jaccard values with significant differentially methylated CpG sites compared to other chromatin states (Fig 5.19A); these states contain more significant CpG sites relative to chromatin state size. The chondrocyte strong enhancer state showed the greatest increase in Jaccard value and therefore hypomethylated CpGs compared to strong enhancer states in hMSCs (Fig. 5.19B), i.e. there are more significantly hypomethylated CpGs in chondrocyte strong enhancers than hMSC strong enhancers. When all CpG probes were intersected with hMSC and differentiated chondrocyte chromatin states, the strong promoter state had the highest Jaccard value (Fig. 5.19C). In the previous analysis, quiescent and repressed states showed a high proportion of total CpG probes with the quiescent state containing roughly a quarter of total probes. However, the Jaccard values for total CpG probes and these chromatin states were among the lowest due to much larger chromatin state sizes (Fig. 5.19D). To summarise, integration of histone ChIP-seq and DNA methylation data suggests that hypomethylation of strong enhancers plays a role in driving chondrogenesis.

5.3.8 Changes in chromatin state and DNA methylation during chondrogenesis are linked

Previously we found that strong enhancers in chondrocytes are disproportionately significantly hypomethylated. The numbers of significant hypomethylated CpG sites and total CpG sites remain the same, yet we see different Jaccard values between hMSC and differentiated chondrocyte chromatin states and CpG sites (Fig. 5.19A; Fig. 5.19C). This illustrates a shifting of chromatin states between hMSCs and differentiated chondrocytes which may coincide with DNA methylation changes. Of the total CpG sites present in the array, 52.5% were located in chromatin states that remained the same during chondrogenesis. In contrast, only 28.5% of significantly hypomethylated CpGs ($q < 0.05$ % 10% $\Delta\beta$) were found in chromatin states that did not change. The overwhelming majority of hypomethylated CpGs (71.5%) were found in chromatin states that change during chondrogenesis. A Chi-square test showed that the distribution of all CpGs and hypomethylated CpGs was not the same between a) chromatin states that remained the same and b) chromatin states that changed during chondrogenesis ($\chi^2 = 1479$, df = 1, p-value = 1.48×10^{-323}; Table 5.2). This suggests that the hypomethylation of CpG sites and changing chromatin state are linked.

Table 5.2 – All CpGs and hypomethylated CpGs ($q < 0.05$ & 10% $\Delta\beta$) during chondrogenesis in chondrocyte chromatin states that remain the same and chromatin states that change. Expected values in brackets next to observed values

	Total	**Chromatin state remains the same**	**Chromatin state changes**
All CpGs	485438	255048 *(253516)*	230390 *(231922)*
Hypomethylated CpGs	6448	1835 *(3367.409)*	4613 (*3080.591*)

We previously deduced that in differentiated chondrocytes, the strong enhancer state contained 41.8% of hypomethylated CpGs (Fig. 5.20). We were interested to investigate the chromatin state changes of these CpG sites to infer how demethylation influences chromatin state shifting in chondrogenesis.

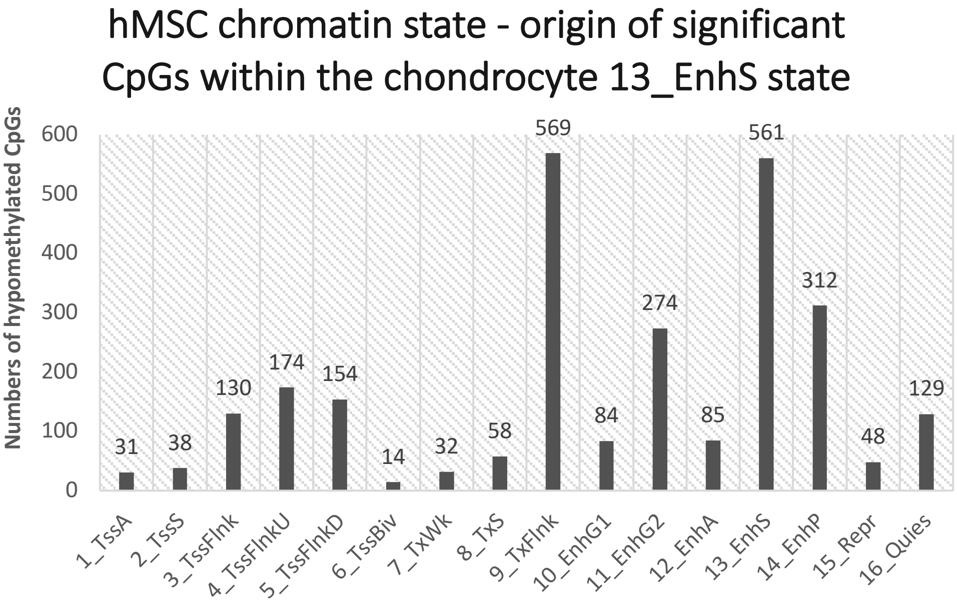


Figure 5.20 - Chromatin states prior to differentiation of hypomethylated CpG sites within chondrocyte 13_EnhS

Chondrocyte strong enhancer states containing hypomethylated CpG sites originate from a range of chromatin states; nearly 80% originate from chromatin states other than strong enhancer with the largest number originating from the transcription flanking state (Fig. 5.59).

Our integrated data shows that shifts in chromatin states during chondrogenesis are accompanied by DNA methylation changes. Whilst we have shown an explicit association between histone modification changes and DNA methylation leading to changes in chromatin state, the direction of these changes is less clear. From our data, it is not possible to deduce whether DNA methylation causes a change in chromatin state or whether histone modifications cause DNA methylation changes.

5.3.9 DNA methylation during adipogenesis and osteoblastogenesis

Following the discovery that H3K27ac and H3K4me1 marked enhancers are hypomethylated during chondrogenesis, DNA methylation during MSC adipogenesis and osteoblastogenesis was investigated in order to establish whether DNA methylation changes were specific to chondrogenesis. DNA 450k methylation arrays were carried out for hMSCs and differentiated adipocytes at 14 days (n = 3) and differentiated osteoblasts at 21 days (n = 3). Differentiation experiments and data collection was carried out by Dr. Ruddy Gomez-Bahamonde. Analysis of the data was performed as for chondrogenesis (functional normalisation in minfi and differential methylation tests using limma). During osteoblastogenesis, there were 2337 significantly differentially methylated CpGs (q < 0.05), of which 2151 were hypomethylated and 186 were hypermethylated. Only 2 CpGs were found to be significantly differentially methylated during adipogenesis and these were hypomethylated. GO terms associated with differentially methylated CpGs during osteoblastogenesis were retrieved using missMethyl (Table 5.6). There were 14 significant GO terms (q < 0.05) found by missMethyl, of which 4 relates to ECM organization and 10 to cell movement. Like cartilage, bone is dense in ECM which provides tissue integrity and a scaffold for cells. The bone ECM is produced by osteoblasts and is composed of different proteins to cartilage and the primary collagen is type I rather than type II (Alford et al, 2015). During osteoblastogenesis, osteoblast precursors migrate to the site of bone formation (Thiel et al, 2017). GO terms reveal that differential DNA methylation occurs at genes responsible for ECM production and cell migration.

Adipogenesis did not yield enough significantly differentially methylated CpGs for GO term analysis. The top 500 most hypomethylated CpGs by beta values (non significant) did not yield any significant GO terms nor did the 500 most hypermethylated CpGs.

Histone modifications were not assayed during adipogenesis or osteoblastogenesis experiments and therefore chromatin states from the Roadmap 18 state model were used in this analysis. Roadmap primary osteoblast (cell type E129) chromatin states were used to determine methylated states during osteoblastogenesis. For adipogenesis, Roadmap primary adipose nuclei (E063) chromatin states were used.

Table 5.3 – GO terms for significantly differentially methylated CpG sites (q < 0.05) during osteoblastogenesis found by missMethyl

GO term	GO term name	FDR
GO:0016477	cell migration	0.000320897
GO:0005578	proteinaceous extracellular matrix	0.000320897
GO:0030198	extracellular matrix organization	0.000482797
GO:0043062	extracellular structure organization	0.000482797
GO:0048870	cell motility	0.000582142
GO:0051674	localization of cell	0.000582142
GO:0040011	locomotion	0.000921287
GO:0006928	movement of cell or subcellular component	0.005619566
GO:0030334	regulation of cell migration	0.005619566
GO:2000145	regulation of cell motility	0.008659225
GO:0051270	regulation of cellular component movement	0.009013122
GO:0051716	cellular response to stimulus	0.019948436
GO:0040012	regulation of locomotion	0.021166516
GO:0009888	tissue development	0.027323105

The most hypomethylated chromatin state in osteoblasts was 9_EnhA1 (equivalent to 13_EnhS in the chondrogenesis 16 state model), with a mean $\Delta\beta$ of -0.018 (Fig. 5.21; Appendix iii, Table 6). In contrast, no adipocyte chromatin states showed a net hypomethylation during adipogenesis (Fig. 5.22; Appendix iii, Table 7). This suggests that osteoblast development, like chondrocyte development, is regulated by the de-methylation of DNA in enhancers. However, DNA methylation does not appear to change during adipogenesis. This is corroborated by a separate study (van den Dungen et al, 2016). One caveat with this analysis is the use of chromatin states not generated from the cell types used to assay DNA methylation. The histone modifications profile of Roadmap primary osteoblasts and primary adipocytes may be different to *in vitro* differentiated osteoblasts and adipocytes.

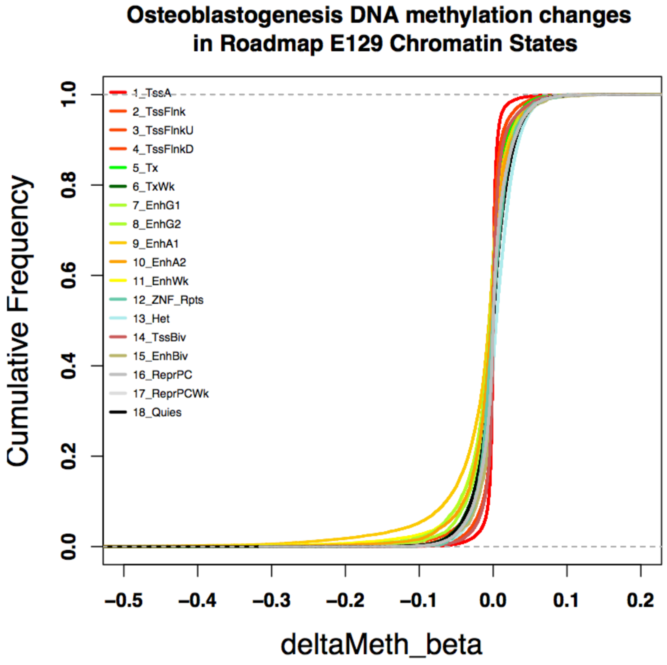


Figure 5.21 – Empirical cumulative distribution frequency plot of DNA methylation changes during osteoblastogenesis within osteoblast (Roadmap E129) chromatin states. All CpG sites were intersected with E129 chromatin states. Plot generated using ggplot2 in RStudio.

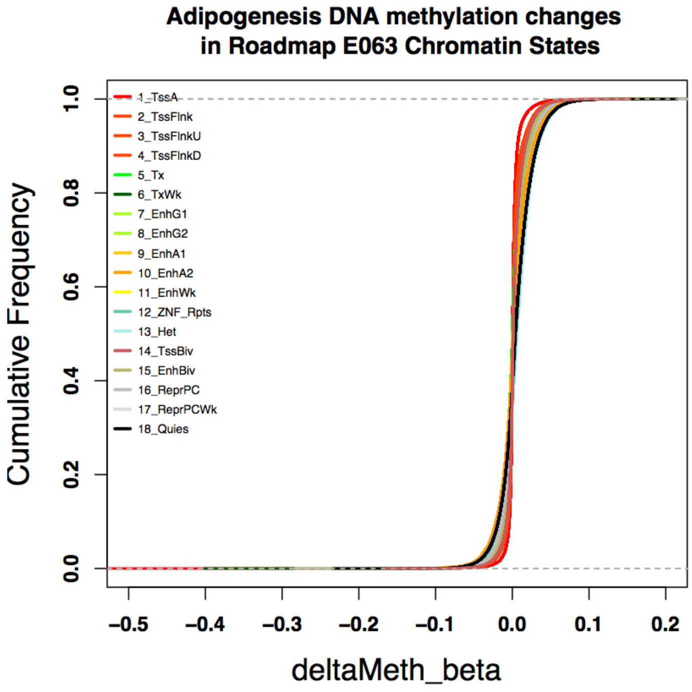


Figure 5.22 - Empirical cumulative distribution frequency plot of DNA methylation changes during adipogenesis within adipocyte (Roadmap E063) chromatin states. All CpG sites were intersected with E063 chromatin states. Plot generated using ggplot2 in RStudio.

5.4 Discussion

Characterisation of chromatin states allows us to view combinations of histone modifications and their co-occurrences in the genome. This approach yields more information than inspecting single histone marks. From the five histone modifications assayed in our chondrogenesis ChIP-seq experiments, 16 chromatin states were characterised including promoter states denoting gene TSS and enhancer regions. We generated two states designated genic enhancers (7_EnhG1 and 8_EnhG2). These consist of enhancer marks alongside H3K36me3. Further investigation is needed to determine whether these states function in a different manner to conventional enhancer states 12_EnhA and 13_EnhS. The Roadmap project also identified genic enhancers but no further information on their function is available. Some chromatin states were more prevalent in the genome than others. The relatively high percentage of quiescent or unmarked state may reflect the low percentage of the genome that is transcribed at any given time. This abundance of quiescent states is also corroborated by the equivalent quiescent state in Roadmap's extended 18 state model (Kundaje et al, 2015). It is thought that only 2% of the human genome codes for proteins and ~5-10% is transcribed (Pertea, 2012). Therefore, the majority of the genome is in an inactive or untranscribed state. Conversely, the 16_Quies state may not be quiescent at all and may be marked with other histone modifications not included in our comparatively limited study of five histone marks. There are numerous histone modifications other than the marks included in our study. Furthermore, the modifications included in our ChIP-seq experiment, and those used to generate chromatin states in the Roadmap project, are only on histone H3. Considering there are three other core histone proteins (H2A, H2B, H4) comprising the nucleosome, our project and indeed even the Roadmap project, only provide a small snapshot into how histone modifications influence the epigenome. A much larger panel of histone modifications is required for a truly comprehensive study of the epigenome.

ChromHMM relies on the user to annotate chromatin state names. Chondrogenesis chromatin state names were designated based on information from Roadmap and ENCODE projects. Where possible, chromatin state names and abbreviations match Roadmap's extended 18 state model. An exception is Roadmap's 11_EnhWk and chondrogenesis 14_EnhP. These two states both exhibit a high level of H3K4me1 enrichment and low enrichment for other marks. Whilst Roadmap have termed this a weak enhancer state, we decided to designate this state a poised enhancer. Studies have shown that H3K27ac marks are required to distinguish poised from active enhancers (Creyghton et al, 2010: Heinz et al, 2015). Therefore, a high H3K4me1 and low H3K27ac enrichment profile seen in chondrogenesis 14_EnhP and Roadmap's 11_EnhWk states corresponds to a histone mark profile more akin to poised enhancers than active enhancers. Chromatin state names were supported by overlap and neighbourhood enrichment probabilities determined by ChromHMM. ChromHMM also outputs transition probabilities of chromatin states which is the likelihood of states (in 200bp bins) transitioning to another adjacent state or remaining the same. This pattern of transitioning states is symbolic of the spreading mechanism employed for the addition of post-translational modifications to histone proteins (Ernst and Kellis, 2010); histone modifications are added to specific amino acids on histone tails, more of the same modification is then added to neighbouring nucleosomes in both directions (Schlissel, 2004). The sequential addition of the histone mark continues until blocked by an insulator or boundary element (Ferrari et al, 2004; Weth et al, 2014). Insulators can also block the interaction between enhancers and gene promoters (Zhao and Dean, 2004). The transition probabilities for the chondrogenesis quiescent state indicates that apart from itself, it is likely to be adjacent to chromatin states that are not within a gene body. This is indicative of boundary elements within the quiescent state that stop the spreading of histone modifications from one gene into another. Boundary elements also lead to the formation of distinctive chromatin states by ensuring histone modifications do not spread along the genome unchecked.

GO term analysis of genes associated with individual chromatin states reveals different functional characteristics of some chromatin states between hMSCs and differentiated chondrocytes. Enhancer chromatin states yielded more cell type specific biological functions than other states, suggesting that enhancers more than any other region regulate cell specific gene transcription during chondrogenesis. This finding fits in line with other studies that have found that gene enhancers drive cell specific gene expression, usually mediated by transcription factor binding (Heinz et al, 2015). Further information is required to establish whether our chondrogenic enhancers contain binding sites for transcription factors or other DNA binding proteins.

To correlate chromatin states derived from our chondrogenesis ChIP-seq and the Roadmap project, we calculated the Jaccard similarity coefficient between comparable states. There were eight chromatin states between the Roadmap's extended 18 state model and our chondrogenesis model that were considered equivalent i.e. displayed similar emission parameters for the same histone modifications. Chromatin state correlations between cells showed that enhancers are more distinct between cell types and cell origins to a greater extent than other chromatin states. One of the aims of this analysis was to assess the similarity of the epigenomes of different cell types. We hoped to see the same cell types from different cell origins cluster together, for example the same cell type from primary cells and from cell lines. The use of cultured cells and cell lines as a proxy for cells *in vivo* is widespread in research. Therefore, it is important to investigate whether immortalising or culturing cells changes their epigenomic make up and to what extent. Epigenetic marks can be influenced by environmental factors and therefore variability in the epigenome could be introduced by culture conditions (McEwen et al, 2013) and number of passages (Noer et al, 2009).

With the exception of blood cells, most other cell types included in the Roadmap Epigenomics project only derived from one cell origin. Consequently, our analysis only reliably shows the epigenomic relationship between cells of different cell origins rather than type. Our hMSC-derived chondrocytes cluster with Roadmap's hMSC-derived chondrocytes on the basis on enhancers marked by H3K27ac and H3K4me1. This indicates that although the chondrocytes originate from different chondrogenesis models, research labs and ChIP-seq assays, they are nonetheless more similar to each other compared to other cell types. This gave us confidence that the disparate chondrogenesis models yielded differentiated chondrocytes that have similar epigenomes, at least when comparing gene enhancers. The chondrogenesis model from which Roadmap's differentiated chondrocytes originated from involved the use of a 3D Alginate scaffold whereas our chondrocytes were differentiated in scaffold-free transwell cultures (Murdoch et al, 2007). Furthermore, we differentiated chondrocytes over 14 days whereas Roadmap's chondrocytes were differentiated over 7 days (Herlofsen et al, 2013). Both chondrogenesis models utilise cultured primary hMSCs and therefore the *in vitro* chondrogenic environment is artificial. From our analysis, we can conclude that both chondrocytes from these two models have a similar enhancer profile which may be chondrocyte specific but without further data it is unknown how similar they are to primary human articular chondrocytes.

The Jaccard index relies on overlapping genomic coordinates to determine similarity. Consequently, even if a chromatin state does not overlap by 1bp in another dataset it will not be taken into account. Considering the plasticity and variability of the epigenome, not to mention the noise that is inherently present in biological data, it may be reasonable to recognize that chromatin states between cell types can still be synonymous if they are only a few bases apart as well as directly overlapping. The Jaccard index cannot compare chromatin states if they do not physically overlap so does not allow for any flexibility. Furthermore, similarity in our analysis is a relative term and it is unknown exactly how similar epigenomes of cell types are expected to be. Additionally, whilst we matched chromatin states between Roadmap and our chondrogenesis data, the two models are not the same and equivalent states are not identical. When comparing weak transcription (Roadmap 6_TxWk and chondrogenesis 7_TxWk), active enhancers (Roadmap 10_EnhA2 and chondrogenesis 12_EnhA), quiescent (Roadmap 18_Quies and chondrogenesis 16_Quies) and repressed (Roadmap 16_ReprPC and 15_Repr) states, our hMSCs and chondrocytes formed their own cluster away from the Roadmap cell types. This could be due to a true difference in the epigenome between these cells and Roadmap's cells or it may simply be due to a mismatch between the chromatin states from Roadmap and chondrogenesis ChromHMM models, resulting in segregation of chondrogenesis samples.

In this chapter, we have explored the correlation between DNA methylation and histone modification. Whilst the DNA 450K array data was re-analysed for the purpose of this project, a full in depth analysis is beyond the scope of this PhD and is the focus of another project. We found that DNA is hypomethylated during chondrogenesis and that these hypomethylated CpG sites mainly occur in the strong enhancer state in chondrocytes. The same study that generated Roadmap's differentiated chondrocyte chromatin states also assayed DNA methylation using reduced representation bisulfite sequencing (RRBS). Hypomethylation at CpG site dense regions during chondrogenesis was observed but DNA methylation at gene promoters was not found to correlate with gene expression (Herlofsen et al, 2013). Unfortunately, DNA methylation at enhancers was not investigated. Furthermore, RRBS may not be an appropriate method to use to assay enhancer DNA methylation as it only measured DNA methylation at CpG dense regions which enhancers typically are not (Calo and Whysocka, 2013). In this chapter, we also found that DNA de-methylation occurs in enhancers during osteoblastogenesis but not adipogenesis. This suggests that DNA hypomethylation in enhancers may regulate enhancers in some differentiation processes but not others. DNA hypomethylation at enhancers also plays a role during intestinal stem cell differentiation into enterocytes (Sheaffer et al, 2014). Enhancer hypomethylation was also found to be important in T cell differentiation (Schmidl et al, 2009). A study found that DNA hypomethylation occurring at transposable elements in the genome led to the gain of enhancer markers to these regions in a tissue specific manner (Xie et al, 2013). This suggests that DNA hypomethylation at enhancer regions play a role in driving cell specific gene transcription, possibly by changing histone modification signatures. This is important during differentiation processes when progenitor cells are acquiring a new cell identity.

DNA methylation was investigated using a 450K methylation array which quantifies methylation levels using probes designed to known CpG sites. This approach is often described as genome wide but whilst probes cover a large degree of the genome, the coverage is far from comprehensive. Our analysis reveals that CpG probes are highly biased towards chromatin states surrounding gene promoters and the proportion of probes outside these regions are small. Whilst the design of CpG probes is logical as traditionally DNA methylation was studied at gene promoters due to CpG islands mainly being found at promoters (Lim and Maher, 2010), more recent research – including this study, has pointed towards a role for DNA methylation within gene bodies and enhancers (Bell et al, 2016; Yang et al, 2014). The limited CpG probes designed to these areas means that using this technology for investigation into enhancer methylation leads to an incomplete picture of DNA methylation within these regions. Additionally, methylation of novel CpG sites cannot be quantified using array based technology. The newer Illumina MethylationEPIC array provides greater coverage at over 850K CpG sites in the genome but whole genome bisulfite sequencing is required for a truly genome wide approach that also assays novel CpG sites.

Our data reveals a correlation between DNA methylation and strong enhancers but not a causation; it is unknown from our data which mechanism acts first or how DNA methylation affects changing chromatin states. Other research suggests that DNA methylation affects histone modifications and precedes histone modification changes (Rose and Klose, 2014; Hashimshony et al, 2003). For both ChIP-seq and DNA methylation experiments, populations of hMSCs was seeded into transwell inserts and cultured in chondrogenic media to induce chondrogenesis. Multiple transwells were used per sample for ChIP-seq. However, although culture conditions were the same, we cannot guarantee that all cells were in the same stage of differentiation when chromatin was harvested even within the same transwell insert. It's likely that the final population of differentiated chondrocytes at day 14 were heterogeneous and even slight differences may affect histone modification patterns. Furthermore, the chondrogenesis experiments for DNA 450K array was carried out at a different time with extra hMSC donors. Therefore, investigating temporal changes in chondrogenesis using these data is not feasible.

In summary, our analysis has found that gene enhancers, particularly strong enhancers denoted by histone marks H3K4me1 and H3K27ac, are more segregated by cell type and are hypomethylated in chondrogenesis. Enhancers are likely to be involved in the regulation of transcription of cell specific gene expression. Indeed, GO term analysis of enhancer chromatin states, rather than promoter or transcription states, revealed enrichment of terms relating to cell specific biological processes. The summarised evidence in this chapter so far points to a unique regulatory role of H3K27ac and H3K4me1 defined strong enhancers to promote cell type specific gene transcription. We have shown for the first time that some CpG sites within gene enhancers become hypomethylated during chondrogenesis. Gene transcription is regulated through epigenetic mechanisms which can be influenced by various factors, both genetic and environmental and there has been considerable interest in the manipulation of gene expression through epigenetic means (Hilton et al, 2015; Dominguez et al, 2016). Further research is needed to elucidate whether *cis*-regulatory elements such as enhancers can be exploited to control gene transcription in a therapeutic or tissue engineering context.

5.5 Conclusion

- Chromatin state learning using ChromHMM identified 16 distinct chromatin states.

- GO term analysis of chromatin states revealed that enhancers are linked to genes with GO terms related to cell specific processes.

- Comparison of chondrogenesis chromatin states to the Roadmap Epigenomics project's 18 state model indicates that gene enhancers are more specific to cell type and are more distinct between cell types compared to other chromatin states.

- DNA is hypomethylated during the transwell model of chondrogenesis.

- Strong enhancers are more hypomethylated compared to other chromatin states during chondrogenesis and osteoblastogenesis but not adipogenesis.

- DNA demethylation during chondrogenesis is accompanied by histone modification changes leading to altered chromatin state.

Chapter 6. Identification and characterisation of super enhancers in chondrogenesis

6.1 Introduction

Cis-regulatory elements such as promoters, transcription factor binding sites and enhancers regulate gene transcription. Promoters are located closely upstream to the TSS of genes and are on average ~500bp in size (Lui and States, 2002). Conversely, gene enhancers can be located large distances away from start sites of genes and may cover large genomic areas. Furthermore, gene enhancers do not necessarily target their nearest gene; often they affect transcription of genes not in their immediate vicinity, in fact, it is thought that the majority of enhancers do not affect their nearest gene (Whalen et al, 2016). Enhancers physically interact with target gene promoters through a chromatin looping mechanism. One study found that only 22% of enhancers physically interact with a neighbouring active TSS (Sanyal et al, 2012). Other features, such as DNA binding proteins, within enhancers may give clues as to which gene promoter an enhancer could act on. Investigating transcription factor binding in enhancers will help elucidate further the regulatory mechanisms involved in chondrogenesis.

In recent years, much attention has been given to a potentially novel subset of enhancers called super enhancers. Super enhancers are loosely defined as multiple gene enhancers in close proximity bound by high levels of active enhancer markers (Pott and Lieb, 2014). Enhancer markers used to define super enhancers can include transcription factors, transcriptional co-activators and histone modifications. H3K27ac is commonly used to define super enhancers, either on its own or in conjunction with other enhancer markers (Hnisz et al, 2013). Transcription factors known to be important in a particular cell type or process may be used to define super enhancers that drive expression of genes specific to a cell type or process. Like typical enhancers, super enhancers are associated with cell specific gene expression rather than housekeeping genes or genes necessary for general cell function. Super enhancers have been implicated in cancer (Sur and Taipale, 2016) and stem cell differentiation (Adam et al, 2015).

Transcription factor binding sites may be present within other *cis*-regulatory regions such as promoters and enhancers. SOX9 is commonly described as the master transcription factor involved in chondrogenesis. Previous evidence has shown that SOX9 binds to both gene enhancers as well as promoters in chondrocytes (Ohba et al, 2015). SOX9 binding during chondrogenesis also coincides with binding of the AP-1 transcription factor, which generally comprises heterodimers from the Jun and Fos family of proteins. The most common heterodimers are c-Jun and c-Fos, encoded by the *JUN* and *FOS* genes respectively (Karin et al, 1995). Co-binding of AP-1 and SOX9 promotes chondrocyte hypertrophy (He et al, 2016), the terminal stage of chondrogenesis (Pacifici et al, 1990). Due to the importance of these transcription factors during chondrogenesis, they are ideal candidates for identifying super enhancers in chondrocytes.

Previous chapters have pointed to an important role of gene enhancers during chondrogenesis. We observed that enhancers are more distinct between cell types compared to other chromatin states and that they are hypomethylated during chondrogenesis and osteoblastogenesis. In this chapter, we further explore the characteristics of gene enhancers in chondrogenesis. With the aid of external mouse Sox9 and Jun ChIP-seq datasets, super enhancers were identified in our differentiated chondrocytes and enhancer activity assessed using luciferase reporter assays. We identified four genes, *LOXL1-4*, with associated enhancers or super enhancers and have begun to characterize their roles in chondrogenesis.

6.2 Aims

- Analyse mouse Sox9 and Jun ChIP-seq datasets. Convert Sox9 and Jun ChIP-seq read co-ordinates from mouse to human.

- Identify chondrocyte super enhancers using our histone ChIP-seq dataset and SOX9 and JUN ChIP-seq peaks.

- Assess activity of identified super enhancers using luciferase reporter assays, with and without SOX9 overexpression

- Compare chondrocyte super enhancers to known super enhancers

- Deplete LOXL1-4 genes in hMSCs using siRNAs to investigate the effect on chondrogenesis

6.3 Results

6.3.1 Sox9 and Jun ChIP-seq analysis

Sox9 is widely considered the principal transcription factor driving chondrogenesis; without Sox9, chondrogenesis cannot occur (Akiyama et al, 2002). The activator protein 1 (AP-1) complex is also involved in promoting chondrogenesis through the activation of the TGFβ signalling pathway (Huang et al, 2005). The AP-1 transcription factor usually comprises heterodimers of Jun and Fos proteins. A study has shown that the Jun protein co-binds with Sox9 during chondrogenesis to promote hypertrophy (Ohba et al, 2015; He et al, 2016). To elucidate how transcription factor binding acts together with histone modifications to influence gene expression in chondrogenesis, we investigated the binding of SOX9 and JUN within chondrocyte chromatin states. Unfortunately, human SOX9 and JUN ChIP-seq data are not publically available. However, SOX9 and JUN ChIP-seq has been performed in mouse rib chondrocytes (Ohba et al, 2015; He et al, 2016). Sox9 and Jun ChIP-seq reads were converted from mouse (mm10) to human (hg38) using UCSC liftOver. Liftover between species can be problematic if the two species are separated by a large evolutionary distance (LoVerso and Cui, 2015). Fortunately, mice and humans are not dissimilar and the mouse is widely used as a model for human disease and development in research. Over 90% of the human and mouse genomes exhibit shared synteny (Chinwalla et al, 2002). To confirm that liftover of SOX9 and JUN peaks to human genome coordinates was appropriate in our analysis, motif discovery using MEME-ChIP (default settings) was performed.

De novo motif discovery using SOX9 peaks called from reads lifted over from mouse to human revealed that the top motif (Fig. 6.2A) matched to a motif belonging to EGR-1 (early growth response protein 1) which is required for stem cell differentiation and has been shown to have a role in chondrogenesis (Spaapen et al, 2013), tenogenesis (Guerquin et al, 2013) and adipogenesis (Boyle et al, 2009). The multiple roles of EGR-1 suggest that it is differentiation specific rather than specific to chondrogenesis alone. Another match for this motif included the Sp/KLF family of transcription factors such as Sp-1 and Sp-2 (Appendix IV, Table 1). This family of general transcription factors typically binds to gene promoters to regulate transcription (Voelkel et al, 2016).

Additional motif matches include other proteins with zinc finger motifs. Zinc finger motif containing proteins commonly bind to biological molecules such as other proteins, RNA, and DNA and have a diverse range of functions including regulation of gene transcription (Ravasi et al, 2003). The second top motif found by MEME belonged to the SOX family of transcription factors, including SOX9 (Fig. 6.1B). The third belonged to the AP-1 transcription factor comprised of a JUN and FOS heterodimer (Fig. 6.1C). The presence of other transcription factor motifs within SOX9 ChIP-seq data indicates a potential interaction between SOX9 and other transcription factors.

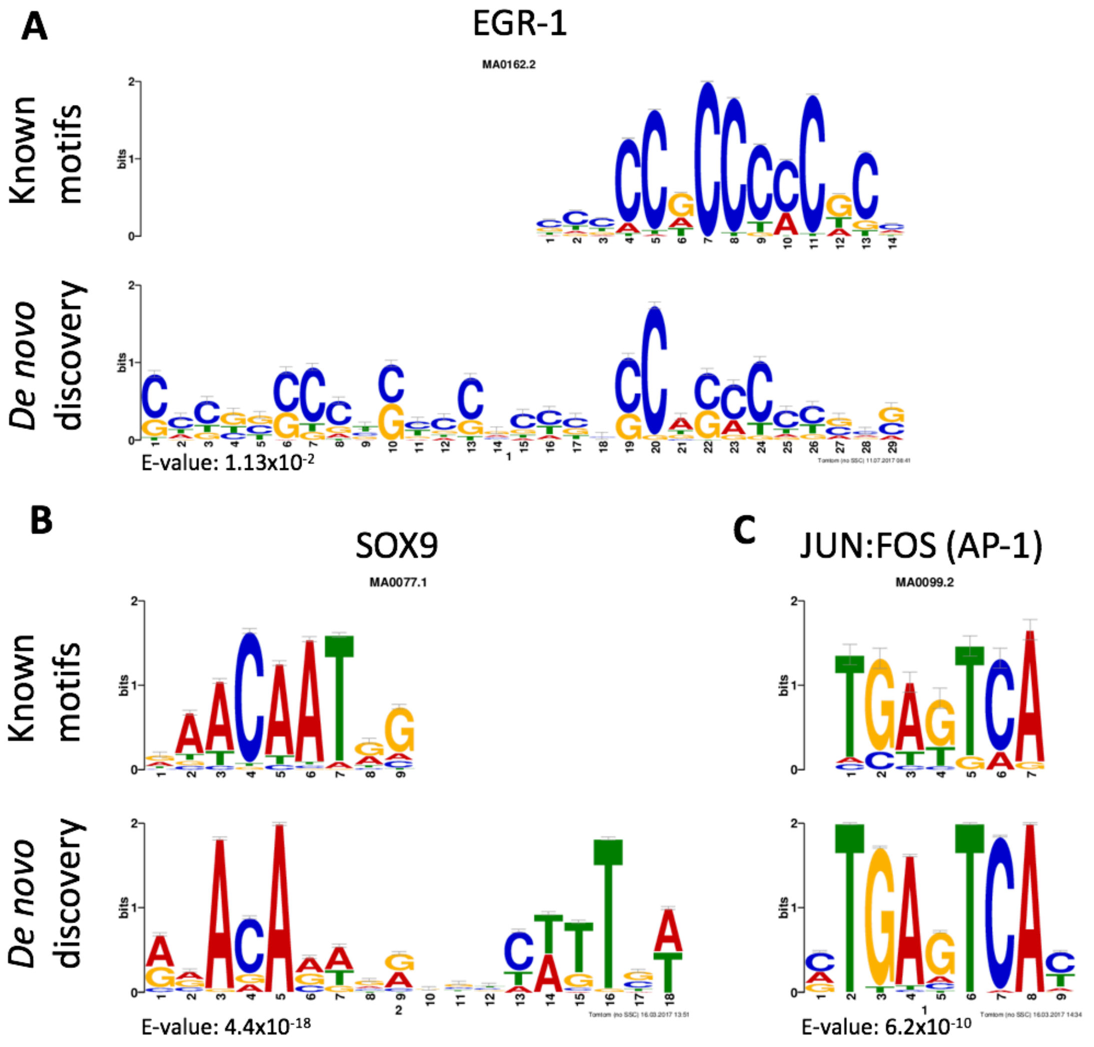


Figure 6.1 – Motifs found in SOX9 ChIP-seq peaks, derived from lifted over mouse Sox9 ChIP-seq reads, by MEME. Peaks called using MACS2 after lifting over aligned reads from mm10 to hg38. (A) Top motif found by de novo motif discovery using MEME ChIP-seq belonged to the EGR-1 transcription factor. (B) the second motif matched the SOX family of transcription factors and includes SOX9 and (C) the third motif matched the AP-1 transcription factor.

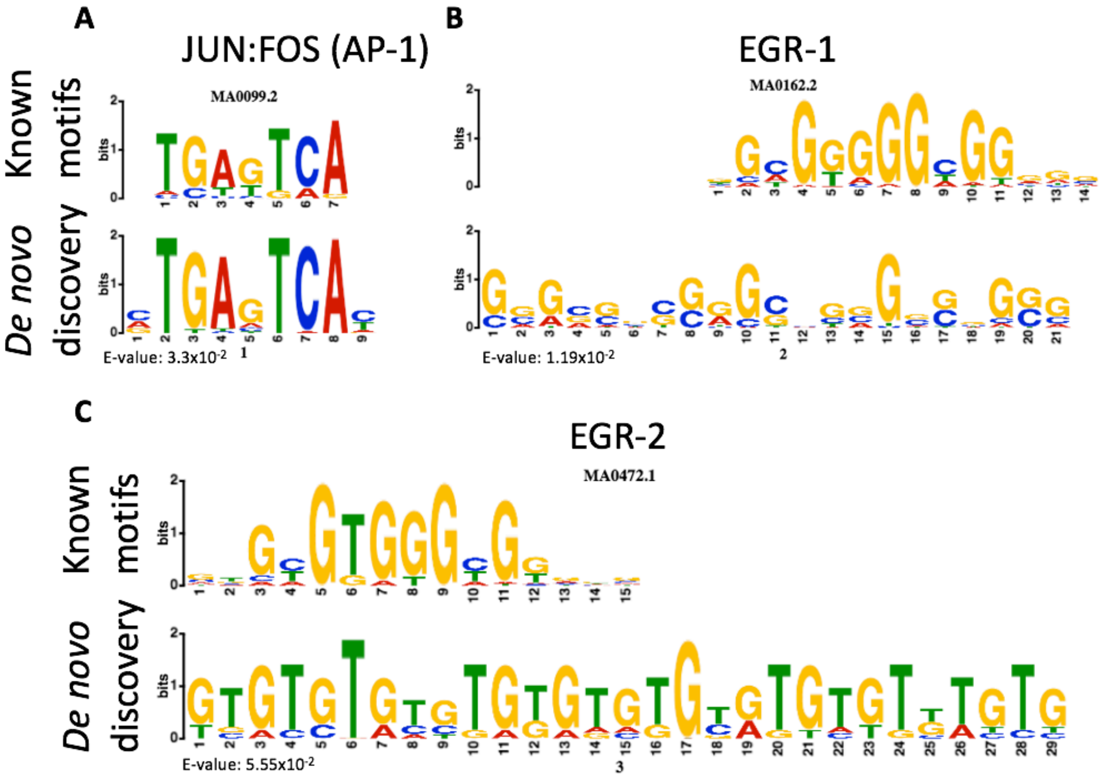


Figure 6.2 - Motifs found in JUN ChIP-seq peaks, derived from lifted over mouse Jun ChIP-seq reads, by MEME. Peaks called using MACS2 after lifting over aligned reads from mm10 to hg38. (A) Top motif found by de novo motif discovery using MEME ChIP-seq belonged to the AP-1 transcription factor. (B) the second motif matched the EGR-1 transcription factor (C) the third motif matched the EGR-2 transcription factor

Motif discovery was also performed on the JUN ChIP-seq peaks called after read liftover from mm10 to hg38. The top motif found matched to the AP-1 transcription factor, of which JUN forms along with FOS (Fig. 6.2A). The second motif found by MEME matched to EGR-1 (Fig. 6.2B) and Sp/KLF transcription factor family members and zinc finger proteins (Appendix IV, Table 2). The third motif matched EGR-2; similar to EGR-1, EGR-2 is involved in development. Although the two EGR transcription factors can be involved in the same processes, they often have opposing effects (Du et al, 2014; Boyle et al, 2009).

The presence of other transcription factor binding motifs found using SOX9 peaks indicates that SOX9 potentially co-binds with other transcription factors. Previous studies have shown that SOX9 forms complexes with other proteins before binding to

DNA (Jo et al, 2014). Motif discovery using SOX9 peaks yields AP-1 binding motifs. However, motif search with JUN peaks did not yield SOX9 binding motifs. These findings corroborate with motif discovery in the original mouse dataset (Ohba et al, 2015; He et al, 2016) and shows that liftover of ChIP-seq reads achieved expected outcomes. He et al found a high degree of overlap between Sox9 and Jun ChIP-seq peaks in chondrocytes. To investigate whether this was the case in our lifted over data, we assessed the overlap rate between hg38 SOX9 and JUN peaks. There were 51082 significant (*q* < 0.05) peaks in the lifted over JUN ChIP-seq data and 39465 significant peaks in the lifted over SOX9 ChIP-seq data. There were 26413 SOX9 peaks with one or more overlapping JUN peaks; a total of 28602 JUN peaks overlapped a SOX9 peak with some overlapping more than one (Fig. 6.3).

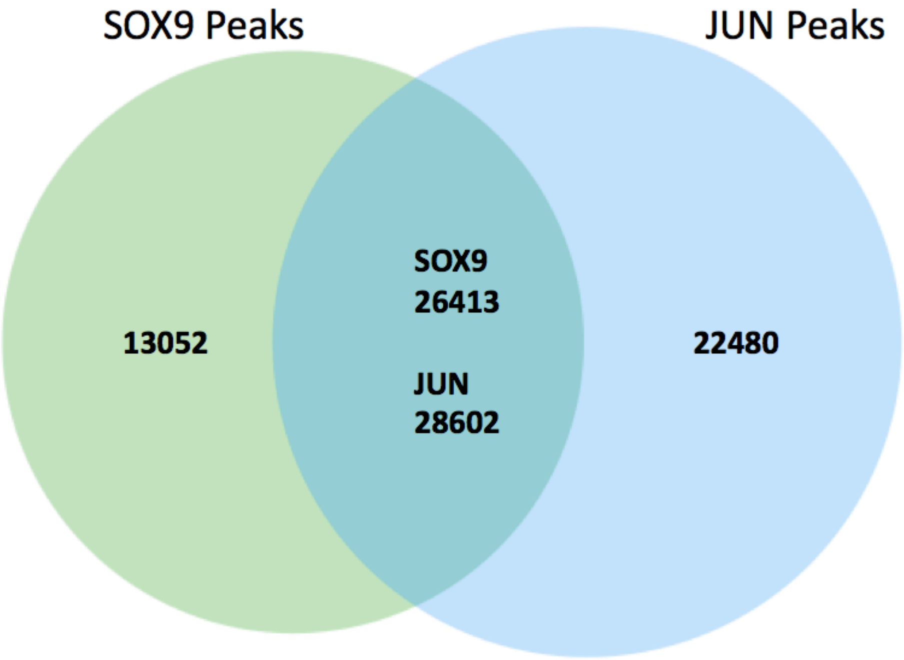


Figure 6.3 – Overlapping SOX9 and JUN peaks. Mouse rib chondrocyte SOX9 and JUN ChIP-seq data (along with respective input controls) were aligned to mouse reference genome mm10; aligned reads were converted into hg38 coordinates using the UCSC liftOver tool. Significant peaks (p < 0.05) were called on lifted over reads using MACS2 peak caller (default settings). Overlapping peaks were assessed using BEDTools intersect and Venn diagram created using the Venneuler package in RStudio.

Peak overlaps suggest that SOX9 and JUN do indeed co-bind, as suggested by *de novo* motif analysis on SOX9 peaks yielding AP-1 motifs, or at least share common binding sites. Transcription factor ChIP-seq was performed on a population of cells and therefore overlapping peaks may reflect heterogeneity of SOX9 and JUN binding in different cells rather than co-binding on the same genome. However, *in situ* hybridization experiments show that Sox9 and Jun co-localize to the same region of the growth plate in mouse tibia (He et al, 2016).

67% of SOX9 peaks had at least one overlapping JUN peak, this shows that the majority of SOX9 binding is accompanied by the AP-1 transcription factor in chondrocytes. SOX9 can bind to gene promoters where it regulates transcription of non-chondrogenic genes and can also bind to enhancers to regulate chondrogenesis and cartilage related genes, termed class I and class II SOX9 binding sites respectively (Ohba et al, 2015). SOX9 peaks with at least one overlapping JUN peak were intersected with chondrocyte chromatin states to investigate where in the chondrocyte genome co-binding of these transcription factors occur (Fig. 6.4).

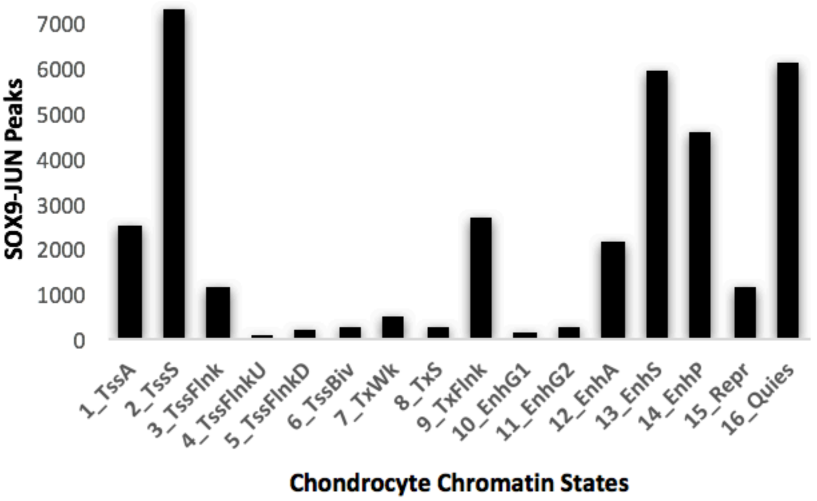


Figure 6.4 – Overlapping SOX9-JUN peaks in chondrocyte chromatin states. Overlaps were found using BEDTools intersect; some peaks span more than one chromatin state, where this was the case both states were included.

The chromatin state with the most overlapping SOX9-JUN was the strong active promoter/TSS state (2_TssS) indicating that highly expressed genes possess SOX9 and AP-1 binding at their promoters. The strong enhancer (13_EnhS) and quiescent (16_Quies) states both have similar numbers of SOX9-JUN overlapping peaks. SOX9 is known to bind to enhancers but it is unclear why a similar number of SOX9-JUN peaks are found in the quiescent state. The quiescent state is the largest chromatin state in the genome, if transcription factor binding was assumed to be random then most binding sites would be located in this state. However, transcription factor binding is not random. Poised enhancers (14_EnhP) also contain a large number of SOX9-JUN overlap peaks. Few SOX9-JUN peaks are found in chromatin states covering gene bodies demonstrating that regulation of genes through SOX9 and AP-1 is primary through binding at the promoter of genes or at distal *cis*-regulatory regions.

GO terms for genes associated with SOX9-JUN peaks in 2_TssS, 13_EnhS and 16_Quies were retrieved using GREAT Ontology after converting hg38 coordinates to hg19 (GREAT does not support hg38). As hypothesised, SOX9-JUN peaks at 2_TssS states regulate genes involved in general cell function whereas binding in the 13_EnhS state regulate cartilage development genes. GO terms for 14_EnhP include cartilage development related terms as well as terms associated with other developmental processes. Terms related to chondrogenesis may reflect a temporal difference between SOX9 binding and activation of poised enhancers and GO terms connected to other differentiation processes could represent poised enhancers which may never be activated in chondrocytes. Interestingly, GO terms for SOX9-JUN peaks present in the 16_Quies state were related to development of other cell types although *regulation of cartilage development* was also among these GO terms (Fig. 6.5). GO terms for SOX9-JUN in chondrocyte 16_Quies state is suggestive that co-binding of SOX9 and AP-1 is also important in other developmental processes. As well as chondrogenesis, SOX9 is known to regulate differentiation of other cell types. SOX9 is also involved in gliogenesis (Stolt et a, 2003; Vong et al, 2009), maintenance of hair follicle stem cells (Kadaja et al, 2014), pancreas development (Seymour, 2014) and more. Apart from differentiation processes, SOX9 is also involved in sex determination in mammals (Koopman, 2001). With such a large repertoire of roles, SOX9 requires the interaction of other proteins and epigenetic mechanisms in order to promote the correct cell process.

SOX9 has multiple roles and binds to different *cis*-regulatory regions in the genome but binding at gene enhancers in chondrocytes yielded the most GO terms related to chondrogenesis compared to other chromatin states. This illustrates the complex relationship between DNA binding proteins and histone modifications. Further investigation is required to fully understand gene regulation by SOX9 and its binding partners. It will be interesting to include SOX9 ChIP-seq data from different cells types in future analyses. Unfortunately, the investigation of SOX9 binding in different cell types is beyond the scope of this project.

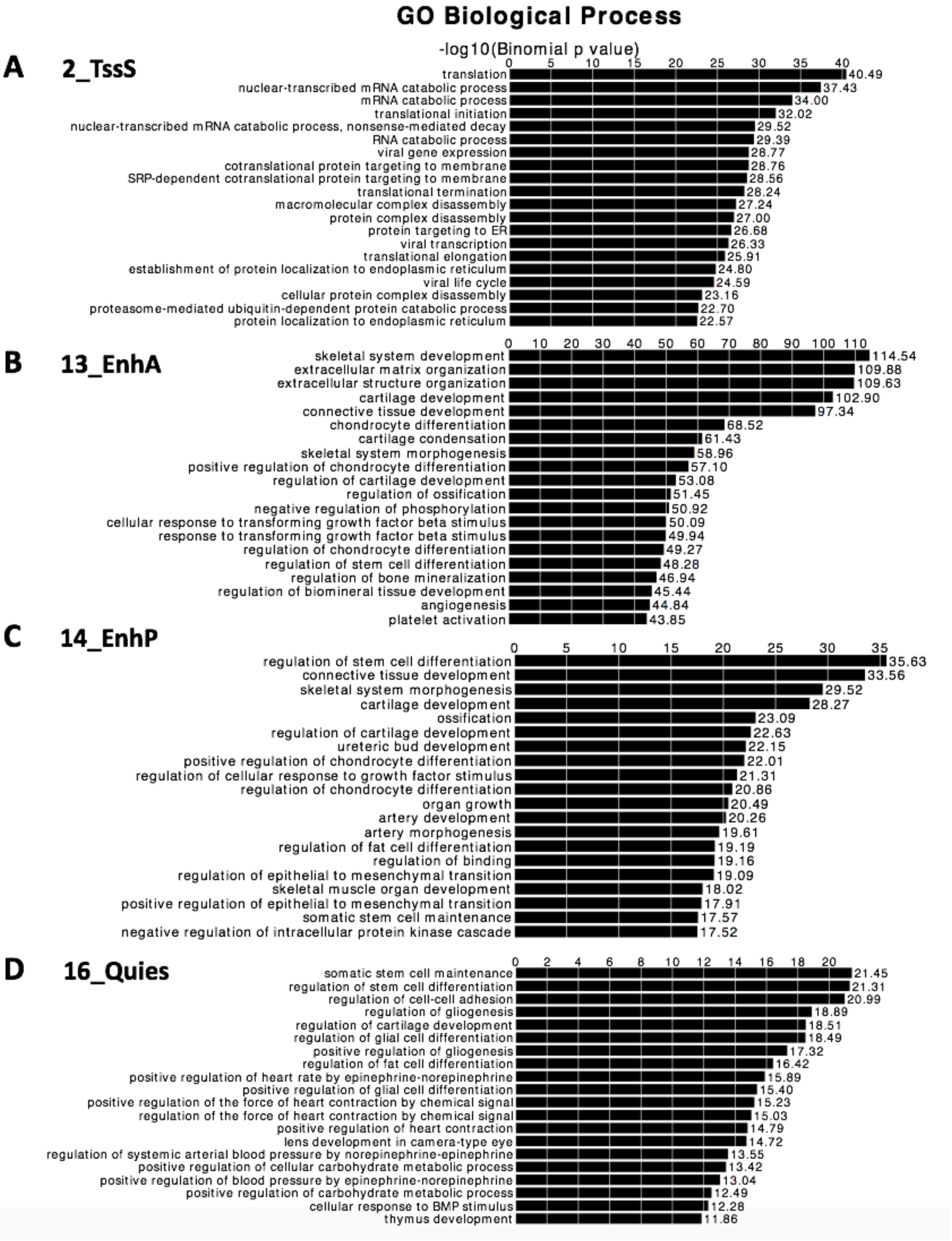


Figure 6.5 – Top 20 significant (FDR < 0.05) GO terms found using GREAT GO Ontology tool for SOX9-JUN overlapping peaks within chondrocyte chromatin states (A) 2_TssS, (B) 13_EnhS and (C) 14_EnhP and (D) 16_Quies

6.3.2 Identification of super enhancers

Following the validation that mouse SOX9 and JUN ChIP-seq data produced similar results when lifted over to humans, we sought to identify super enhancers in chondrocytes using these transcription factors as enhancer markers. The general method for identifying super enhancers involves a three step process (Pott and Lieb, 2015). First, typical enhancers must be identified. Active enhancers are marked by H3K27ac and this histone mark is usually used to classify enhancer locations. Other enhancer markers such as the Mediator 1 protein complex which is also associated with active enhancers (Kuras et al, 2003) can also be used for this first step. The second step clusters individual enhancers within 12.5kb of each other. The third step involves another enhancer marker which is intersected with clustered enhancers and ranked on ChIP-seq peak signal. Clustered enhancers also exhibiting high levels of the enhancer marker used for the third step are classed as super enhancers. Determining the enhancer marker enrichment threshold for super enhancers is somewhat arbitrary; enhancers are ranked and plotted on the basis of the second enhancer marker and the signal value cut off is the point of the curve where the slope equals 1 (Fig. 6.6). Super enhancers may be defined using different markers for step one and step three or using the same marker for both steps (Hnisz et al, 2013).

To identify chondrocyte super enhancers, strong enhancer states (13_EnhS) in our chondrocyte chromatin states were used; this constitutes step one in the method outlined above. As an extra control, we also stipulated that strong enhancer states must also be an enhancer state in Roadmap BM-MSC derived chondrocytes (cell type E049). The level of SOX9 peak enrichment was used to define super enhancers from typical enhancers. SOX9 peaks must also have an overlapping JUN peak, although the JUN enrichment signal was not directly taken into account. SOX9 peaks meeting these criteria were plotted by ranking SOX9 signal enrichment (MACS2 signal value score) and an enrichment value threshold above 19 was used to define super enhancers (Fig. 6.6). Using this method, 746 super enhancers were identified in chondrocytes.

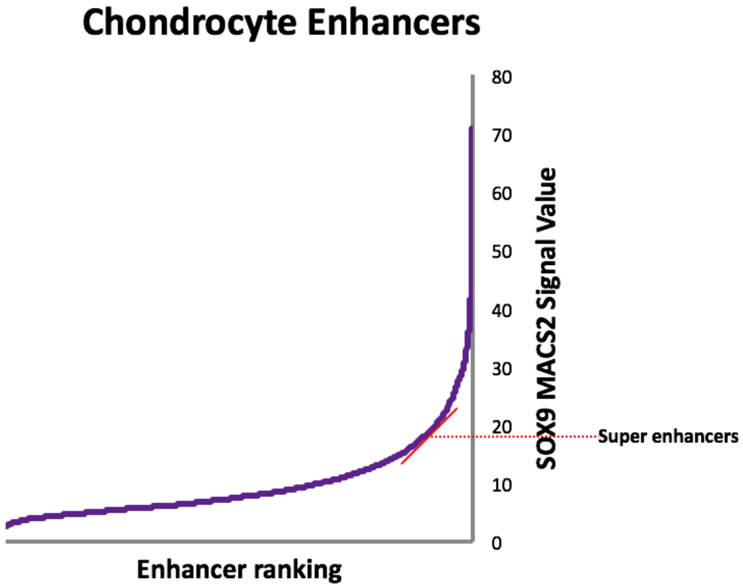


Figure 6.6 – Identification of super enhancers in chondrocytes. Chromatin states characterised as strong enhancers (13_EnhS) were used to identify all enhancers. SOX9-JUN overlapping peaks were intersected with 13_EnhS states. Strong enhancer regions were ranked on the basis on SOX9 peak enrichment. The SOX9 enrichment cut off for super enhancers was defined as above 19, this is the point of the curve where the slope = 1. There were 746 identified in chondrocytes.

DNA methylation levels in super enhancers did not differ significantly from typical enhancers (*p* > 0.05; Fig. 6.7); only the level of SOX9 enrichment separates super enhancers from typical enhancers in our dataset. GO terms for super enhancers were retrieved using the GREAT GO tool (Fig. 6.8). As expected, almost all terms were related to cartilage development. These results were similar to GO terms for the strong enhancer state (Chapter 5) but are more chondrogenesis specific. Super enhancers yielded 8 GO terms containing the words “cartilage” or “chondrocyte” whereas the strong enhancer state contained 2 terms with these words.

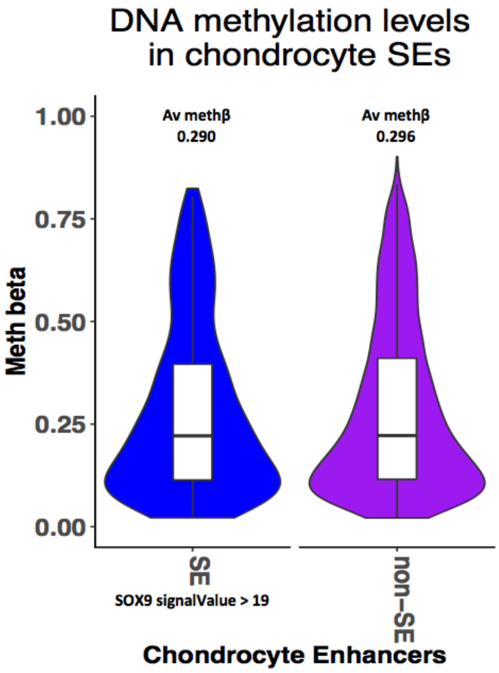


Figure 6.7 – DNA methylation levels (β values) in SOX9 defined super enhancers and SOX9 bound typical enhancers

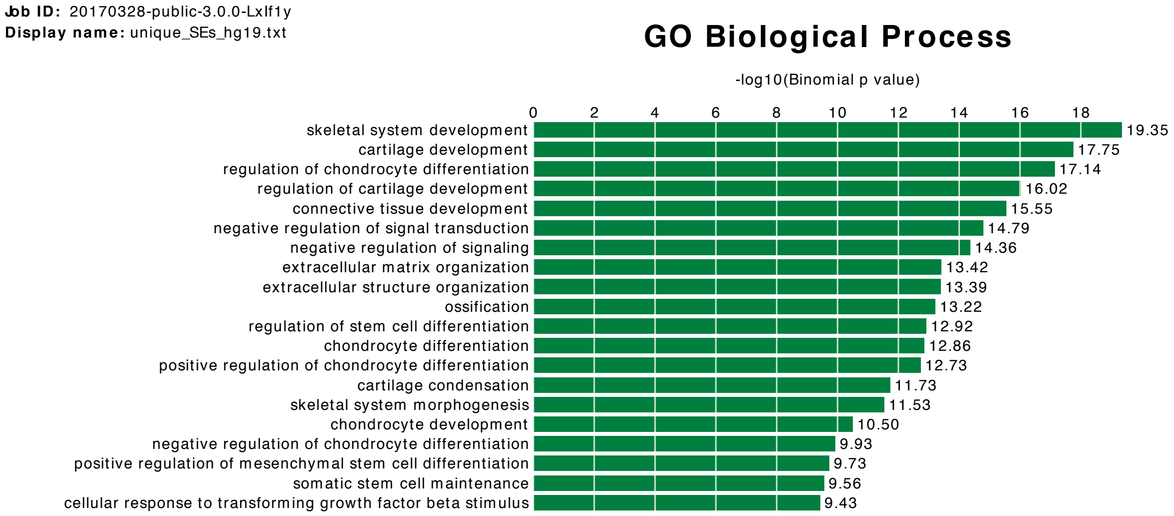


Figure 6.8 – Biological process GO terms for chondrocyte super enhancers. Super enhancer co-ordinates were converted from hg38 to hg19; GO terms were retrieved using GREAT (default settings).

6.3.3 Assessment of enhancer activity using a luciferase reporter assay

Following the generation of a list of super enhancers, some were selected for further validation based on their proximity to cartilage related genes (Table 6.1). SOX9 peak regions within super enhancers were cloned into a PGL3-promoter vector and transfected into the SW1353 and HEK293T cell lines with and without SOX9 overexpression. The PGL3-promoter vector contains a luciferase gene; potential enhancers may be cloned upstream of the luciferase gene and SV40 promoter. After transfection into cells, luciferase activity may be measured as a proxy for enhancer activity.

Table 6.1 – Super enhancer segments selected to be cloned into PGL3-promoter vectors. Regions selected are SOX9 peaks within a super enhancer. Super enhancer segments were given a number for simpler referencing; no. 2 is absent due to the unsuccessful cloning of the region.

Insert no.	**Locus (hg19)**	**Length**	**Nearest gene**	**Predicted gene target**
1	chr2:74810421-74811389	968	DOK1	LOXL3
3	chr16:69924387-69924932	545	WWP2	WWP2
4	chr16:17452680-17453167	487	XYLT1	XYLT1
5	chr12:104879495-104880612	1117	CHST11	CHST11
6	chr1:183922069-183922664	595	COLGALT2	COLGALT2

Initially, enhancer luciferase assays were performed using the SW1353 cell line. This chondrosarcoma cell line is commonly used in cartilage research and is often considered a chondrocyte cell line (Tew et al, 2007; Gebauer et al, 2005; Santoro et al, 2015). Of the five enhancer regions successfully cloned into a PGL3-promoter vector, three showed significantly increased luciferase activity compared to an empty vector control (Fig. 6.9A). One of the aims of this experiment was to observe whether overexpression of SOX9 affected enhancer activity. Although three enhancer constructs increased luciferase activity (insert no. 1, 3, 6), overexpression of SOX9 did not increase this further. We hypothesized that endogenous SOX9 expression in SW1353 may be sufficient to saturate enhancers and therefore overexpression of SOX9 may not influence luciferase activity. Consequently, the experiment was repeated in a human embryonic kidney cell line, HEK293T, which should exhibit little SOX9 expression (Blache et al, 2004). The three enhancer constructs that increased luciferase activity in SW1353 cells also increased in HEK293T cells, plus an extra enhancer construct (insert no. 4). For inserts no. 1, 3 and 6 in HEK293T cells, overexpression of SOX9 significantly increased luciferase activity further compared to an empty vector control (Fig. 6.9B). Interestingly, overexpression of SOX9 led to a decrease of luciferase activity in enhancer construct no. 5. A protein immunoblot was performed to confirm SOX9 overexpression (Fig. 6.10). There appears to be more endogenous SOX9 present in SW1353 cells compared to HEK293T cells without overexpression of SOX9. However, endogenous SOX9 was also observed in HEK293T cells showing that HEK293T cells do indeed express SOX9.

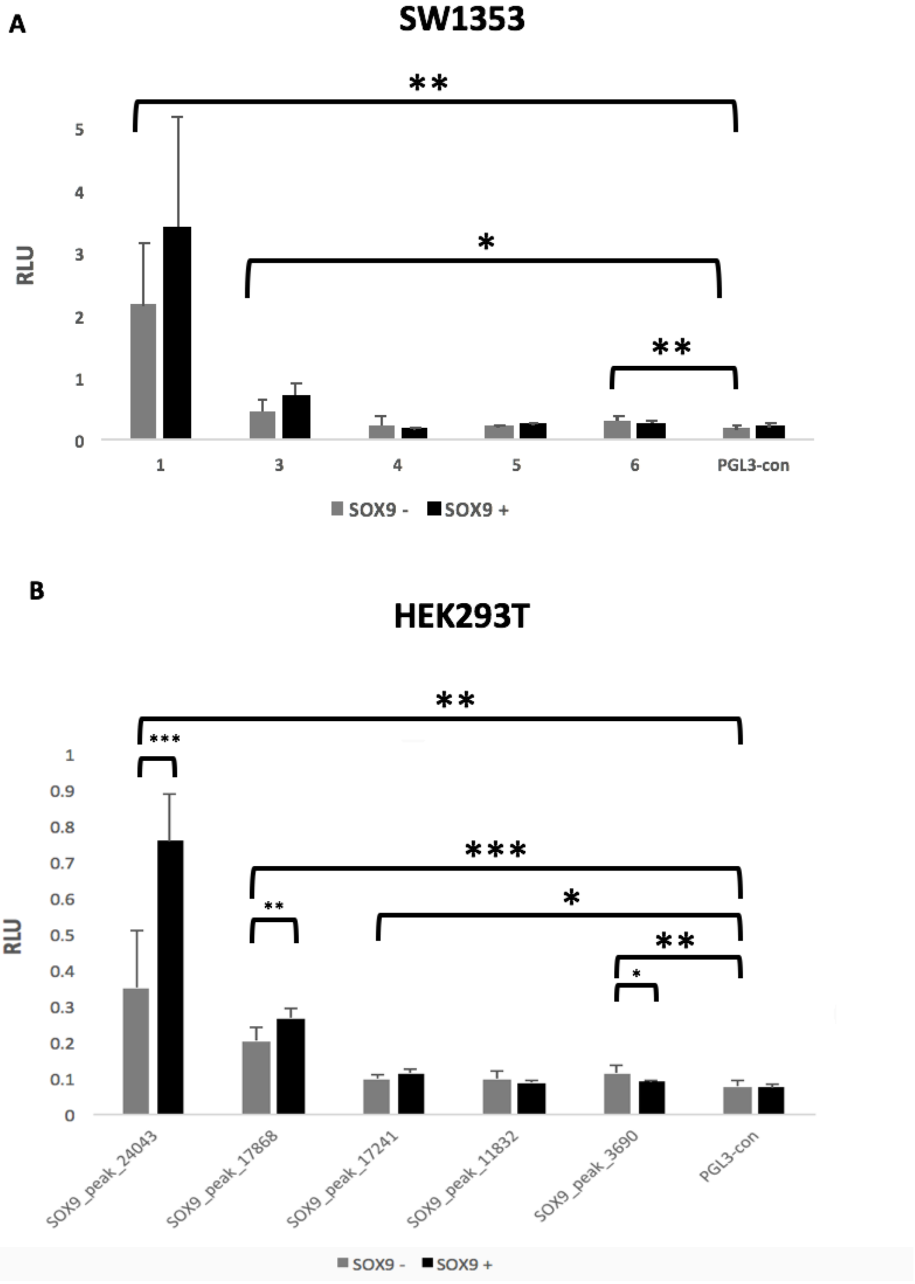


*Figure 6.9 – Enhancer luciferase assays performed in (A) SW1353 and (B) HEK293T cell lines. Small regions of super enhancers containing a SOX9 peak were cloned into PGL3-promoter vectors, transfected into SW1353 and HEK293T cell lines for 24hrs and the relative light unit (RLU) measured. Light units were normalised against a Renilla control. A t-test (two tailed, unpaired, equal variances not assumed) was performed to test for statistical significance between enhancer constructs and empty vector control, and SOX9 overexpression vs empty SOX9 control. Error bars show the standard deviation. Significant contrasts are marked by lines, significant p values are indicated by asterisks: * < 0.05, ** < 0.01, *** < 0.001.*

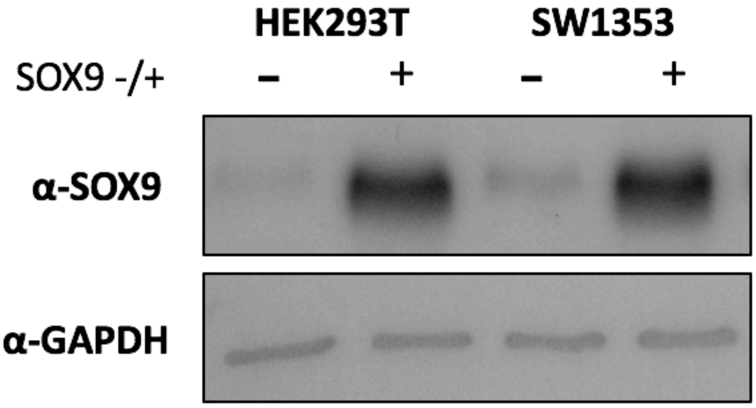


Figure 6.10 – Immunoblot of SOX9 overexpression in HEK293T and SW1353 cell lines. A SOX9 expression plasmid or empty vector control was transfected into cells for 24hrs. Protein was extracted and an immunoblot performed to detect levels of SOX9 protein. The SOX9 protein has a molecular weight of 56kDa. An immunoblot against GAPDH was used as a protein loading control.

In summary, enhancer luciferase reporter assays showed that at least three out of five segments of separate super enhancers assayed exhibited independent enhancer activity in SW1353 and HEK293T cell lines. Overexpression of SOX9 increased luciferase activity further in two reporter assays in HEK293T cells but not in SW1353 cells, possibly due to endogenous SOX9 being present in SW1353 cells. Unfortunately, luciferase reporter assays are unable to define which gene is being directly targeted by the cloned enhancer. Table 6.1 includes the nearest gene to the super enhancer and its predicted target gene. In all cases except one, the predicted target gene is also the nearest gene. As enhancers do not always target the nearest gene, super enhancers next to a gene upregulated in chondrogenesis were chosen as these were deemed more likely to be targeted by the super enhancers. The nearest gene to the super enhancer containing insert no. 1 is DOK1 (Fig. 6.11). However, DOK1 does not have an established role in chondrogenesis. The gene downstream of DOK1 is LOXL3 which is involved in collagen crosslinking (Lee and Kim, 2006). Consequently, we predicted that this super enhancer targets LOXL3 rather than DOK1. Further characterisation and functional analysis is required to validate this hypothesis.

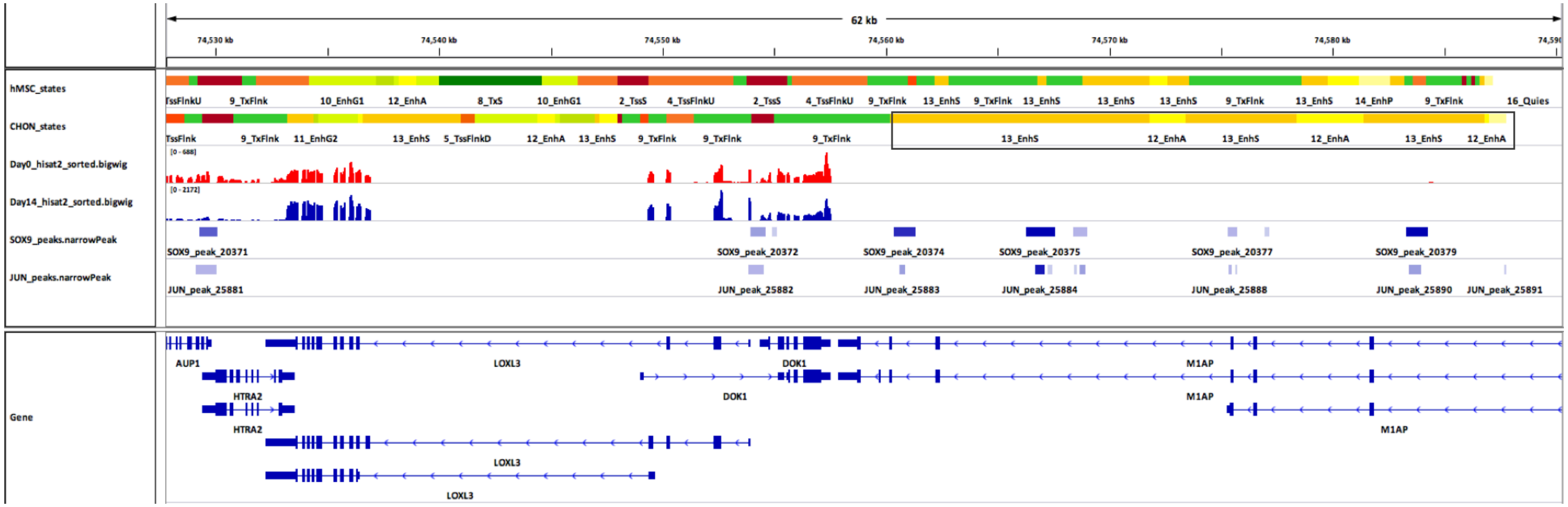

62 kb
74,530 kb
74,540 kb
74,550 kb
74,560 kb
74,570 kb
74,580 kb
74,59
hMSC_states
CHON_states
Day0_hisat2_sorted.bigwig
Day14_hisat2_sorted.bigwig
SOX9_peaks.narrowPeak
JUN_peaks.narrowPeak
Gene
TssFlnkU 9_TxFlnk 10_EnhG1 12_EnhA 8_TxS 10_EnhG1 2_TssS 4_TssFlnkU 2_TssS 4_TssFlnkU 9_TxFlnk 13_EnhS 9_TxFlnk 13_EnhS 13_EnhS 13_EnhS 9_TxFlnk 13_EnhS 14_EnhP 9_TxFlnk 16_Quies
TssFlnk 9_TxFlnk 11_EnhG2 13_EnhS 5_TssFlnkD 12_EnhA 13_EnhS 9_TxFlnk 9_TxFlnk 9_TxFlnk 13_EnhS 12_EnhA 13_EnhS 12_EnhA 13_EnhS 12_EnhA
[0 - 688]
[0 - 2172]
SOX9_peak_20371 SOX9_peak_20372 SOX9_peak_20374 SOX9_peak_20375 SOX9_peak_20377 SOX9_peak_20379
JUN_peak_25881 JUN_peak_25882 JUN_peak_25883 JUN_peak_25884 JUN_peak_25888 JUN_peak_25890 JUN_peak_25891
AUP1
LOXL3
DOK1
M1AP
HTRA2
DOK1
M1AP
HTRA2
M1AP
LOXL3
LOXL3

Figure 6.11 – IGV screenshot of potential LOXL3 super enhancer. RNA-seq read coverages, SOX9 and JUN peaks shown below chromatin states.

6.3.4 Super enhancer cell type comparisons and gene target predictions

There have been many studies investigating super enhancers since their discovery in 2012 and databases have been created to store information about super enhancers. One such database is dsSUPER which contains super enhancers found in the mouse and human genomes (Khan and Zhang, 2016). Super enhancers in 86 Roadmap Epigenomics human cells and tissues were identified using H3K27ac (Hnisz et al, 2013) and deposited into dbSUPER. The overlap tool within dbSUPER was used to assess the overlap between chondrocyte super enhancers and those in other cell types available in the database. Chondrocyte super enhancers coordinates were converted from hg38 to hg19 because dbSUPER only supports hg19. Super enhancers with at least a 10% overlap were considered present in both cell types (default setting).

The cell type that exhibited the most overlap with our chondrocyte super enhancers was the U87 cell line with 287 overlaps (Table 6.2). The U87 cell line was originally thought to derive from a human glioblastoma but their origin has since been called into question (Dolgin, 2016; Allen et al, 2016). Therefore, U87 may be an aberrant and potentially contaminated cell line and its super enhancer similarities to chondrocytes a coincidence. However, recent studies have shown that U87 cells may have MSC-like qualities (Oh et al, 2017; Svensson et al, 2017) which could explain why they possess so many super enhancers in common with those identified in chondrogenesis. As super enhancers are widely considered to be associated with cell type specific processes, it is unknown exactly why the U87 cell line would have so many in common with chondrocytes.

The second cell type with the most matches to chondrocyte super enhancers was primary osteoblasts. Osteoblasts derive from MSCs and share many similarities to chondrocytes (Gomez-Picos and Eames, 2015). During development, bone tissue can originate from cartilage via endochondral ossification, a process that occurs after the terminal hypertrophic stage of chondrogenesis (Yang et al, 2014; Zhou et al, 2014). One of the criteria to define our super enhancers was the presence of a SOX9 and JUN overlapping peak; co-binding of these proteins at enhancers is proposed to promote chondrocyte hypertrophy (He et al, 2016). Therefore, it is unsurprising that osteoblast super enhancers overlap many chondrocyte super enhancers. Although adipocytes also derive from MSCs, chondrocytes are more similar to osteoblasts and this is reflected in the number of super enhancers the cells have in common. These results corroborate with the findings of Chapter 5 which showed that enhancers from similar cell types tended to cluster together. This is further evidence that gene enhancers can be used to infer relationships between different cell types.

Table 6.2 – dbSUPER overlap of chondrocyte super enhancers

Cell type	Overlap with chondrocyte super enhancers
U87	286
Osteoblasts	212
Adipose nuclei	127

An important caveat to consider when using public databases and tools is the variety of different methods used to define enhancers and super enhancers. Furthermore, databases can suffer from lack of recent annotation and caution must be taken to avoid concluding that an absence of data is a negative result.

Previously we observed that the LOXL3 gene may be associated with a super enhancer. The potential super enhancer is upstream of LOXL3 within the M1AP gene and adjacent to the DOK1 gene. We predicted that this super enhancer is acting on LOXL3 rather than DOK1 due to LOXL3 being involved in collagen crosslinking. Enhancers target gene promoters through chromatin looping. Physical looping of chromatin occurs within topologically associating domains (TADs), whose boundaries are marked by insulator or boundary elements (Whalen et al, 2016). Chromatin conformation studies such as Hi-C are used to elucidate the genome wide chromatin interactions and characterize TADs. Although Hi-C data is not available for chondrocytes, it has been performed in a variety of other cell types. The 3D genome browser is a repository of high throughput chromatin conformation studies and can be used to view chromatin interaction (Wang et al, 2017). This was used to investigate whether the potential LOXL3 super enhancer could interact with the LOXL3 gene promoter. The H1-ESC is stem cell line and can be induced to differentiate into chondrocytes (Brown et al, 2014). The potential LOXL3 super enhancer is within the same TAD as the LOXL3 promoter in H1-MSCs showing that is possible for LOXL3 to be regulated by this super enhancer (Fig. 6.12A). We observed that this interaction could also occur in an unrelated cell line, GM12878 (a lymphoblastoid cell line; Fig. 6.12B). TAD boundaries can change between cell types and chondrocyte chromatin conformation data is required to confirm whether this interaction occurs in chondrocytes. However, other genes are also within the same TAD and further experiments are required to validate which gene is regulated by this super enhancer in chondrocytes.

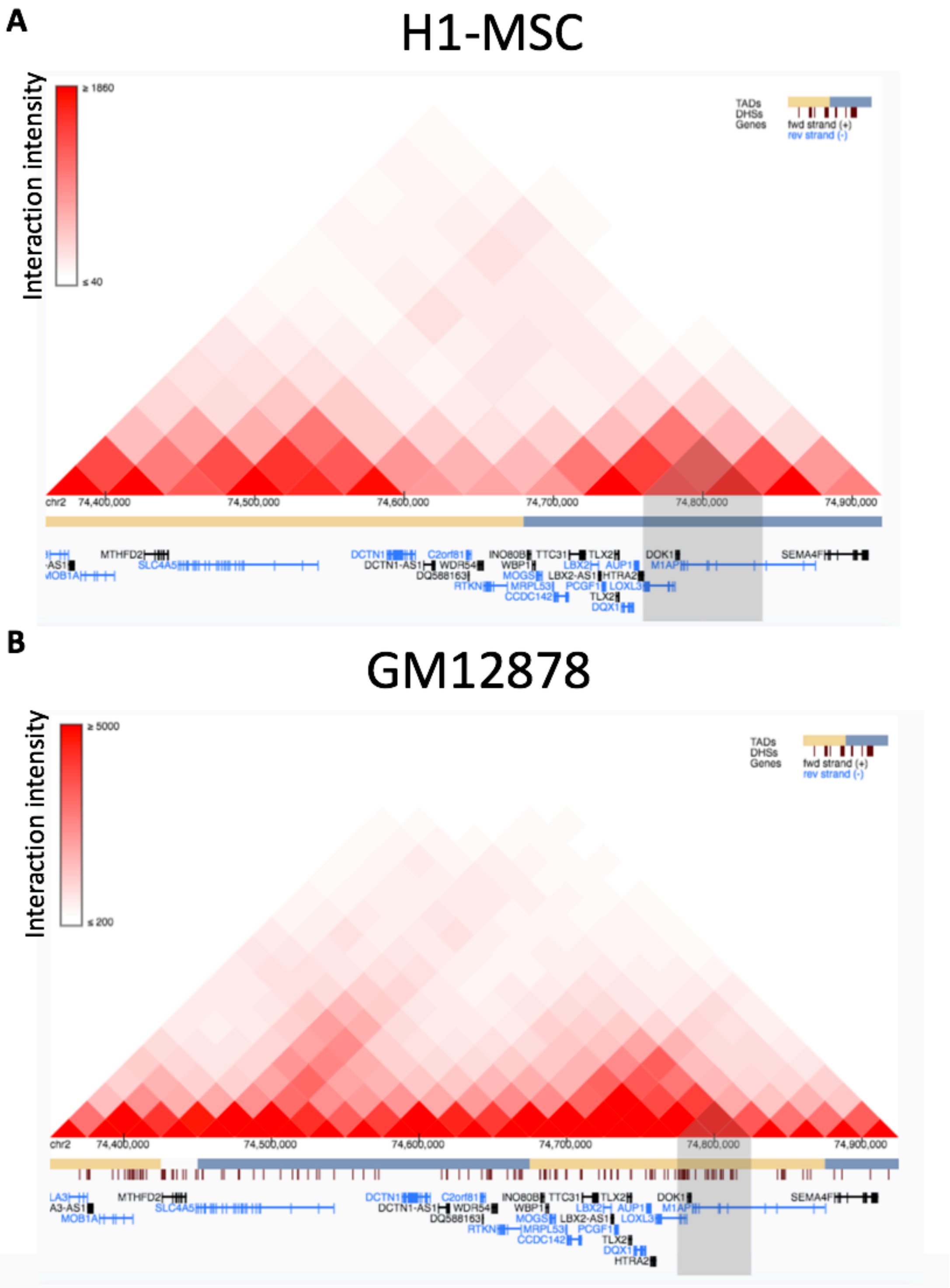


Figure 6.12 – Hi-C interaction matrix of LOXL3 in (A) H1-MSC (human MSC) and (B) GM12878 (human lymphoblastoid cell line). Enhancer-LOXL3 promoter interaction is shaded. Red pyramids indicate the intensity of interaction, maroon lines are DNAseI hypersensitivity sites (DHS data not available for H1-MSC) and yellow/blue regions are separate TADs. Image generated using Hi-C 3D genome browser (Wang et al, 2017).

Enhancer-promoter prediction tools aim to identify likely gene targets of enhancers, usually based on existing data such as chromatin conformation and open chromatin studies. One such tool is PETModule (Zhao et al, 2016) which considers factor such as distance, conservation, GO terms and correlation of DNAseI hypersensitivity site signals in enhancers and promoters from existing datasets to generate a probability of enhancer-promoter pairing. This tool was chosen because the authors claim it to be superior to existing tools and because of its simple input requirement of enhancer coordinates, whereas other tools such as IM-PET (He et al, 2014) require more data input and preprocessing of data. PETModule was run using all 746 identified chondrocyte super enhancers (in hg19; hg38 not supported) using default parameters.

As a case study, we again considered the potential LOXL3 super enhancer. Although the super enhancer is closest to the promoter of the DOK1 gene, we predicted that this super enhancer is actually regulating LOXL3 due to its important role in collagen crosslinking. PETModule output suggests that this super enhancer does indeed regulate LOXL3 instead of DOK1. However, it is equally as likely to regulate other genes in the vicinity (Table 6.3). Enhancer-promoter target prediction tools are a helpful aid but can struggle with predictions when multiple genes are in close proximity in the genome; prediction scores can be very similar. These tools represent the first generation of enhancer-promoter prediction software and there is much scope for improvement. Manuals for these tools can be incomplete or vague and they are not user friendly even for researchers familiar with using command line tools.

Table 6.3 – PETModule output for the predicted LOXL3 super enhancer.

GENE	PROB	SE_CHR	SE_START	SE_END
AUP1	1	chr2	74787327	74813927
DQX1	1	chr2	74787327	74813927
LBX2-AS1	1	chr2	74787327	74813927
LOXL3	1	chr2	74787327	74813927
PCGF1	1	chr2	74787327	74813927
SLC4A5	1	chr2	74787327	74813927
TLX2	0.99	chr2	74787327	74813927
WDR54	0.98	chr2	74787327	74813927
HTRA2	0.97	chr2	74787327	74813927
LBX2	0.97	chr2	74787327	74813927
MOGS	0.97	chr2	74787327	74813927
WBP1	0.97	chr2	74787327	74813927
DOK1	0.96	chr2	74787327	74813927

6.3.5 siRNA depletion of LOXL1-4 genes

The *LOXL* (lysl oxidase) family of genes have an established role in collagen and elastin crosslinking. This family comprises *LOX* and the *LOX*-like members 1-4; although all are involved in crosslinking different collagens, they appear to have disparate roles depending on cell type. The different family members are implicated in a range of pathologies involving aberrant ECM formation. Disruption of the *Lox* gene in mice caused death shortly before or after birth due to defective cardiovascular development, likely due to impaired elastic fibre formation (Maki et al, 2002). Single nucleotide polymorphisms in *LOXL1* are associated with glaucoma and pseudoexfoliation syndrome, a disease characterised by abnormal crosslinking in the ECM of the eye (Schlötzer-Schrehardt and Naumann, 2006; Thorleifsson et al, 2007). *LOXL2* is required for chondrogenesis and depletion of *LOXL2* inhibits *SOX9* expression, suggesting it also acts as a transcriptional regulator separate from its role as a collagen cross linker (Iftikhar et al, 2011). Knockout of Loxl3 in mice leads to cleft palate formation and spinal deformity due to abnormal cartilage development (Zhang et al, 2015). Mutations in *LOXL3* was observed in a family with the cartilage disease Stickler syndrome (Alzahrani et al, 2015).

Apart from LOXL2, there have not been any studies on the function of LOXL proteins in chondrogenesis despite their role in collagen crosslinking and chondropathies. From our chondrogenesis RNA-seq data, we observed gene expression changes of *LOXL* genes from day 0 to day 6. *LOXL1* is downregulated during chondrogenesis whereas *LOXL2*, *LOXL3* and *LOXL4* were upregulated (Table 6.4). Following the observation that *LOXL3* may potentially be regulated by a super enhancer, we investigated whether siRNA depletion of *LOXL3* mRNA would influence chondrogenesis. Family members *LOXL1*, *LOXL2* and *LOXL4* were also investigated. hMSCs (donor 2454e) were transfected with siRNAs for 48hrs before inducing chondrogenesis in transwell inserts. RNA was extracted prior to differentiation (day 0) and at day 7 of chondrogenesis. Gene expression was measured using RT-qPCR. All siRNAs successfully depleted mRNA levels of their target genes at day 0 compared to controls (Fig. 6.13A).

Table 6.4 – Gene expression changes of LOXL1-4 at day 0 and day 6 of chondrogenesis. Gene expression reported in TPM, quantified by Salmon (quasi-mapping mode).

	DAY 0	DAY 6	LOG2FC
LOXL1	67.86678	0.6980193	-6.603294828
LOXL2	29.899996	61.474932	1.039852943
LOXL3	28.364686	83.169144	1.55195249
LOXL4	6.10088	13.4118	1.136413615

At day 7 of chondrogenesis, *LOXL1-4* genes remained depleted (Fig. 6.13B). *LOXL1* expression decreased from day 0 to day 7. Gene expression of *LOXL1-4* measured by RT-qPCR largely corroborated our RNA-seq results although RNA-seq was performed at RNA extracted at day 6 and RT-qPCR at day 7. Gene expression of *SOX9* and *ACAN* were also assayed at day 7 (Fig. 6.14), unfortunately *SOX9* and *ACAN* expression were not assayed at day 0. Expression of SOX9 decreased when *LOXL2* was depleted, as previously reported (Iftikhar et al, 2011). We also observed a decrease of *SOX9* when *LOXL1* and *LOXL3* were depleted although the highest decrease was seen with siLOXL2. Although *LOXL1* is normally downregulated in chondrogenesis, further decreases using siRNA treatment resulted in *SOX9* downregulation, suggesting that *LOXL1* regulates chondrogenesis in a concentration dependent manner. Interestingly, the same pattern of gene expression decreases with depletion of *LOXL1-4* was also seen in *ACAN*. This is possibly due to downregulation of *SOX9*; *ACAN* is a direct target of *SOX9* in chondrogenesis (Zhang et al, 2015).

To determine whether siRNA depletion of *LOXL1-4* has an effect on glycosaminoglycan production, a GAG assay was performed at day 7 of chondrogenesis. This assay measures the levels of proteoglycans such as aggrecan and glycoaminoglycans such as hyaluronic acid (Frazier et al, 2008). hMSCs were differentiated into chondrocytes using a pellet model of chondrogenesis (Zhang et al, 2010).

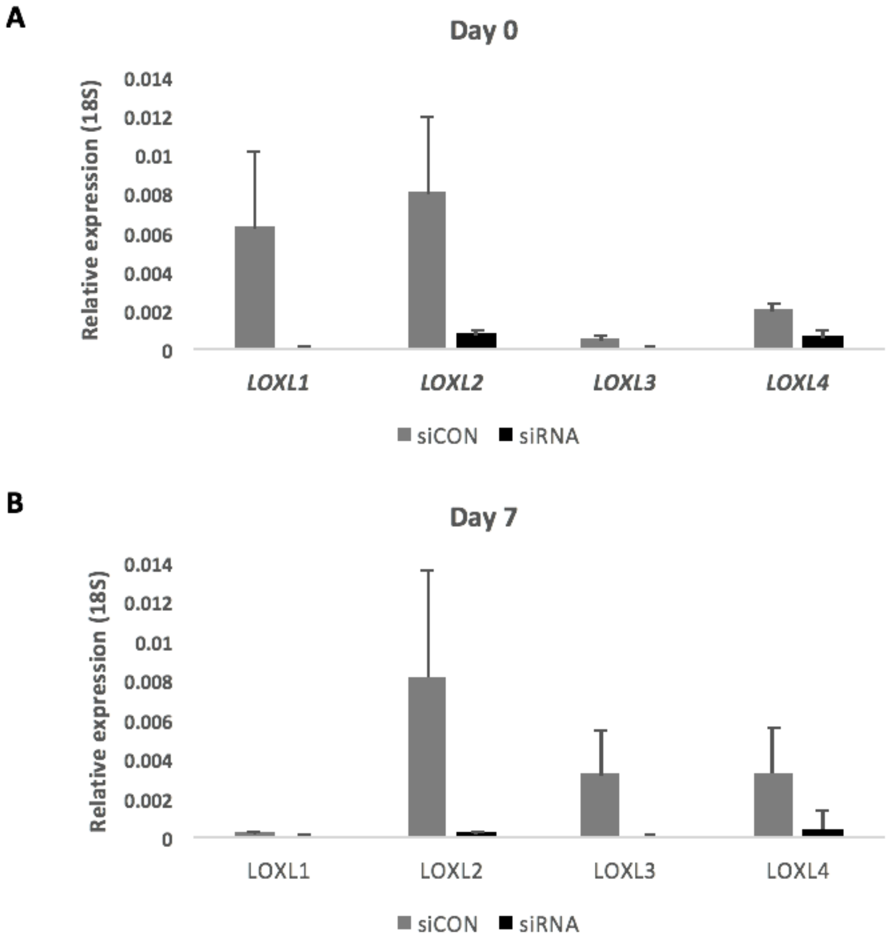


Figure 6.13 – Gene expression of LOXL1-4 in (A) hMSCs after transfection for 48hrs with siRNAs against LOXL1-4. (B) Gene expression of LOXL1-4 in day 7 chondrocytes; 48hr siRNA transfected hMSCs were differentiated in transwell inserts over 7 days in chondrogenic media. RNA was extracted and gene expression was measured using RT-qPCR after cDNA synthesis. Gene expression was normalised to an 18S control. Experiments performed at n = 2 (hMSC donor 2454e). Errors bars represent the standard error of the mean (SEM).

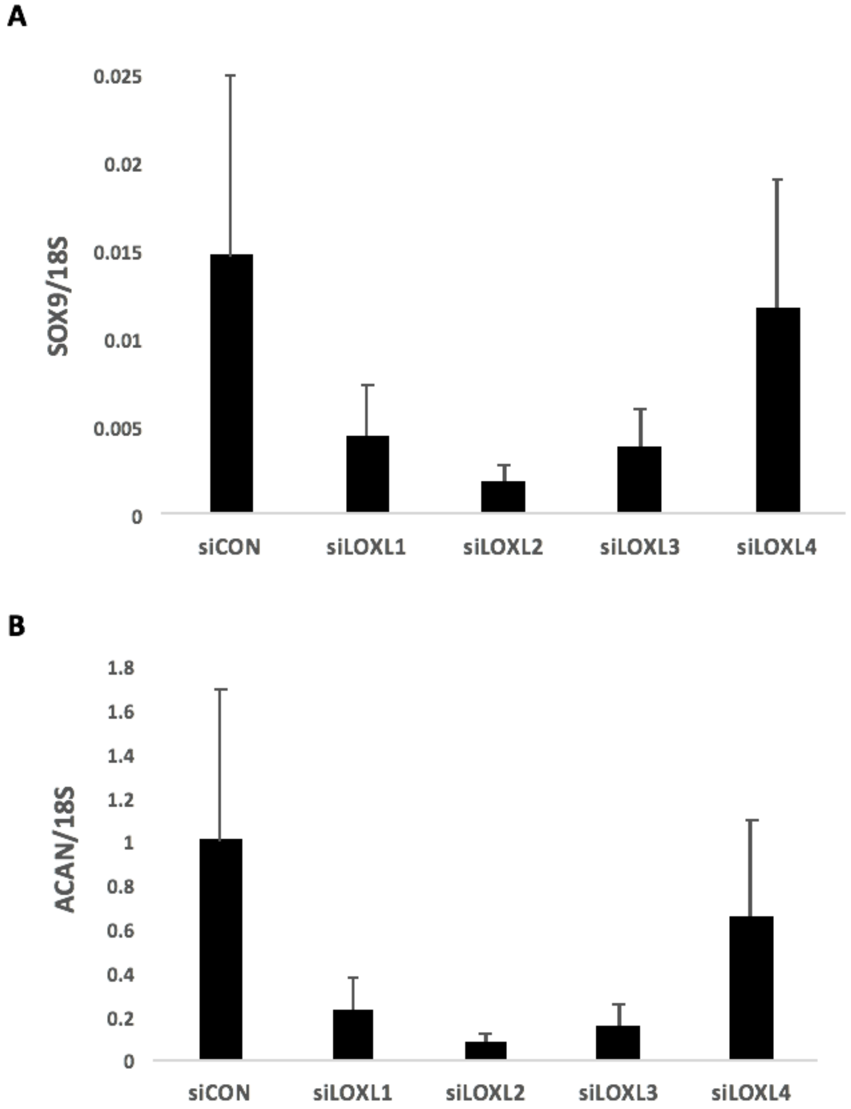


Figure 6.14 – Gene expression of (A) SOX9 and (B) ACAN at day 7 chondrogenesis after hMSC transfection with siRNAs against LOXL1-4 for 48hrs. RNA was extracted and gene expression was measured using RT-qPCR after cDNA synthesis. Gene expression was normalised to an 18S control. Experiments performed at n = 2 (hMSC donor 2454e). Errors bars represent the SEM.

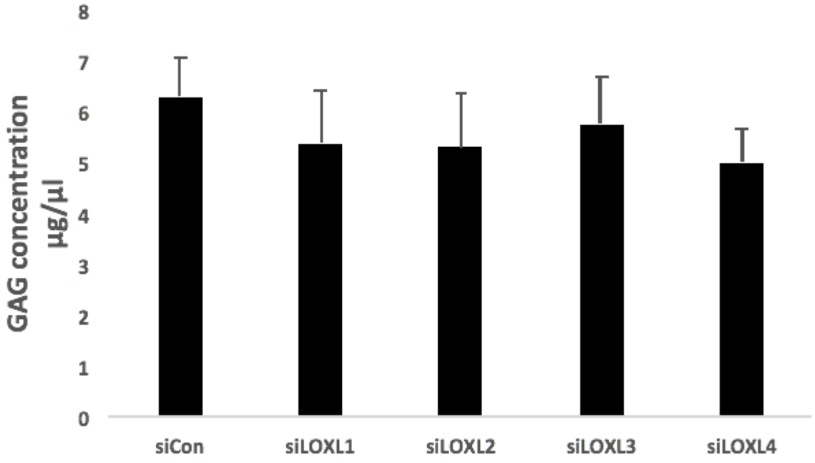


Figure 6.15 – GAG assay performed at day 7 of chondrogenesis after hMSC transfection with siRNAs against LOXL1-4 for 48hrs. hMSCs were differentiated into chondrocytes in pellet cultures. Experiments were performed at n = 6 (hMSC donor 2454e). Errors bars represent the standard deviation.

Despite *ACAN* being downregulated in chondrogenesis with siLOXL1, siLOXL2 and siLOXL3 treatment in hMSCs, concentrations of glycoaminoglycans did not change (Fig. 6.15). Although this could suggest the upregulation of other proteoglycans and glycoaminoglycans as a response, this may be due to the differences in the two chondrogenesis models used; gene expression changes were measured in RNA extracted from the transwell model whereas the GAG assay was performed using a pellet culture chondrogenesis model. To avoid discrepancies between models affecting outcomes, the same model should be used in future.

To summarize this section, siRNA knockdown of *LOXL1, LOXL2* and *LOXL3* in hMSCs appeared to result in downregulation of SOX9 and ACAN during chondrogenesis. This may suggest that *LOXL1-3* have a role in transcriptional regulation, previously undefined for *LOXL1* and *LOXL3*, or that collagen crosslinking in the ECM feeds back to regulate gene transcription. ECM properties are known to influence gene expression in chondrocyte differentiation (Allen et al, 2012). GAG concentrations did not change with siRNA treatment. The four *LOXL* genes have well defined roles in collagen crosslinking. Unfortunately, we were unable to quantify or measure collagen crosslinking within the timeframe of this project. These experiments were preliminary but the results merit further investigation to fully elucidate the roles of these four *LOXL* genes in chondrogenesis.

6.4 Discussion

De novo motif discovery using SOX9 peaks derived from lifted over ChIP-seq data revealed that the top second motif found belonged to the SOX family of proteins. However, the top motif displayed matches to other transcription factors such as SP-1 and EGR-1. These transcription factors - whilst important for chondrogenesis, are not exclusive to chondrogenesis or MSC differentiation in general. SP-1 and EGR-1 have roles in regulating differentiation of many other different types of stem cells. Transcription factor binding can help elucidate features that are specific to differentiation in general, MSC or otherwise, and binding that is exclusive to chondrogenesis. Motif searching has offered an insight into the intricate interplay of transcription factors involved in chondrogenesis. SOX9 has been shown to bind at gene promoters as well as gene enhancers (Ohba et al, 2015), with binding at enhancers leading to upregulation of chondrogenesis specific genes. The SOX9 protein functions as a homodimer and does not act alone (Bernard et al, 2003). SOX9 can also dimerize with other binding partners to regulate gene transcription. The majority of SOX9 binding in chondrocytes is accompanied by the JUN family of proteins, which makes up the AP-1 transcription factor along with its binding partner FOS. There is evidence that SOX9 forms a complex with other transcription factors first before binding to DNA as a protein complex (Jo et al, 2014). SOX9 can activate or repress genes at different stages of chondrogenesis depending on its binding partner. For example, SOX9 binds to GLI proteins to repress *COL10A1* prior to chondrocyte hypertrophy; *COL10A1* expression is a marker for chondrocyte hypertrophy. Hypertrophy of chondrocytes is the final stage of chondrocyte differentiation and precedes ossification (van der Kraan and van den Berg, 2012). *COL10A1* is known to be regulated by an enhancer (Chambers et al, 2002). Chondrocytes resident in permanent cartilage, such as the articular cartilage found in synovial joints, do not enter this terminal stage except in cartilage diseases, particularly osteoarthritis (von der Mark et al, 1992). The repression of *COL10A1* by SOX9-GLI heterodimers prevent cells from entering this stage (Leung et al, 2011). In contrast, SOX9 and AP-1 transcription factor binding at enhancers promotes hypertrophy (He et al, 2016). These studies highlight the importance of gene regulation through enhancers in chondrocyte development.

Although SOX9 and JUN binds to many chromatin states across the genome, binding in chondrocyte enhancers yielded the most GO terms related to chondrocyte differentiation. This is further evidence that gene enhancers are the main *cis*-regulatory feature driving chondrogenesis. We also observed that SOX9-JUN binding in chondrocyte quiescent state yielded GO terms related to development of other cells and tissues. More research is needed to elucidate how SOX9 and AP-1 binding sites change between cell types and processes.

Motif searching using transcription factor ChIP-seq peaks derived from lifted over reads (mm10 to hg38) yielded similar results as motif searching in the original mouse data (Ohba et al, 2015; He et al, 2016). Transcription factor DNA binding motifs are highly conserved between species (Hemberg and Kreiman, 2011). Liftover of genome coordinates from different species may not yield such synonymous results if the regions exhibit low synteny or if the species are not closely related. Furthermore, some loss of data is to be expected. Nonetheless, this shows that it is indeed possible to use data from another (albeit closely related) species to uncover biologically useful results. This has important implications for genomic research, and bioinformatics as a research tool. Although the cost of next-generating sequencing has dramatically decreased in recent years, it still remains a monumental task for one research group to gather all the datasets required for large scale projects. The sharing of datasets can facilitate ‘omics research and reduces the burden of generating data for a single group. Mice are widely used as a proxy for human research where the use of human cells and tissue is prohibitive. Many findings in mice can also be applied to humans, especially for highly conserved processes such as development (Breschi et al, 2017). The use of a mouse dataset in this analysis can therefore be considered analogous to the use of mouse models in laboratory based research.

The term “super enhancer” was coined in 2012 by Richard Young (Whyte et al, 2013). The same year, stretch enhancers, defined as long enhancers, were described by Francis Collins (Parker et al, 2013). Super enhancers and stretch enhancers may be largely synonymous. Since their identification, there has been much interest in super enhancers and how they define a cell type and contribute to diseases such as cancer (Loven et al, 2014; Niederriter et al, 2015). However, there is some controversy surrounding the existence of super enhancers. Critics have claimed that super enhancers do not exist and are simply a series of typical enhancers that happen to be in the same genomic region (Pott and Lieb, 2014). Super enhancers may be defined by different enhancer markers and therefore, depending on the markers used, reported super enhancers may be different even within the same cell type. There are online databases containing super enhancers identified in different cell types; however, the lack of consensus about what constitutes a super enhancer means these repositories can be considerably heterogeneous. Furthermore, the mechanism of how super enhancers act is currently unknown, contributing to the uncertainty of super enhancer definition. However, there have been many studies investigating super enhancers and their role in human development and disease and their existence has been well documented. There is no doubt that clusters of enhancers are present in the genome. Consequently, the question should not be whether they exist, but rather what do they do and how do they do it? There are two main mechanisms proposed for how super enhancers regulate gene transcription. It may be that each individual enhancer has an additive effect yet still show enhancer activity independent of other enhancers within the super enhancer (Hay et al, 2016). The other mechanism proposed is that individual enhancers within a super enhancer exhibit little to no independent enhancer activity but act synergistically with other neighbouring enhancers to regulate gene expression i.e. greater than the sum of its parts (Dukler et al, 2017). It could also be that different super enhancers have different mechanisms of action. Regardless of whether super enhancers constitute a separate class of enhancers, they have been shown to be important regulators of cell specific processes and therefore merit investigation.

In this study, enhancer luciferase reporter assays yielded mixed results. The entire super enhancers defined in our dataset were too large to clone into a vector and therefore only small sections were selected. Some regions of super enhancers exhibited independent enhancer activity suggesting that they do not require other enhancers within the same super enhancer to regulate gene transcription. However, others did not show any independent enhancer activity. Plasmid constructs are an artificial environment for gene enhancers and the plasmid DNA is not surrounded by histone proteins as is the case with endogenous chromatin. Vectors are unlikely to possess the correct endogenous chromatin domains and therefore the outcomes of enhancer luciferase reporter assays may not accurately reflect the process that occurs in endogenous chromatin. Furthermore, the cloned enhancer and luciferase gene may be a lot closer in the vector compared to the enhancer and its target promoter in the genome. Overexpression of *SOX9* did increase luciferase activity of some enhancer constructs in HEK293T cells but not in SW1353 cells. The presence of JUN binding peaks was also stipulated in our super enhancer identification criteria. Therefore, it will be interesting to observe whether overexpression of *JUN* (and its binding partner *FOS*) would also increase enhancer activity. Conversely, it will also be of interest to deplete *SOX9* and *JUN-FOS* to elucidate whether this decreases enhancer activity.

There are far more enhancers present in the genome than genes (Pennacchio et al, 2013) and genes may be regulated by more than one enhancer and one enhancer may regulate multiple genes (Mohrs et al, 2001). There is evidence that some enhancers are partially redundant and act together to fine tune gene expression (Moorthy et al, 2017). Redundancy in gene enhancers may reflect their role in driving important cell specific processes; loss of function of one gene enhancer may also be less detrimental if there are several backups. Gene enhancers generally are not evolutionarily conserved in mammals and evolve more rapidly than protein coding genes and gene promoters (Villar et al, 2015). Exceptions are enhancers involved in regulating development genes which are often highly conserved among vertebrates (Plessy et al, 2005; Visel et al, 2008).

As more and more data is being gathered by researchers, it is crucial to understand how data can be consolidated; it is also important to consider what is and what is not appropriate to integrate. Groups use different model organisms and systems to study a biological process and therefore, although data is being generated, it may not necessarily be suitable to apply the data to a different project. The use of high throughput chromatin conformation studies to study higher order chromatin structure is increasing in popularity due to cheaper next-generation sequencing (Schmitt et al, 2016). This technology is also used to validate and detect enhancer-promoter interactions. Hi-C and 4C studies have been performed on a number of cell types, but so far not in chondrocytes. Confirming enhancer-promoter interactions in this project has proved impossible without further relevant data. Although we have shown that some small sections of super enhancers show independent enhancer activity using luciferase reporter assays, it is unknown which genes they directly target.

We have consistently shown throughout this project that gene enhancers are cell type specific and distinct between cell types. Therefore, it was not considered entirely appropriate to integrate chromatin conformation data from other cell types to confirm chromatin interactions in chondrocytes. If enhancers are distinct between cell types and the main mechanism of enhancers is through physical looping of chromatin, then it follows that chromatin conformation will be different between cell types with different enhancers. Without chromatin conformation data or other validation, the next best approach is to link enhancers to the nearest gene or use other prediction methods. Some groups have proposed to link enhancers to the nearest upregulated or active gene as a better approach to linking to any nearest gene (Blattler et al, 2014; Zhang et al, 2013). Linking to the nearest upregulated gene necessitates the availability of an accompanying genome wide transcriptome dataset. Computational tools have been designed to aid the prediction of enhancers to target genes, authors of these tools have claimed high accuracy and reliability. However, predictive tools are just that and experimental validation of gene targets of enhancers is still required to pinpoint their gene target and ascertain its importance in the biological process being studied.

siRNA depletion of *LOXL1*, *LOXL2* and *LOXL3* during chondrogenesis led to downregulation of *SOX9* and *ACAN* gene expression; the decrease in *ACAN* expression may be an indirect effect of *SOX9* downregulation. This may point to an additional role of *LOXL1-3* as transcriptional regulators, potentially independent of their established role as collagen and elastin crosslinkers. Previous studies have found that the LOXL2 protein has a role in transcriptional regulation (Iturbide et al, 2014). However, the ECM in cartilage is not inert and can influence gene expression. Both ECM stiffness and mechanical stress affects gene expression and turnover of ECM proteins (Allen et al, 2012; Maldonado and Nam, 2013). Therefore, the downregulation of *SOX9* may be caused by abnormal collagen crosslinking when *LOXL1-3* are depleted. As *SOX9* regulates many chondrogenesis genes, we anticipate that other genes are also affected by *LOXL1-3* depletion. Further experiments are required to fully characterize this potential novel role of *LOXL1-3* genes as transcriptional regulators or to ascertain whether changes to collagen crosslinking is feeding back to regulate gene expression. *LOXL3* is predicted to be associated with a super enhancer. As siRNA depletion of mRNA levels of these genes led to decreased *SOX9* and *ACAN* expression, it will be interesting to investigate whether targeting the enhancers to disrupt gene expression would also lead to a downregulation of *SOX9* and other genes during chondrogenesis.

To conclude this chapter, we have utilized external cross-species datasets to aid the interrogation of our chondrogenesis histone ChIP-seq experiment. Using this integrated data, we identified chondrocyte super enhancers. Sections of super enhancers were confirmed to exhibit independent enhancer activity using luciferase reporter assays. We used siRNAs to deplete the *LOXL* family of genes and found that *LOXL1-3* may act to regulate the transcriptional programme necessary for chondrogenesis. Further investigations are required to fully elucidate super enhancer mechanisms and the genes they target.

6.5 Conclusion

- Analysis of mouse Sox9 and Jun ChIP-seq data lifted over to human coordinates achieved expected outcomes. Lifted over *SOX9* and *JUN* co-bind at strong promoters, strong enhancers and quiescent chromatin states.

- We identified 746 chondrocyte super enhancers, defined by strong enhancer states and JUN and SOX9 co-binding.

- Luciferase activity assays showed that some regions of super enhancers displayed independent enhancer activity

- siRNA depletion of *LOXL1*, *LOXL2* and *LOXL3* resulted in the downregulation of *SOX9* and *ACAN*, indicating that these genes could act as transcriptional regulators – a role not previously described for *LOXL1* or *LOXL3*.

Chapter 7. General discussion

7.1 Summary and main findings

The epigenome of an *in vitro* model of chondrogenesis was investigated by generating a histone modification ChIP-seq dataset (H3K4me3, H3K4me1, H3K27ac, H3K27me3 and H3K36me3). RNA-seq of the same chondrogenesis model was also generated. Histone mark enrichments correlated with the level of gene expression but a change in enrichment of single marks did not correlate with gene expression changes. The software ChromHMM was used to characterise 16 distinct chromatin states using the histone ChIP-seq dataset; *cis*-regulatory elements such as promoters and enhancers were included. Pathway analysis showed that enhancers were associated with cell type specific GO terms. Similarity analysis with Roadmap epigenomes found that enhancers are more distinct between cell types compared to other chromatin states. DNA 450k methylation arrays were performed for chondrogenesis, osteoblastogenesis and adipogenesis. Enhancers marked by H3K4me1 and H3K27ac were found to be de-methylated during chondrogenesis and osteoblastogenesis but not adipogenesis. Chondrocyte super enhancers were identified with the aid of publically available mouse Sox9 and c-Jun ChIP-seq datasets. Selected regions of chondrocyte super enhancers were cloned into PGL3-promoter vector and luciferase assays showed that some regions exhibited independent enhancer activity whilst others did not. The *LOXL* family of genes have previously defined roles as enzymes involved in collagen cross-linking. We observed that *LOXL3* may be associated with a super enhancer. We found that siRNA depletion of *LOXL1*, *LOXL2* and *LOXL3* resulted in the downregulation of *SOX9* and *ACAN* during chondrogenesis. The *LOXL* family of genes may potentially have a role in transcriptional regulation that is independent of their roles as collagen cross-linkers.

7.2 Limitations and caveats

We differentiated hMSCs into chondrocytes using a scaffold-free transwell model of chondrogenesis. Chondrocytes were isolated from the cartilage-like disc that forms in the transwell insert using enzyme digestion prior to chromatin extraction. However, enzymatic digestion at 37^{o}C can lead to altered gene expression (Hayman et al, 2006) and potentially histone modification changes. Enzyme incubation times were optimised to minimise potential disruption to cells. However, it is unknown if and to what extent the enzyme digestion steps had on the isolated chondrocytes. hMSC donors are variable and therefore multiple donors are required to minimise any individual donor effects. In this project, ChIP-seq was performed using two hMSC donors. Whilst two independent biological replicates are considered enough by ENCODE (Landt et al, 2012), other studies have found that three or more replicates reduces technical noise (Yang et al, 2014). Using additional hMSC donors and thus more replicates was not possible due to financial constraints. QC of ChIP-seq data found that one replicate was inferior to the other, possibly due to excess PCR amplification during the library preparation step.

It is impossible to confirm co-localisation of histone marks in the same cell in ChIP-seq data generated from a population of heterogeneous cells. Consequently, single cell or sequential ChIP-seq (ChIP using fragments from a previous ChIP) may be better options for assaying co-occurrence of histone marks and transcription factors in the same cell.

We observed that differential DNA methylation during chondrogenesis occurs in enhancer regions more than promoters. However, probes on the 450K array are biased towards promoters; less than 5% of probes were located in chondrocyte enhancers marked by H3K4me1 and H3K27ac. There are 28 million CpG sites in the human genome (Lövkvist et al, 2016). Therefore, genome-wide bisulfite sequencing is a more suitable method to assay all CpG sites in the genome. Furthermore, material used for chondrogenesis RNA-seq and DNA 450k array was collected from different experiments. Although the same chondrogenesis model was used, cells may be subject to slightly different conditions. Additionally, cells in a population may be at slightly different stages of chondrogenesis upon nucleic acid isolation.

To identify super enhancers, mouse rib chondrocyte Sox9 and c-Jun ChIP-seq reads were converted into human genome co-ordinates prior to peak calling. The MACS2 signal value of SOX9 peaks derived from mouse were used to identify super enhancers in differentiated chondrocytes. Whilst subsequent pathway analysis of chondrocyte super enhancers yielded GO terms associated with chondrogenesis, the use of mouse datasets in lieu of human is not ideal. A human SOX9 or JUN ChIP-seq dataset was not publically available. UCSC liftover between different genome assemblies relies on DNA sequence homology and synteny (Kuhn et al, 2012). The SOX family of transcription factors is highly conserved both at the protein level (Kamachi and Kondoh, 2013) as well as the gene level (Jager et al, 2011) including the consensus binding motif. Our liftover and motif analysis confirmed the presence of the SOX9 (and JUN) motif in syntenic human genomic regions. However, lifted over data only encompasses SOX9 (and JUN) peaks that are conserved between human and mouse. The use of data in this way naturally excludes human specific binding sites and the mouse specific data is lost during the liftover process. Whilst we acknowledge this disadvantage and the limitations of the method, it can also be argued that the conserved binding sites between human and mouse are the most important for development as evolutionary conservation is a positive indicator of functional importance (Georgi et al, 2013).

Another chondrogenesis histone ChIP-seq dataset is available from Roadmap Epigenomics, generated from a 3D alginate scaffold model of chondrogenesis. The associated publication became available early on during this project (Herlofsen et al, 2013) and we contacted the authors for access to the raw data. However, this was not provided and did not become publically available until the publication of Roadmap's flagship paper in 2015. We have provided integrated DNA 450k methylation data and histone ChIP-seq of the same model of chondrogenesis. Whilst we could have used the existing chondrogenesis histone ChIP-seq data once it was released, integrating data from the same model from the same laboratory reduces batch effects and eliminates the presence of model specific effects. Herlofsen et al also attempted to integrate DNA methylation and cDNA microarray data with their histone ChIP-seq data. They observed a correlation between potential H3K4me1 and H3K27ac marked enhancers with gene expression but did not observe a correlation of DNA methylation with gene expression. However, Herlofsen et al used reduced representation bisulfite sequencing (RRBS) to quantify DNA methylation which only effectively assays CpG dense regions like promoters. This is in agreement with our analysis as we also did not observe major DNA methylation changes at gene promoters. Crucially, Herlofsen et al did not integrate the RRBS data with the histone ChIP-seq data, although it is unclear whether DNA methylation at enhancers would be detected using the RRBS method. Therefore, our study offers a valuable insight into the connection between histone modifications and the DNA methylome during chondrogenesis. Despite being from a different *in vitro* model, similarity analysis showed that our chondrocyte enhancers clustered with chondrocyte enhancers from Herlofsen et al. This shows that there is a distinct chondrocyte epigenome which can be identified regardless of culture methods and conditions. However, neither the Herlofsen et al. study nor this project defined a complete reference epigenome for human chondrocytes according to new guidelines from the International Human Epigenome Consortium (IHEC) due to the lack of H3K9me3 mark in our study and absence of whole genome bisulfite sequencing (WGBS) in both studies. IHEC consists of multiple international epigenomics consortia such as ENCODE, Roadmap Epigenomics, European BLUEPRINT, Canadian Epigenetics Environment and Hong Kong Epigenomics project (Stunnenberg et al, 2016). IHEC aims to co-ordinate the efforts of the various member consortia to streamline epigenomics research by defining global guidelines and sharing data, methods and bioinformatics tools. Consequently, WGBS is required in chondrocytes to satisfy the IHEC stipulations for a minimal reference epigenome.

7.3 Future work

In this project, potential enhancers and super enhancers have been identified. Future work includes validation of enhancer-promoter pairings and possible modulation of chondrogenesis by targeting *cis*-regulatory elements. The CRISPR-Cas9 genome editing system may be modified to direct transcriptional co-activators or repressors to chosen loci in the genome. This may also help to elucidate the mechanism of how super enhancers regulate transcription. Chromatin conformation assays can be used to confirm predicted connections and discover new interactions. Single interactions may be assayed using the 3C method or genome wide using Hi-C.

Further computational work could include network analysis of enhancers to elucidate the association between different enhancers and the genes they potentially regulate. This could also classify genes that are regulated by similar epigenetic mechanisms. Some regions of the genome originally thought to be non-coding are in fact, transcribed. There is evidence that enhancers may be transcribed (Li et al, 2016). Enhancer RNAs (eRNAs) are RNA molecules transcribed from the nucleotide sequence of enhancers and may regulate genes both in *cis* and *trans* (Lam et al, 2014). Identification of eRNAs in chondrocytes and investigation into whether they are transcribed may be incorporated into future studies. We have investigated briefly the DNA methylation changes in chromatin states of osteoblasts and adipocytes during differentiation of hMSCs. Further larger studies involving more samples and other differentiation processes would be of interest to elucidate the mechanisms ubiquitous to all differentiation processes, and those unique to chondrogenesis. Traditional DNA methylation assays such as the 450k array do not distinguish between 5mC and 5hmC. The 5hmC mark is associated with gene activation during chondrogenesis (Taylor et al, 2016). Therefore, it will be interesting to include this data for more complete picture of the epigenomic changes during chondrogenesis.

We had hoped to visualise and quantify collagen crosslinking after depletion of *LOXL1-4* genes but this was not feasible within the timeframe of this project. Therefore, subsequent studies could use techniques such as atomic force microscopy to visualise collagen fibrils and mass spectrometry methods to quantify crosslinking (Eyre et al, 2008).

Performing ChIP-seq in an *in vitro* model of chondrogenesis opens up the possibility of further ChIP-seq studies in normal human articular chondrocytes or diseased chondrocytes. Comparisons between normal and diseased chondrocytes may uncover the epigenetic mechanisms associated with disease. With this project, we have built the foundations for other studies into the epigenetics of chondrocytes.

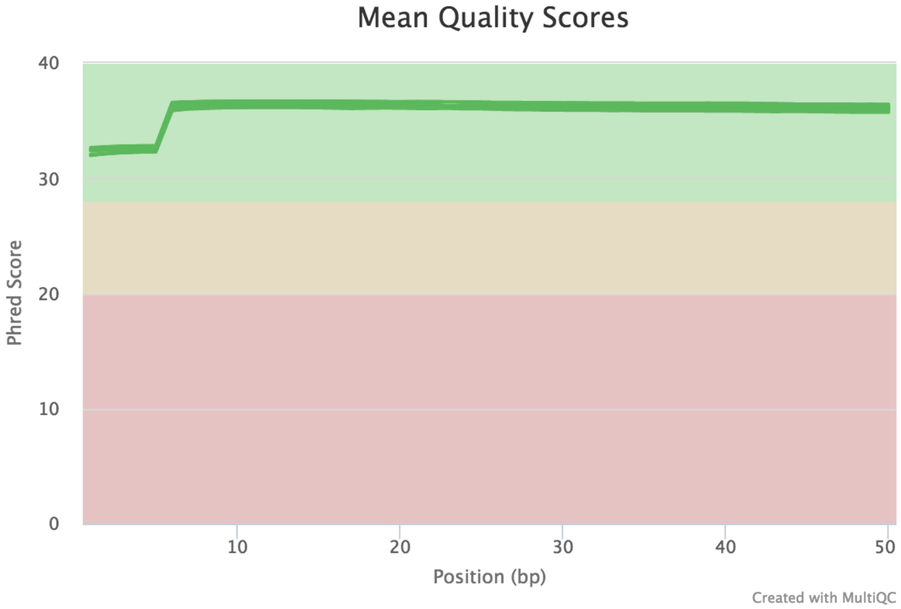
7.4 Contribution to epigenomics research

Genome-wide methods involving high throughput sequencing are becoming more widespread and there are many consortia established to generate and share sequencing data. Although a histone ChIP-seq dataset for differentiated chondrocytes exists as part of the Epigenomics Roadmap project, this project offers a more in depth investigation into the chondrocyte epigenome. We hope to publish this work and make our data publically available so other research groups can benefit from the findings.

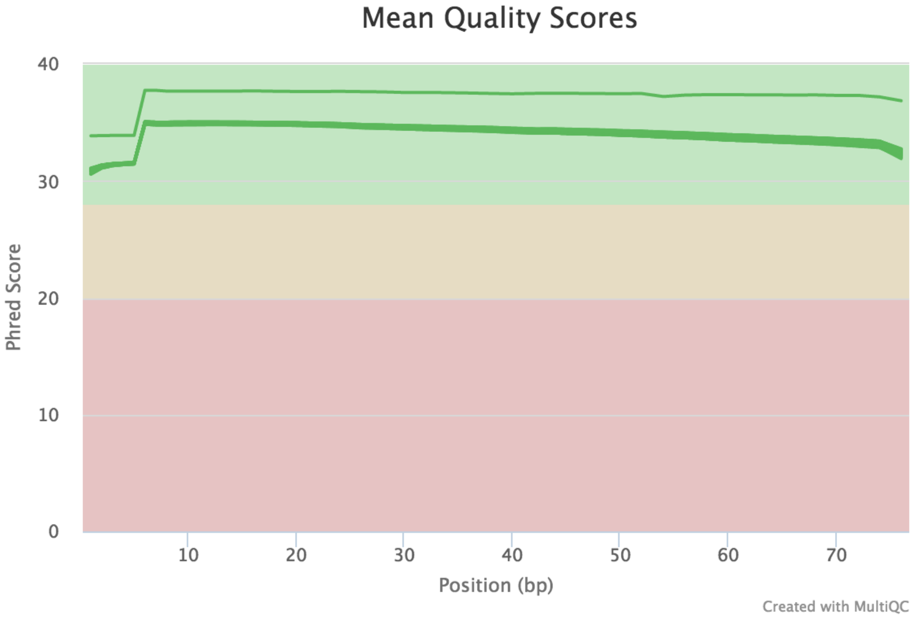
The general areas of biology this project belongs to includes molecular biology, genomics and its subfield, epigenomics. How epigenetic mechanisms regulate biological processes and contribute to disease has received increasing attention over the past 20 years (Ebrahim, 2012). The non-coding majority of the genome was previously thought to contain noise of no biological importance. It is now clear that so called “junk DNA” is not junk at all and contains features necessary for the regulation of gene transcription (Pallazo and Lee, 2015) and also has implications for genome evolution (Juan et al, 2013). Epigenetic marks can be found in both coding and non-coding regions of the genome. This is the first study, to our knowledge, to observe a global de-methylation of enhancers during chondrogenesis. Super enhancers are reported to target genes involved in cell identity. In this project, we have identified super enhancers in differentiated chondrocytes. Chondrocyte super enhancers are not currently part of the repertoire of super enhancers present in the dbSUPER database (Khan and Zhang, 2016). Once published, we can recommend that these be included in future releases.

This PhD project has contributed to the body of literature documenting the epigenetic marks and regulatory elements present within the genome. This project utilised a bioinformatics approach to analyse, integrate and interpret data. Modern bioinformatics is a relatively new and emerging field and is evolving at a fast rate. Bioinformatics was originally seen as a biological tool rather than a field in its own right (Hagen, 2000). However, bioinformatics is now ubiquitous in biology and has become a fundamental discipline (Kanehisa and Bork, 2003; Hogeweg, 2011). An important distinction must be made between a bioinformatics researcher or scientist, and a bioinformatics developer or engineer. The bioinformatics researcher applies tools to analyse biological data, discover new information, and generate or confirm hypotheses. In contrast, bioinformatics developers generally focus on designing and building new tools (Smith, 2015). Whilst there can be overlap between the two, most bioinformaticians can be broadly defined as one or the other and the skillsets required for each role are different. This project used and adapted existing bioinformatics tools to answer biological questions rather than developing a new tool or resource. However, this project has illustrated how integral bioinformatics methods are to ‘omics research.

Appendix i



Appendix I Figure 1 – MultiQC report of read quality scores for all histone ChIP-seq samples from hMSC donor 8a (replicate 1). All read samples achieved a quality score above 30.



Appendix I Figure 2 – MultiQC report of read quality scores for all histone ChIP-seq samples from hMSC donor 2454e (replicate 2). The mean quality score across samples was between 30 and 40.

Appendix I Table 1 – MutiQC report of percentage duplication, GC content and total reads of histone ChIP-seq samples from hMSC donor 8a (replicate 1).

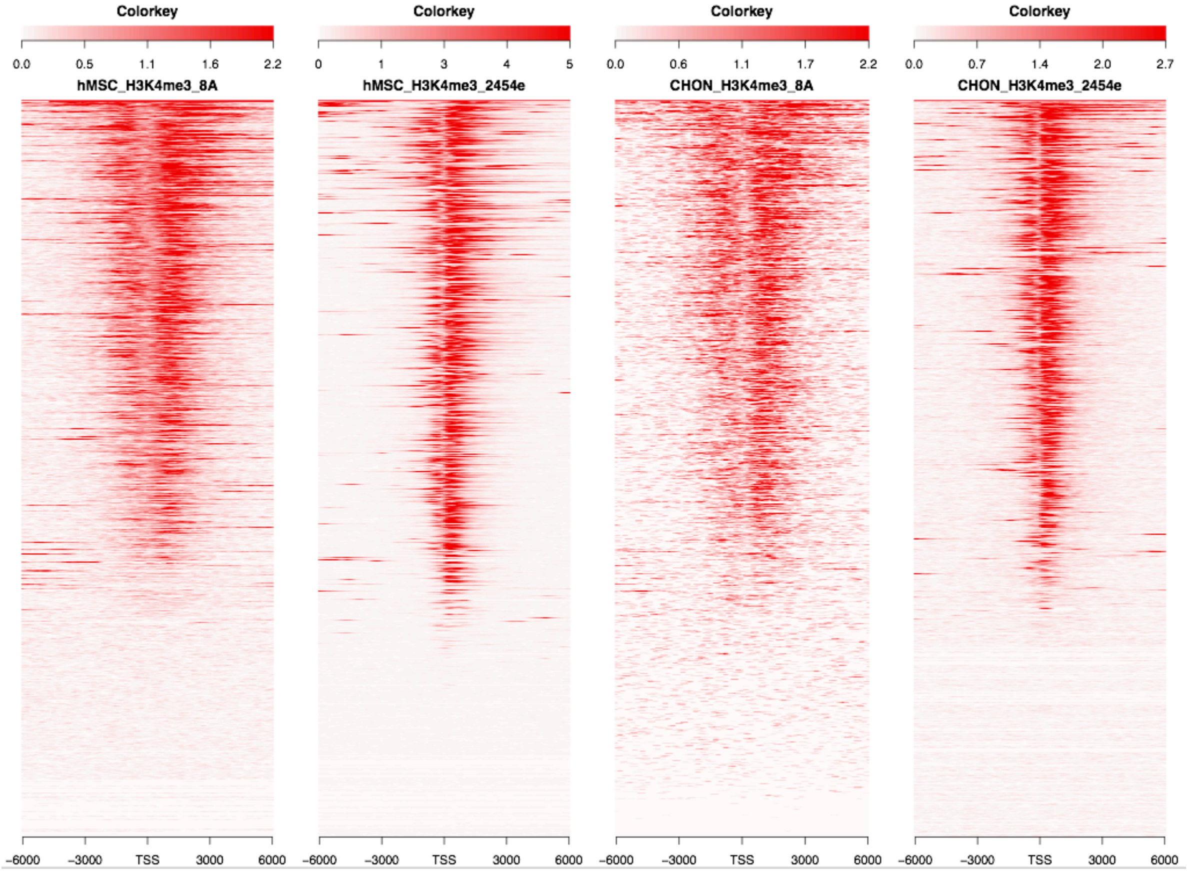
Sample Name	Sample ID	% Dups	% GC	M Seqs
CHON H3K4me1	CHE_sample_10_CAGATCTG_L007_R1_001	73.6%	39%	57.6
CHON H3K27ac	CHE_sample_11_GCCAATGT_L007_R1_001	44.2%	38%	52.8
CHON H3K27me3	CHE_sample_12_CTTGTACT_L004_R1_001	92.6%	41%	56.2
CHON H3K36me3	CHE_sample_13_CTTGTACT_merged	91.5%	39%	64.5
CHON IgG	CHE_sample_14_CTTGTACT_L007_R1_001	41.5%	40%	54.2
MSC input	CHE_sample_1_TGACCACT_merged	60.6%	36%	54.8
MSC H3K4me3	CHE_sample_2_TGACCACT_L004_R1_001	40.0%	41%	54.4
MSC H3K4me1	CHE_sample_3_CAGATCTG_L006_R1_001	81.9%	39%	65.9
MSC H3K27ac	CHE_sample_4_GCCAATGT_L006_R1_001	25.9%	42%	43.8
MSC H3K27me3	CHE_sample_5_GCCAATGT_L004_R1_001	89.2%	36%	56.5
MSC H3K36me3	CHE_sample_6_GCCAATGT_merged	91.7%	39%	56.1
MSC IgG	CHE_sample_7_CTTGTACT_L006_R1_001	26.3%	40%	46.3
CHON input	CHE_sample_8_CAGATCTG_merged	74.3%	38%	61.9
CHON H3K4me3	CHE_sample_9_CAGATCTG_L004_R1_001	81.6%	40%	55.9

Appendix I Table 2 - MultiQC report of percentage duplication, GC content and total reads of histone ChIP-seq samples from hMSC donor 2454e (replicate 2)

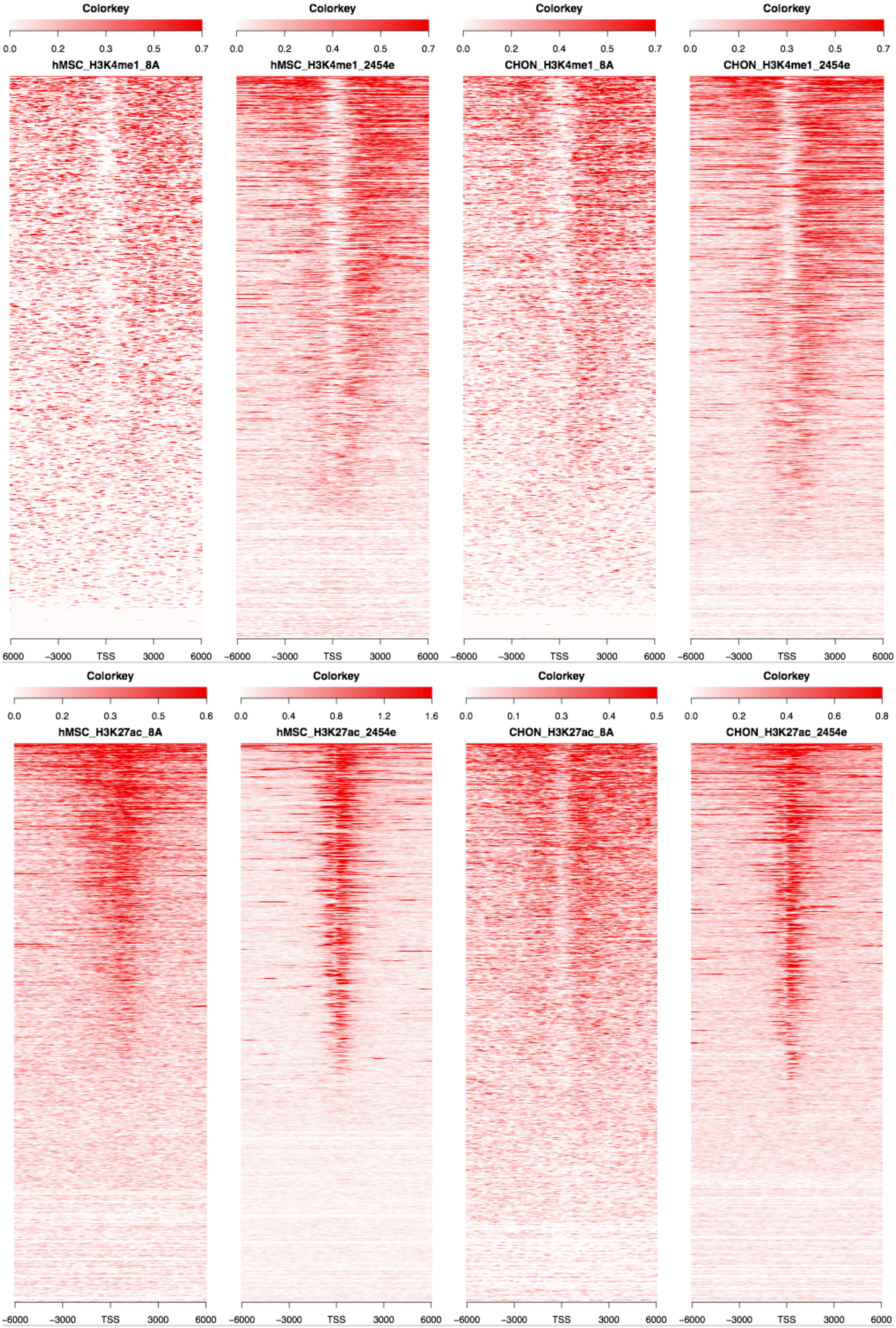
Sample Name	Sample ID	% Dups	% GC	M Seqs
CHON H3K4me3 (reseq)	08_S1_L001_R1_001	4.5%	44%	18.6
MSC input	KC01-David-Young_S1_L001_R1_001	2.9%	40%	9.7
MSC input	KC01-David-Young_S1_L002_R1_001	2.9%	40%	9.6
MSC input	KC01-David-Young_S1_L003_R1_001	3.0%	40%	10.1
MSC input	KC01-David-Young_S1_L004_R1_001	3.0%	40%	10.1
MSC H3K4me3	KC02-David-Young_S2_L001_R1_001	4.5%	48%	20.7
MSC H3K4me3	KC02-David-Young_S2_L002_R1_001	4.5%	48%	20.6
MSC H3K4me3	KC02-David-Young_S2_L003_R1_001	4.7%	48%	21.5
MSC H3K4me3	KC02-David-Young_S2_L004_R1_001	4.7%	48%	21.4
MSC H3K4me1	KC03-David-Young_S3_L001_R1_001	1.7%	42%	7.9
MSC H3K4me1	KC03-David-Young_S3_L002_R1_001	1.6%	42%	7.8
MSC H3K4me1	KC03-David-Young_S3_L003_R1_001	1.7%	43%	8.1
MSC H3K4me1	KC03-David-Young_S3_L004_R1_001	1.7%	43%	8.1
MSC H3K27ac	KC04-David-Young_S4_L001_R1_001	2.1%	44%	9.3
MSC H3K27ac	KC04-David-Young_S4_L002_R1_001	2.1%	44%	9.3
MSC H3K27ac	KC04-David-Young_S4_L003_R1_001	2.3%	44%	9.7
MSC H3K27ac	KC04-David-Young_S4_L004_R1_001	2.2%	44%	9.6
MSC H3K27me3	KC05-David-Young_S5_L001_R1_001	3.8%	43%	13.1
MSC H3K27me3	KC05-David-Young_S5_L002_R1_001	3.8%	43%	13.0
MSC H3K27me3	KC05-David-Young_S5_L003_R1_001	3.9%	43%	13.6
MSC H3K27me3	KC05-David-Young_S5_L004_R1_001	3.9%	43%	13.6
MSC H3K36me3	KC06-David-Young_S6_L001_R1_001	4.9%	43%	16.0
MSC H3K36me3	KC06-David-Young_S6_L002_R1_001	5.0%	43%	15.9
MSC H3K36me3	KC06-David-Young_S6_L003_R1_001	5.0%	43%	16.6
MSC H3K36me3	KC06-David-Young_S6_L004_R1_001	4.9%	43%	16.5
CHON input	KC07-David-Young_S7_L001_R1_001	2.9%	40%	10.4
CHON input	KC07-David-Young_S7_L002_R1_001	2.8%	40%	10.2
CHON input	KC07-David-Young_S7_L003_R1_001	2.9%	40%	10.7
CHON input	KC07-David-Young_S7_L004_R1_001	2.9%	40%	10.6
CHON H3K4me3	KC08-David-Young_S8_L001_R1_001	1.8%	44%	3.4
CHON H3K4me3	KC08-David-Young_S8_L002_R1_001	1.8%	44%	3.3
CHON H3K4me3	KC08-David-Young_S8_L003_R1_001	1.9%	44%	3.5
CHON H3K4me3	KC08-David-Young_S8_L004_R1_001	1.8%	44%	3.5
CHON H3K4me1	KC09-David-Young_S9_L001_R1_001	2.5%	42%	13.3
CHON H3K4me1	KC09-David-Young_S9_L002_R1_001	2.4%	42%	13.2
CHON H3K4me1	KC09-David-Young_S9_L003_R1_001	2.6%	42%	13.9
CHON H3K4me1	KC09-David-Young_S9_L004_R1_001	2.5%	42%	13.8
CHON H3K27ac	KC10-David-Young_S10_L001_R1_001	2.6%	42%	13.5
CHON H3K27ac	KC10-David-Young_S10_L002_R1_001	2.6%	42%	13.3
CHON H3K27ac	KC10-David-Young_S10_L003_R1_001	2.7%	42%	14.0
CHON H3K27ac	KC10-David-Young_S10_L004_R1_001	2.7%	42%	13.9
CHON H3K27me3	KC11-David-Young_S11_L001_R1_001	3.5%	42%	9.7
CHON H3K27me3	KC11-David-Young_S11_L002_R1_001	3.5%	42%	9.6
CHON H3K27me3	KC11-David-Young_S11_L003_R1_001	3.7%	42%	10.0
CHON H3K27me3	KC11-David-Young_S11_L004_R1_001	3.6%	42%	9.9
CHON H3K36me3	KC12-David-Young_S12_L001_R1_001	6.1%	41%	13.0
CHON H3K36me3	KC12-David-Young_S12_L002_R1_001	6.2%	41%	12.9
CHON H3K36me3	KC12-David-Young_S12_L003_R1_001	6.3%	41%	13.5
CHON H3K36me3	KC12-David-Young_S12_L004_R1_001	6.2%	41%	13.5

Appendix I Table 3 – Read numbers for ChIP-seq samples from both replicates.

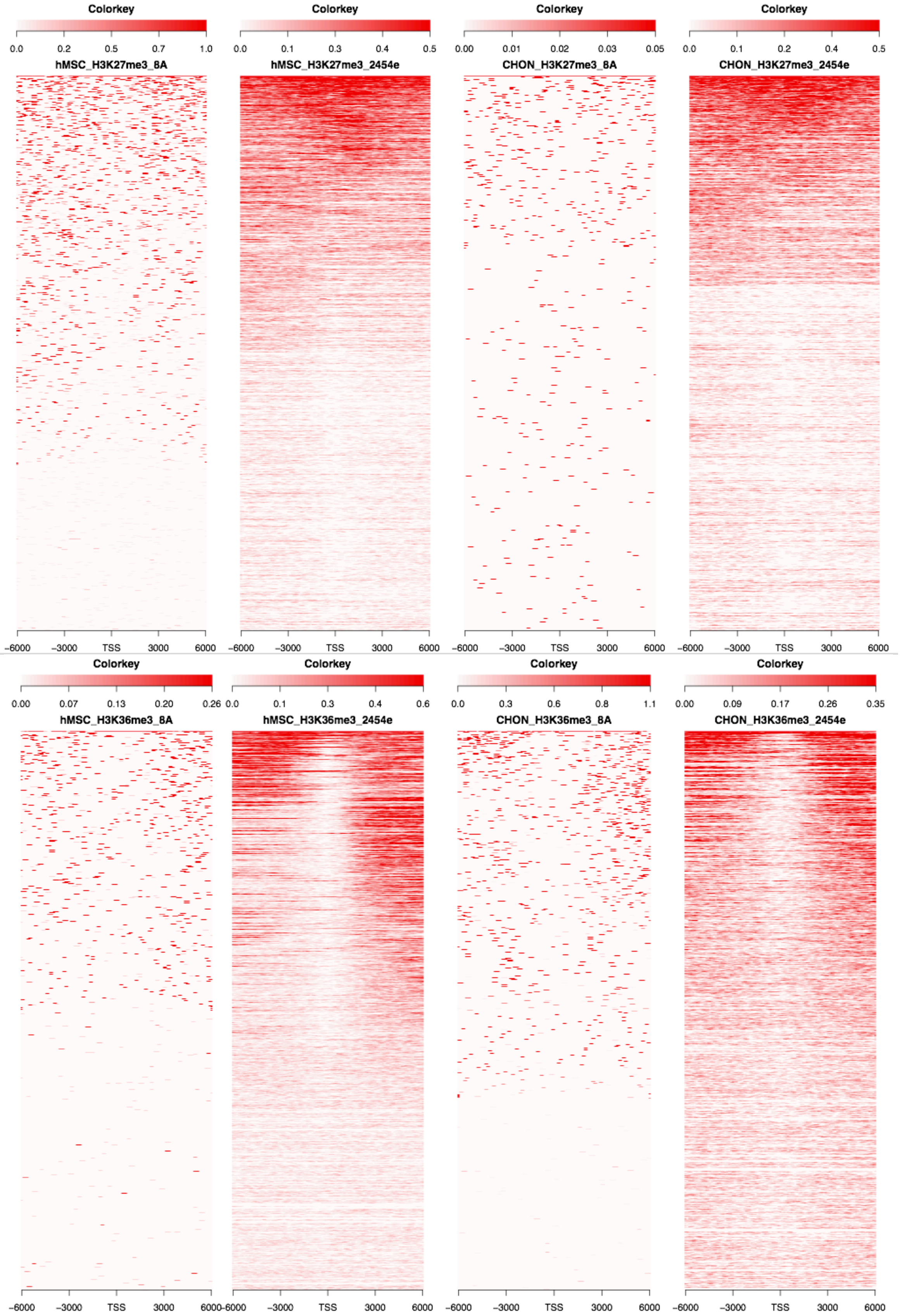
Replicate	**Sample name**	**total reads**	**reads mapped**	**% mapped**	**RPM scale**	**Duplicate mapped reads**	**Uniquely mapped reads**
1	MSC_input	54781975	53567485	97.78	0.018668041	32337393	21230092
1	MSC_H3K4me3	54445014	53322473	97.94	0.018753819	21879648	31442825
1	MSC_H3K4me1	65871548	59759649	90.72	0.016733699	50705654	9053995
1	MSC_H3K27ac	43783953	43190153	98.64	0.023153426	11186827	32003326
1	MSC_H3K27me3	56478748	40585867	71.86	0.024639119	37171899	3413968
1	MSC_H3K36me3	56134283	30653142	54.61	0.032623083	29045392	1607750
1	CHON_input	61900480	59199642	95.64	0.016891994	43738430	15461212
1	CHON_H3K4me3	55885851	53115817	95.04	0.018826784	44172842	8942975
1	CHON_H3K4me1	57644630	54247873	94.11	0.018433902	41224176	13023697
1	CHON_H3K27ac	52782375	51879594	98.29	0.019275401	23602573	28277021
1	CHON_H3K27me3	56168560	21819215	38.85	0.045831163	20627085	1192130
1	CHON_H3K36me3	64468380	39477404	61.24	0.025330946	36724666	2752738
2	MSC_input	39546203	38857682	98.26	0.025734937	1634899	37222783
2	MSC_H3K4me3	84082773	82398018	98	0.012136214	11644251	70753767
2	MSC_H3K4me1	31948658	31379659	98.22	0.031867778	1115069	30264590
2	MSC_H3K27ac	37842038	37118603	98.09	0.026940669	2006616	35111987
2	MSC_H3K27me3	53281894	52109519	97.8	0.019190352	2973659	49135860
2	MSC_H3K36me3	65120119	63548834	97.59	0.01573593	4111910	59436924
2	CHON_input	41956694	41325393	98.5	0.024198197	1893984	39431409
2	CHON_H3K4me3	32248437	31633522	98.09	0.031612035	1382639	30250883
2	CHON_H3K4me1	54224635	53290514	98.28	0.018765066	3080637	50209877
2	CHON_H3K27ac	54669205	53504496	97.87	0.018690018	3064249	50440247
2	CHON_H3K27me3	39173561	38335139	97.86	0.026085728	1966391	36368748
2	CHON_H3K36me3	52824567	51546571	97.58	0.019399933	3620732	47925839



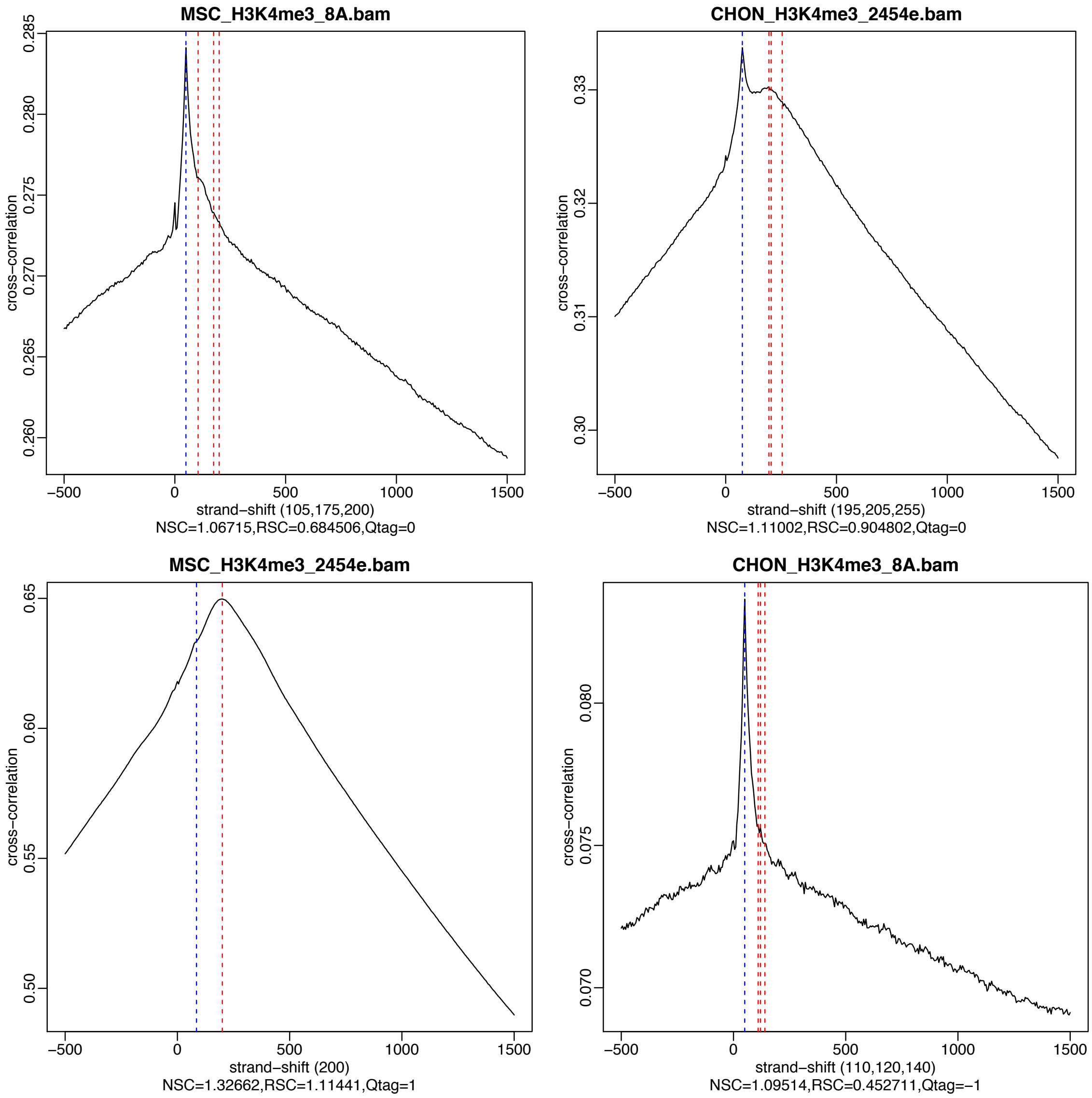
Appendix I Figure 3 – H3K4me3 read density plot. H3K4me3 reads are mainly located close to the TSS of genes.



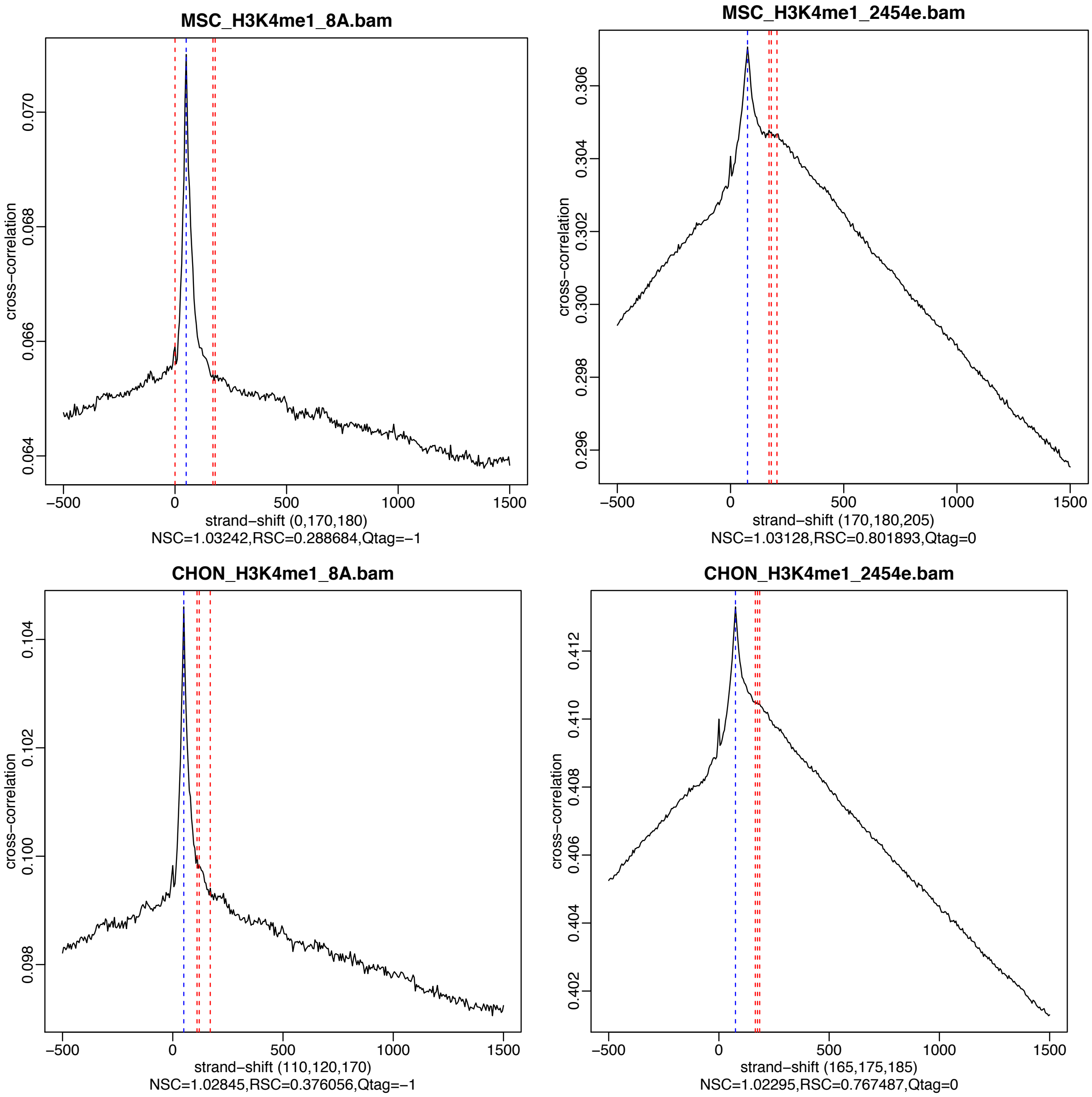
Appendix I Figure 4 – Read densities for enhancer marks H3K4me1 and H3K27ac. High densities of H3K4me1 sample reads are located on either side of the TSS. H3K27ac reads were seen at the TSS as well as around the TSS.



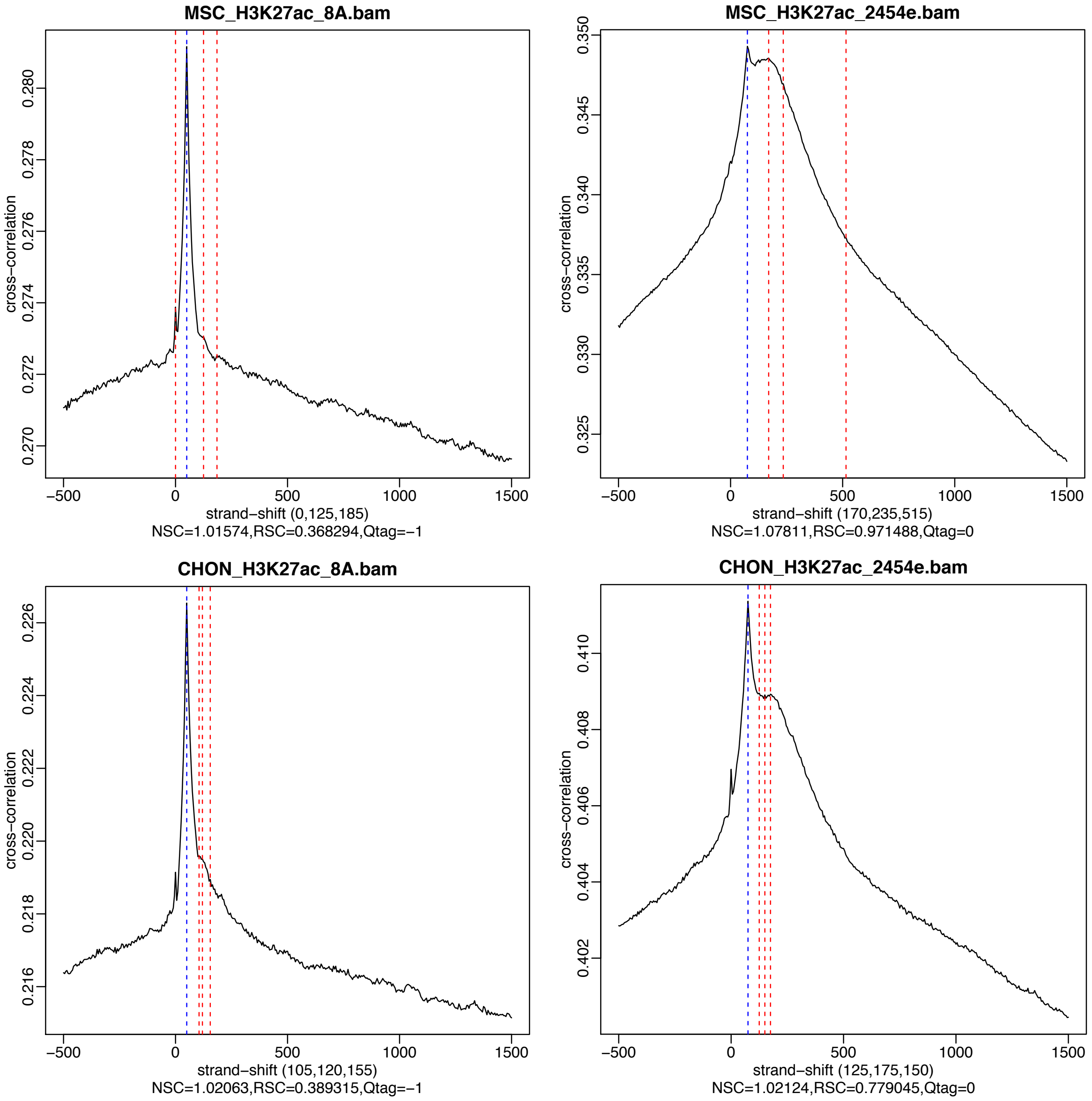
Appendix I Figure 5 – Read densities of H3K27me3 and H2K36me3 samples. Samples from donor 071508A had fewer reads than 2454e in the region shown (+/- 6000bp from TSS).



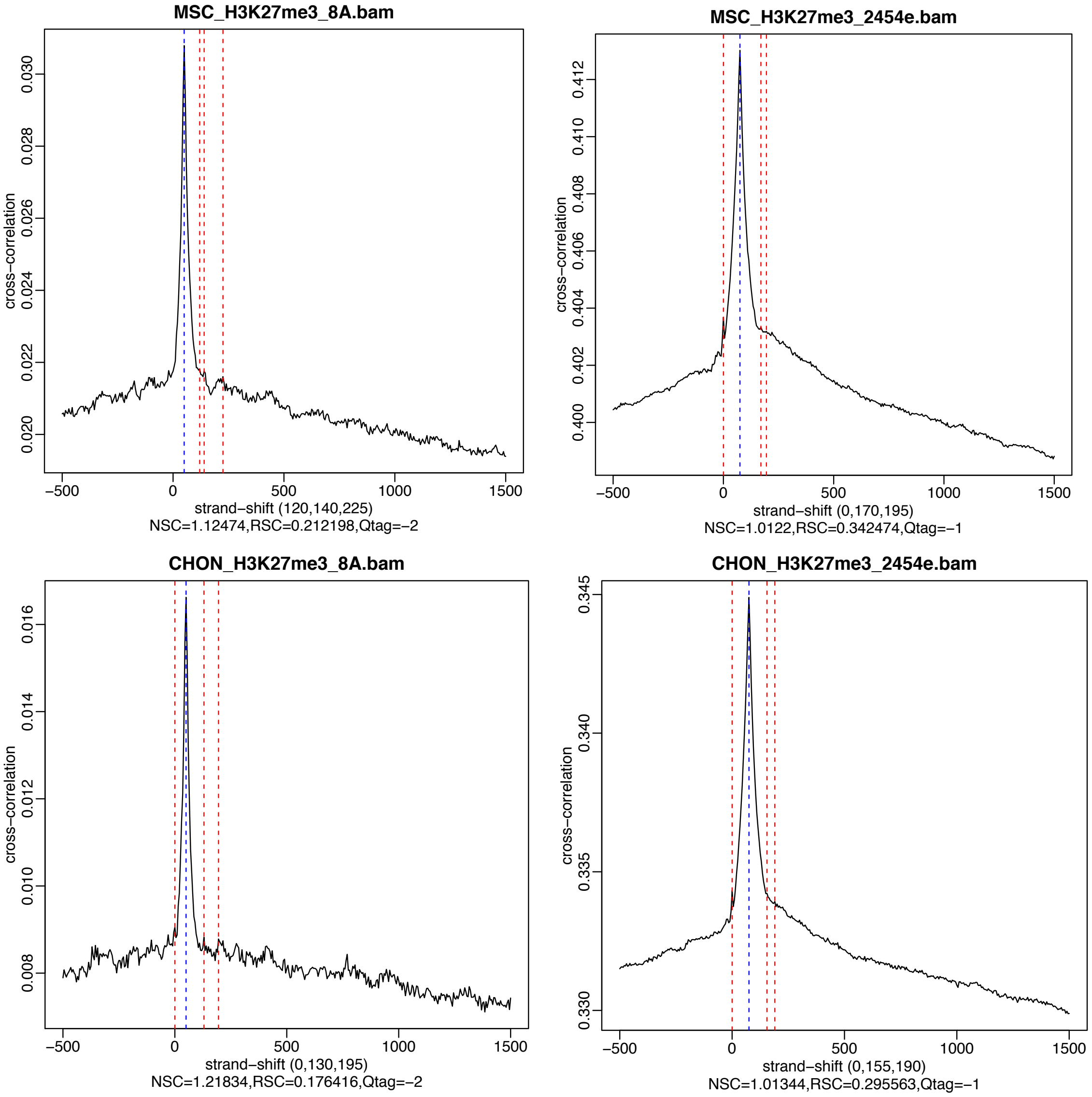
Appendix I Figure 6 – NSC and RSC plots for H3K4me3 sample. The blue dotted line shows the phantom peak, corresponding to read length. The red dotted lines show the best estimations of fragment length, with values given in brackets below the plot. NSC, RSC and quality values are also given below the plots. Qtags are based on RSC values and ranges from -2 to 2, with higher values indicating better quality



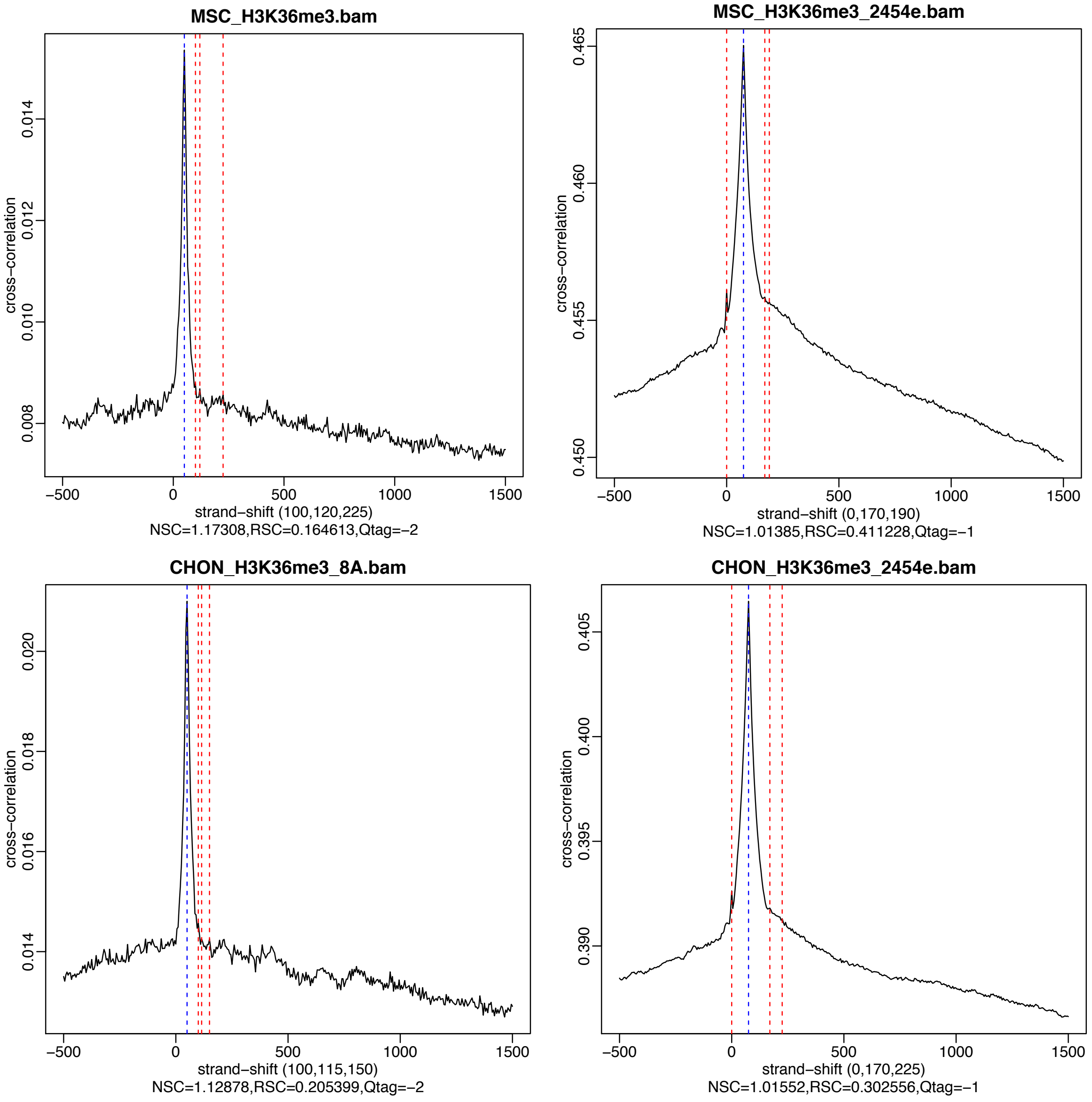
Appendix I Figure 7 – NSC and RSC plots for H3K4me1 samples



Appendix I Figure 8 – NSC and RSC plots for H3K27ac samples

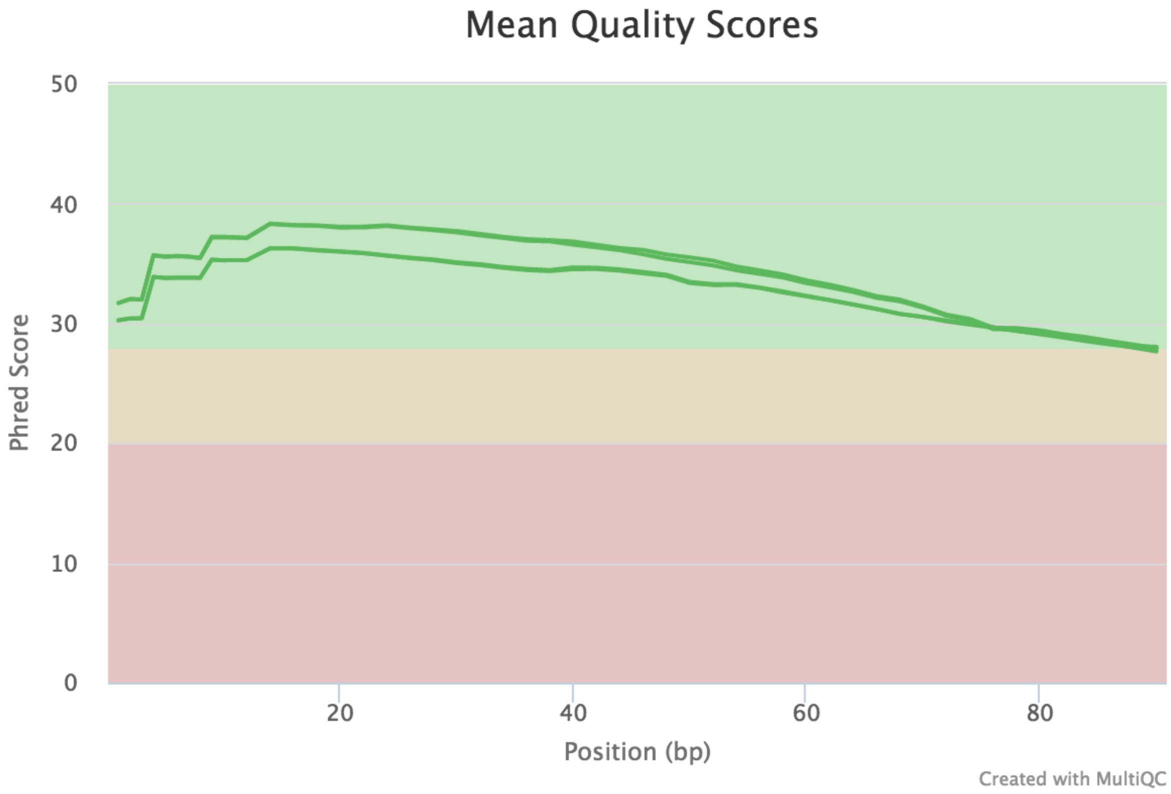


Appendix I Figure 9 – NSC and RSC plots for H3K27me3 samples



Appendix I Figure 10 – NSC and RSC plots for H3K36me3 plots

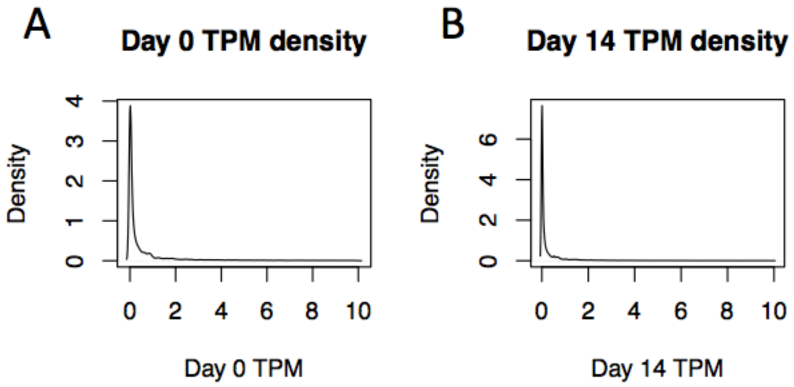
Appendix ii



Appendix ii Figure 1 – Quality scores after trimming for chondrogenesis RNA-seq samples. FastQC was performed individually for each sample before summarising using MultiQC

Appendix ii Table 1 – MultiQC general statistic report of chondrogenesis RNA-seq paired end reads after trimming

SAMPLE NAME	% DUPS	% GC	M SEQS
TRIMMED_DAY0_1	54.4%	50%	52.8
TRIMMED_DAY0_2	55.8%	52%	52.8
TRIMMED_DAY14_1	68.6%	52%	47.5
TRIMMED_DAY14_2	69.3%	54%	47.5



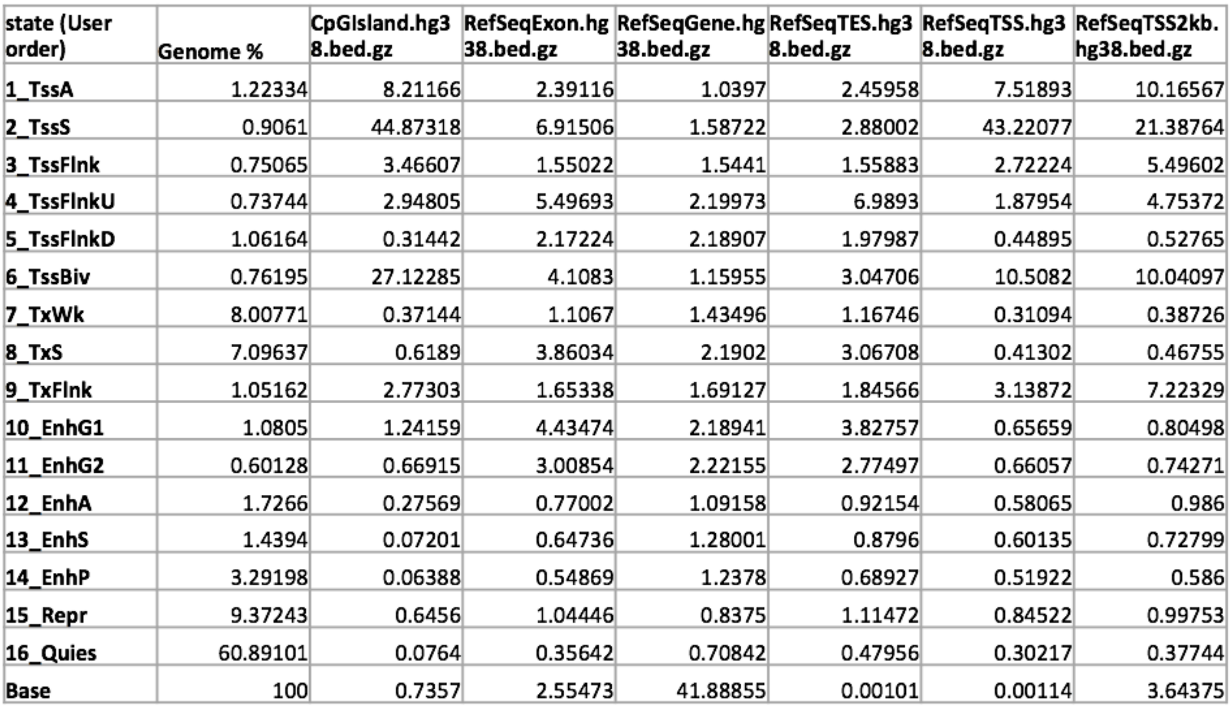
Appendix ii Figure 2 – Gene expression (TPM) density plots at for chondrogenesis samples day 0 (A) and day 14 (B)

Appendix iii

Appendix iii Table 1 – Chromatin state emission parameters. Output by ChromHMM

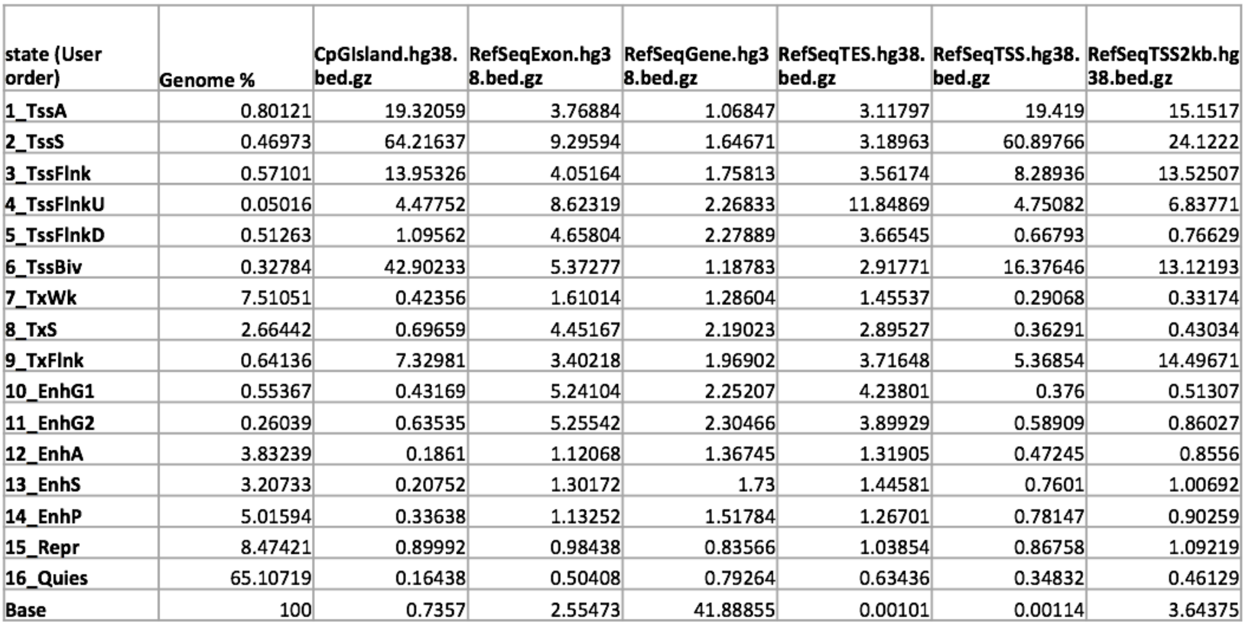
state (User order)	H3K4me3	H3K4me1	H3K27ac	H3K27me3	H3K36me3
1	0.816024963	0.072049299	0.038605152	0.020246479	0.016770888
2	0.992609078	0.056259084	0.939698616	0.003763232	0.00425547
3	0.902509445	0.88670947	0.158882985	0.009916558	0.015085619
4	0.941394687	0.673942642	0.559298287	0.003135921	0.900152465
5	0.022041174	0.597236878	0.132133207	0.001149502	0.645764563
6	0.792054808	0.261267951	0.032311347	0.860093383	0.067220417
7	0.004623883	0.005450119	0.008333466	0.03397753	0.26372209
8	0.003027268	0.006516287	0.023162381	0.004902518	0.939459581
9	0.958937942	0.971008904	0.951671278	8.99E-04	0.025467969
10	0.007150536	0.065843825	0.541856976	8.41E-04	0.893499464
11	0.071764218	0.865330674	0.858478997	5.10E-04	0.807737025
12	0.00555704	0.063212405	0.474481827	0.005120965	0.020984179
13	0.046431507	0.901302483	0.837542311	6.24E-04	0.012490048
14	0.016378588	0.669189546	0.065174242	0.005394518	0.006239919
15	0.006275049	0.004557815	0.002835327	0.516957667	0.015594047
16	0.001318563	0.002559291	0.00222194	0.013031156	0.004085968

Appendix iii Table 2 – Chromatin state overlap enrichment categories in hMSCs

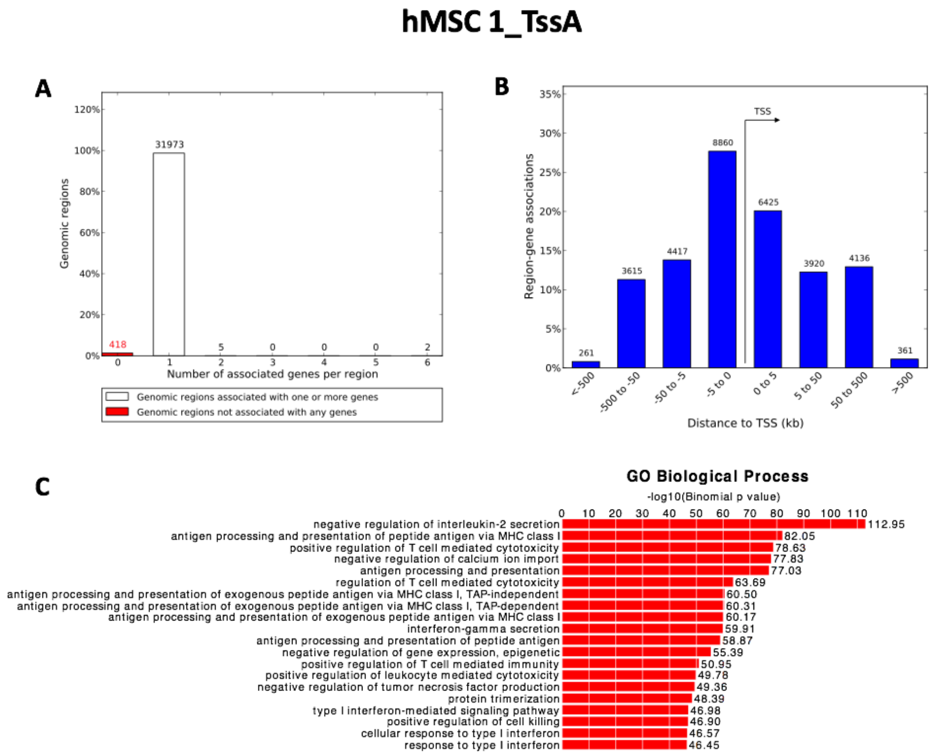


state (User order)	Genome %	CpGIsland.hg38.bed.gz	RefSeqExon.hg38.bed.gz	RefSeqGene.hg38.bed.gz	RefSeqTES.hg38.bed.gz	RefSeqTSS.hg38.bed.gz	RefSeqTSS2kb.hg38.bed.gz
1_TssA	1.22334	8.21166	2.39116	1.0397	2.45958	7.51893	10.16567
2_TssS	0.9061	44.87318	6.91506	1.58722	2.88002	43.22077	21.38764
3_TssFlnk	0.75065	3.46607	1.55022	1.5441	1.55883	2.72224	5.49602
4_TssFlnkU	0.73744	2.94805	5.49693	2.19973	6.9893	1.87954	4.75372
5_TssFlnkD	1.06164	0.31442	2.17224	2.18907	1.97987	0.44895	0.52765
6_TssBiv	0.76195	27.12285	4.1083	1.15955	3.04706	10.5082	10.04097
7_TxWk	8.00771	0.37144	1.1067	1.43496	1.16746	0.31094	0.38726
8_TxS	7.09637	0.6189	3.86034	2.1902	3.06708	0.41302	0.46755
9_TxFlnk	1.05162	2.77303	1.65338	1.69127	1.84566	3.13872	7.22329
10_EnhG1	1.0805	1.24159	4.43474	2.18941	3.82757	0.65659	0.80498
11_EnhG2	0.60128	0.66915	3.00854	2.22155	2.77497	0.66057	0.74271
12_EnhA	1.7266	0.27569	0.77002	1.09158	0.92154	0.58065	0.986
13_EnhS	1.4394	0.07201	0.64736	1.28001	0.8796	0.60135	0.72799
14_EnhP	3.29198	0.06388	0.54869	1.2378	0.68927	0.51922	0.586
15_Repr	9.37243	0.6456	1.04446	0.8375	1.11472	0.84522	0.99753
16_Quies	60.89101	0.0764	0.35642	0.70842	0.47956	0.30217	0.37744
Base	100	0.7357	2.55473	41.88855	0.00101	0.00114	3.64375

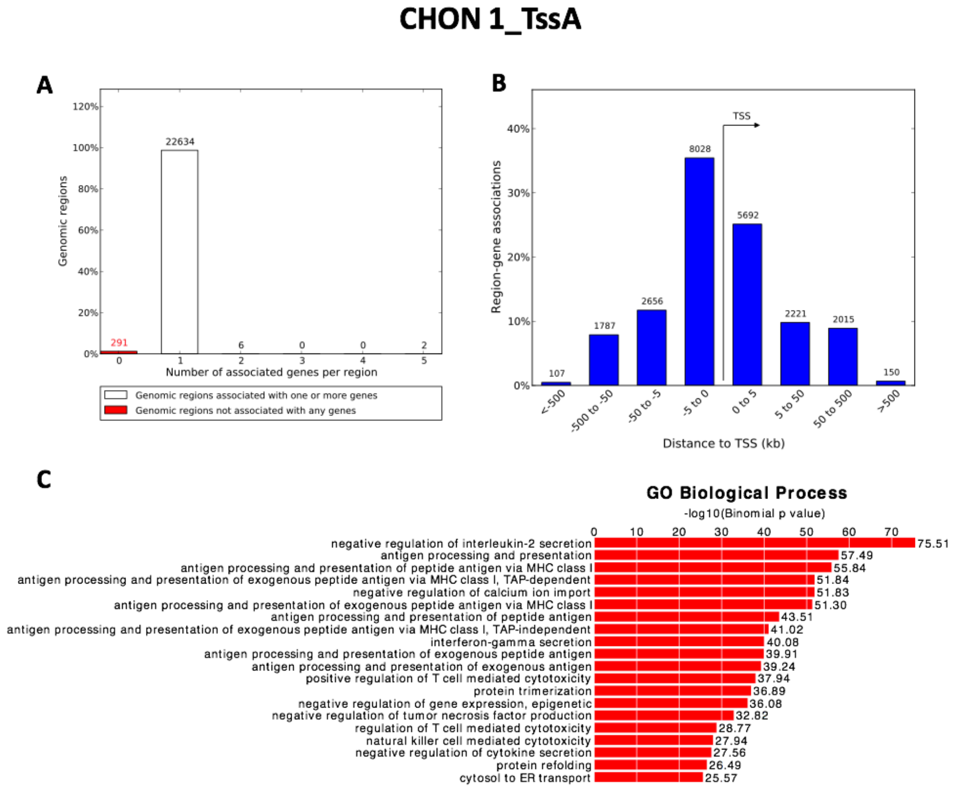
Appendix iii Table 3 – Chromatin state overlap enrichment categories in chondrocytes



state (User order)	Genome %	CpGIsland.hg38.bed.gz	RefSeqExon.hg38.bed.gz	RefSeqGene.hg38.bed.gz	RefSeqTES.hg38.bed.gz	RefSeqTSS.hg38.bed.gz	RefSeqTSS2kb.hg38.bed.gz
1_TssA	0.80121	19.32059	3.76884	1.06847	3.11797	19.419	15.1517
2_TssS	0.46973	64.21637	9.29594	1.64671	3.18963	60.89766	24.1222
3_TssFlnk	0.57101	13.95326	4.05164	1.75813	3.56174	8.28936	13.52507
4_TssFlnkU	0.05016	4.47752	8.62319	2.26833	11.84869	4.75082	6.83771
5_TssFlnkD	0.51263	1.09562	4.65804	2.27889	3.66545	0.66793	0.76629
6_TssBiv	0.32784	42.90233	5.37277	1.18783	2.91771	16.37646	13.12193
7_TxWk	7.51051	0.42356	1.61014	1.28604	1.45537	0.29068	0.33174
8_TxS	2.66442	0.69659	4.45167	2.19023	2.89527	0.36291	0.43034
9_TxFlnk	0.64136	7.32981	3.40218	1.96902	3.71648	5.36854	14.49671
10_EnhG1	0.55367	0.43169	5.24104	2.25207	4.23801	0.376	0.51307
11_EnhG2	0.26039	0.63535	5.25542	2.30466	3.89929	0.58909	0.86027
12_EnhA	3.83239	0.1861	1.12068	1.36745	1.31905	0.47245	0.8556
13_EnhS	3.20733	0.20752	1.30172	1.73	1.44581	0.7601	1.00692
14_EnhP	5.01594	0.33638	1.13252	1.51784	1.26701	0.78147	0.90259
15_Repr	8.47421	0.89992	0.98438	0.83566	1.03854	0.86758	1.09219
16_Quies	65.10719	0.16438	0.50408	0.79264	0.63436	0.34832	0.46129
Base	100	0.7357	2.55473	41.88855	0.00101	0.00114	3.64375

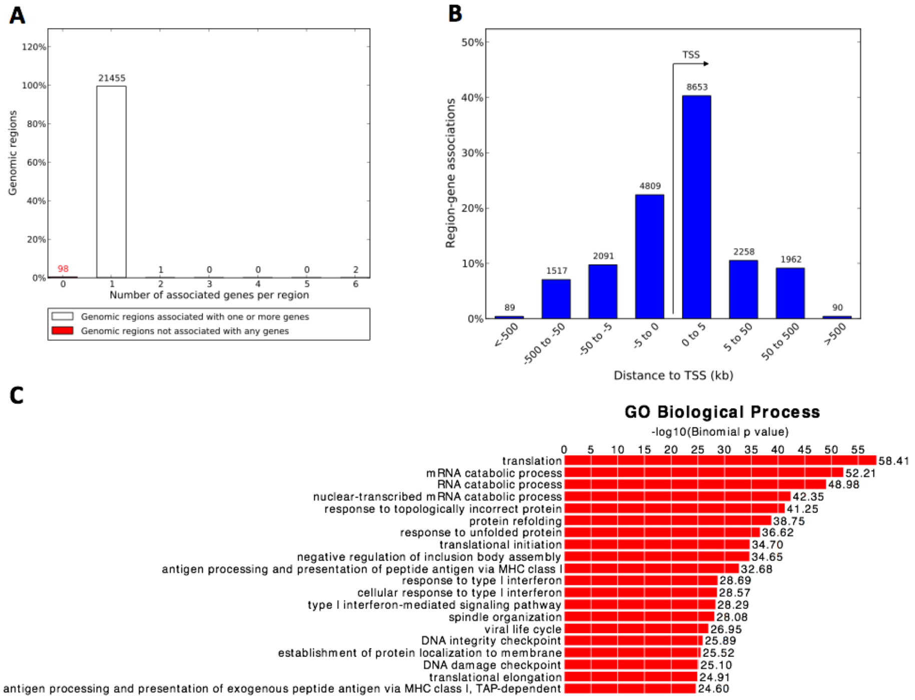


Appendix iii Figure 1 - GREAT gene ontology analysis for hMSC 1_TssA chromatin state. (A) number of associated genes, (B) distance to TSS and (C) biological process GO terms



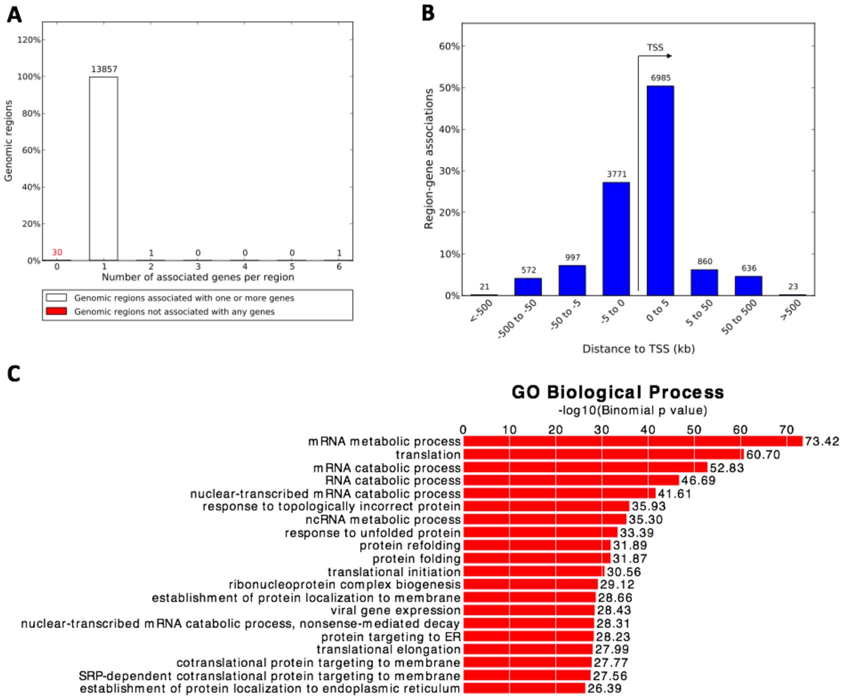
Appendix iii Figure 2 - GREAT gene ontology analysis for CHON 1_TssA chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 2_TssS



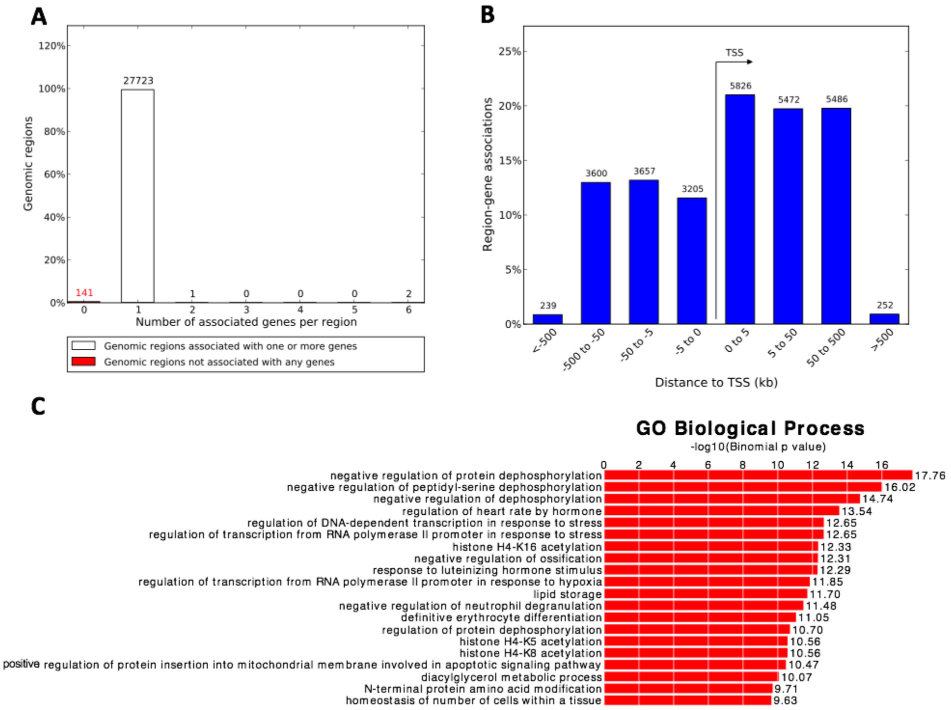
Appendix iii Figure 3 - GREAT gene ontology analysis for hMSC 2_TssS chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 2_TssS



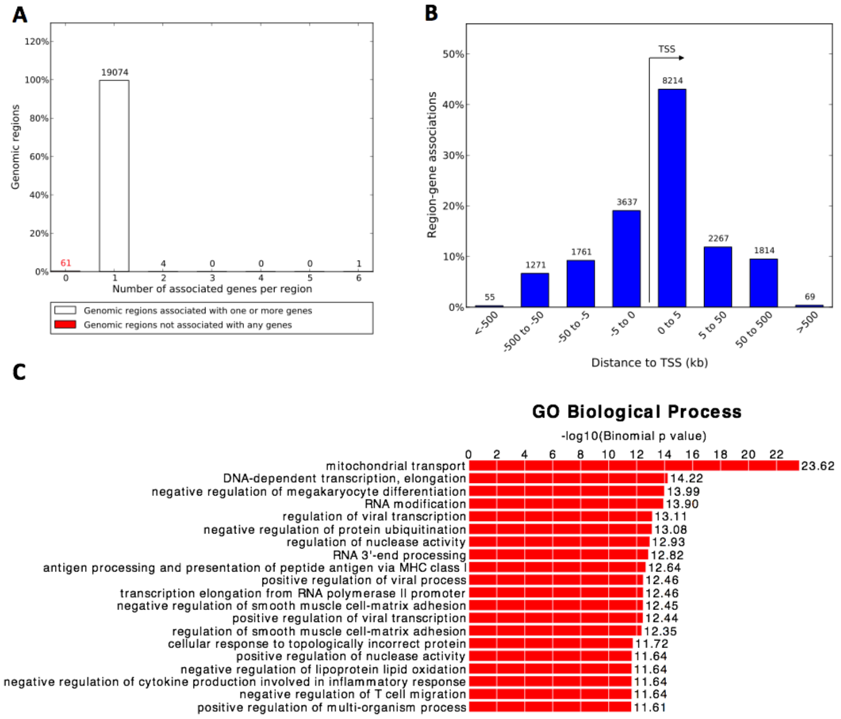
Appendix iii Figure 4 - GREAT gene ontology analysis for CHON 2_TssS chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 3_TssFlnk

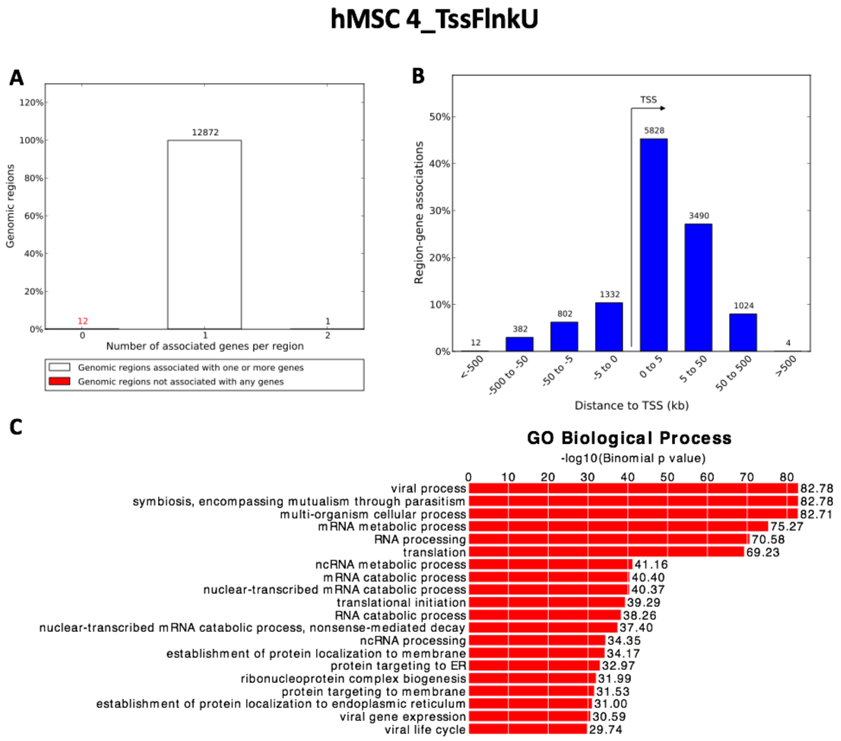


Appendix iii Figure 5 - GREAT gene ontology analysis for hMSC 3_TssFlnk chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

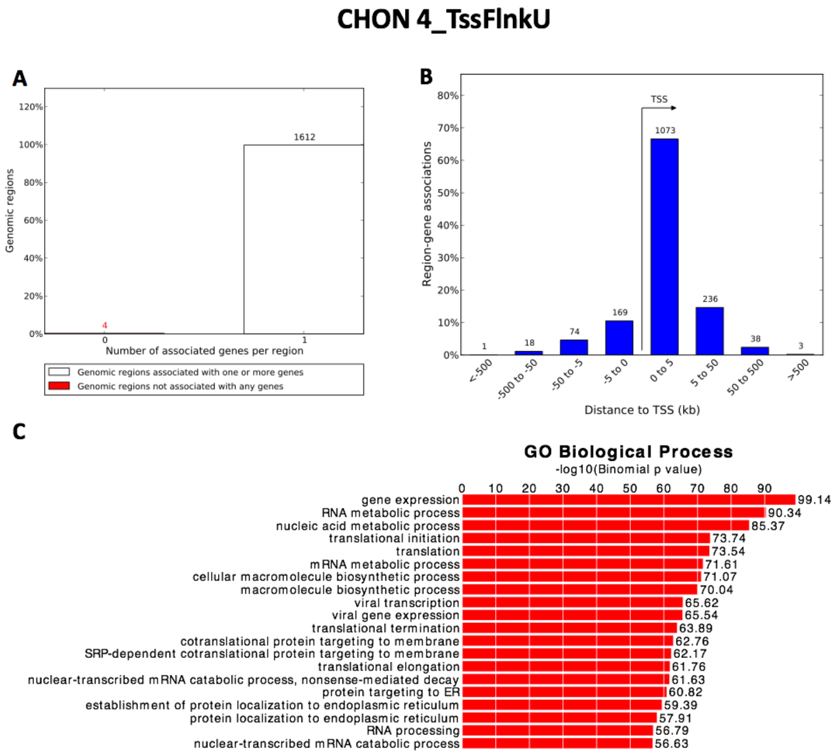
CHON 3_TssFlnk



Appendix iii Figure 6 - GREAT gene ontology analysis for hMSC 3_TssFlnk chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

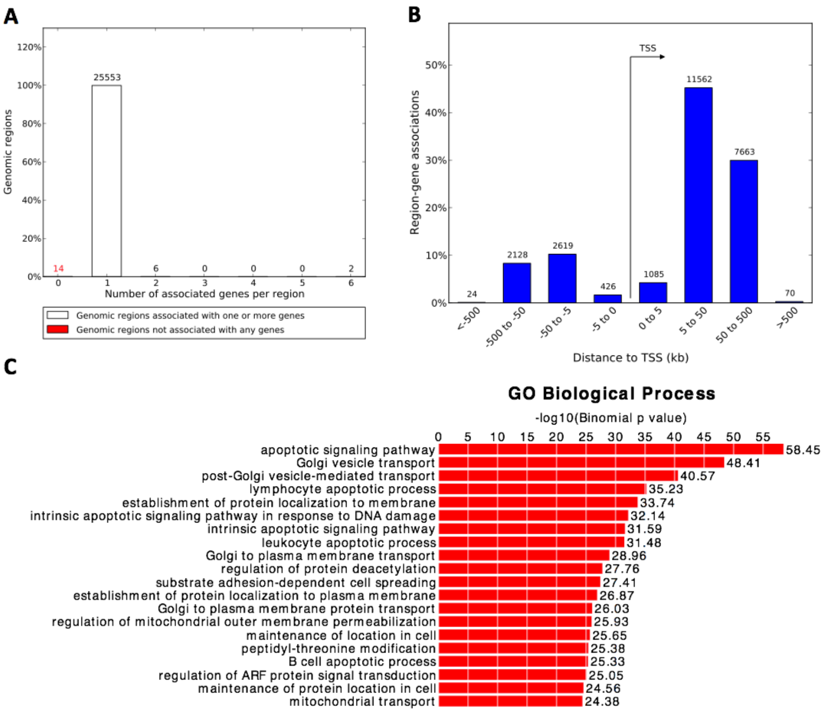


Appendix iii Figure 7 - GREAT gene ontology analysis for hMSC 4_TssFlnkU chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms



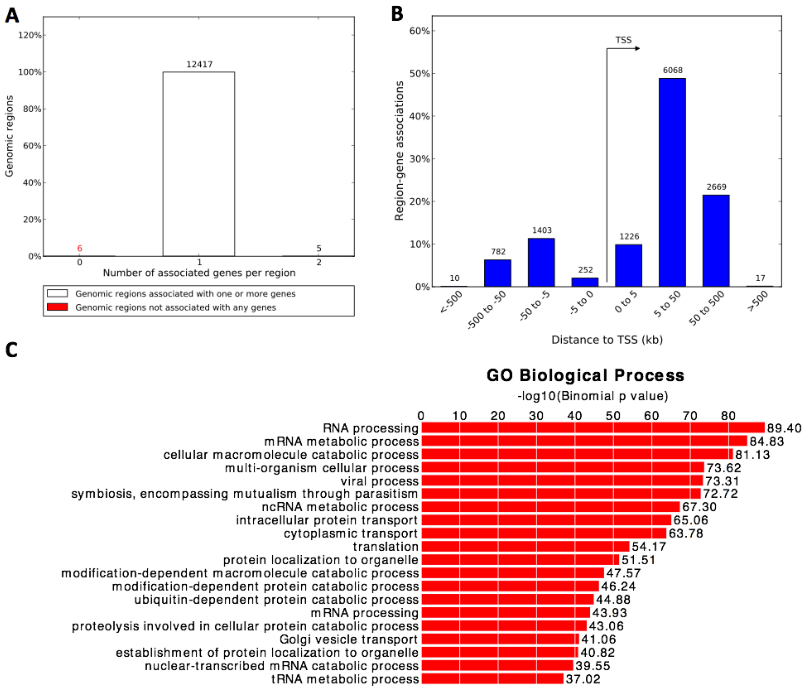
Appendix iii Figure 8 - GREAT gene ontology analysis for CHON 4_TssFlnkU chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 5_TssFlnkD



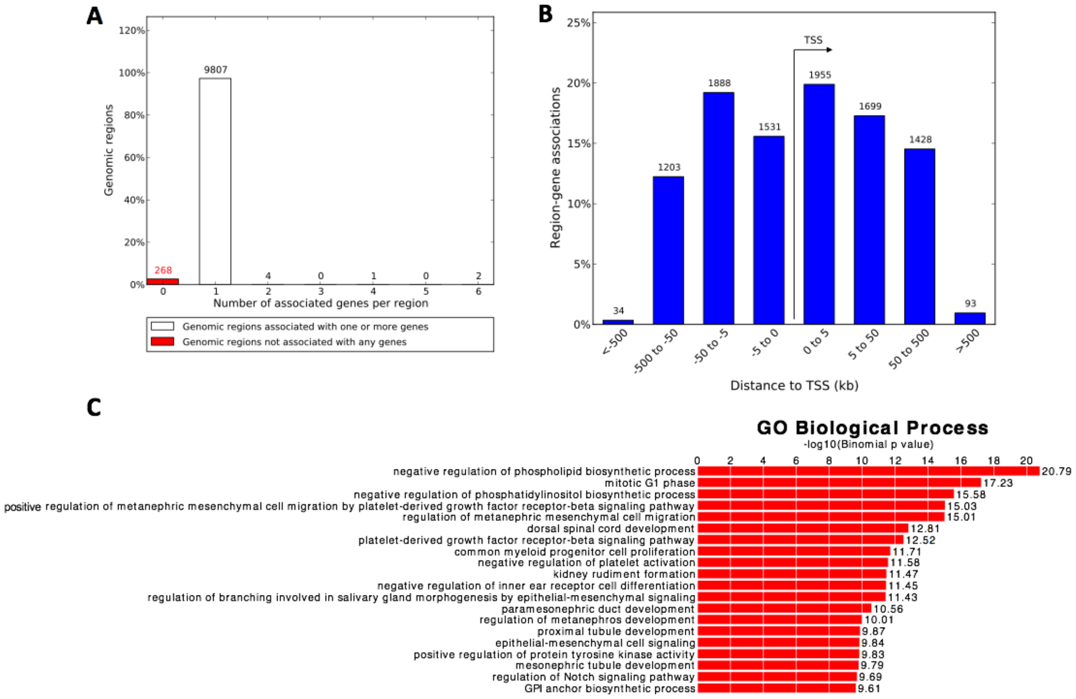
Appendix iii Figure 9 - GREAT gene ontology analysis for hMSC 5_TssFlnkD chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 5_TssFlnkD



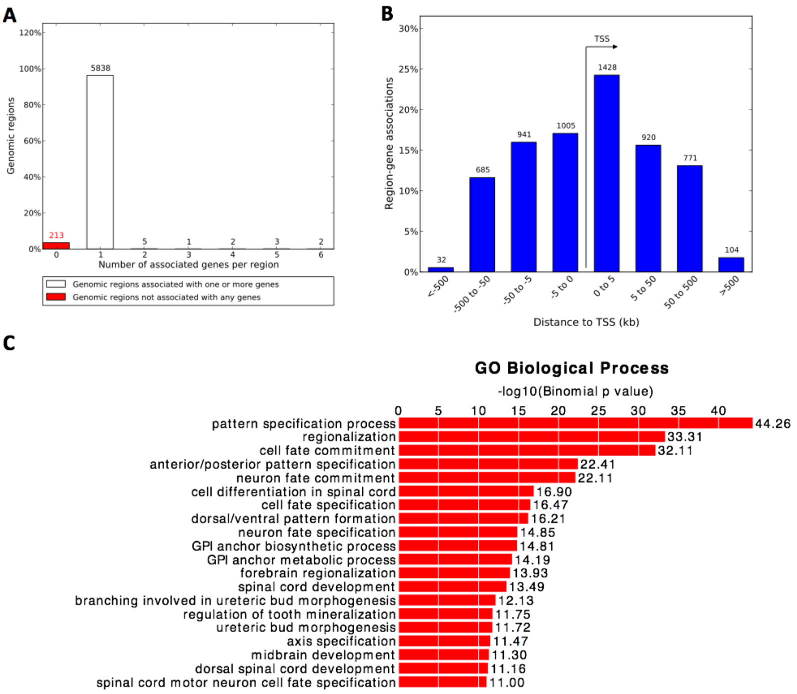
Appendix iii Figure 10 - GREAT gene ontology analysis for CHON 5_TssFlnkD chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 6_TssBiv



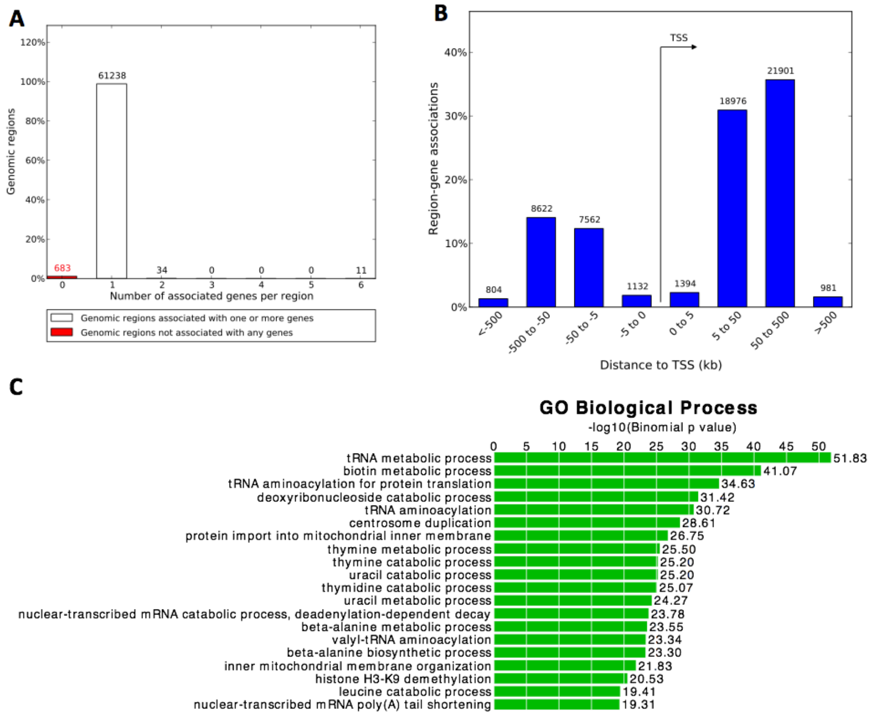
Appendix iii Figure 11 - GREAT gene ontology analysis for hMSC 6_TssBiv chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 6_TssBiv



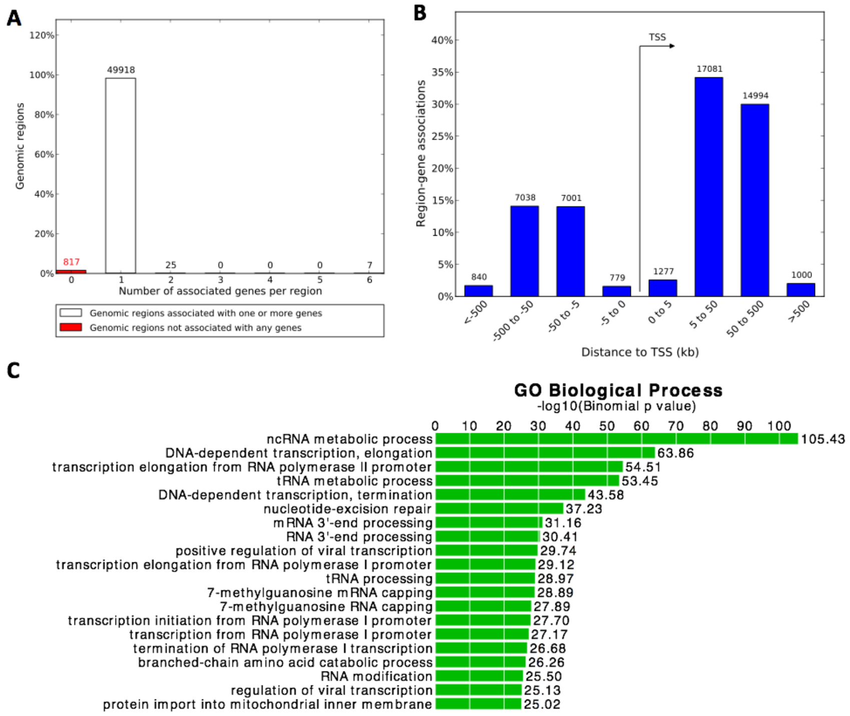
Appendix iii Figure 12 - GREAT gene ontology analysis for CHON 6_TssBiv chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 7_TxWk



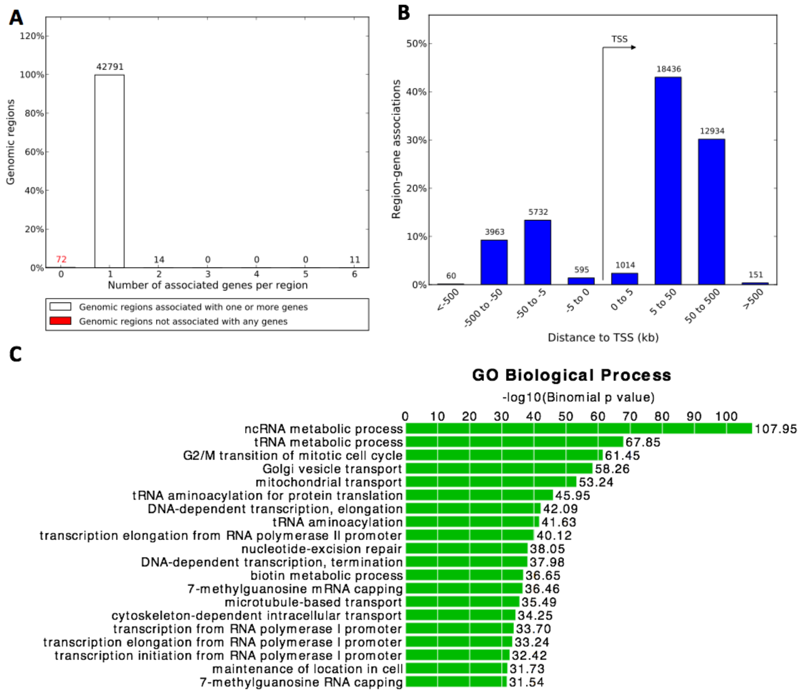
Appendix iii Figure 13 - GREAT gene ontology analysis for hMSC 7_TxWk chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 7_TxWk



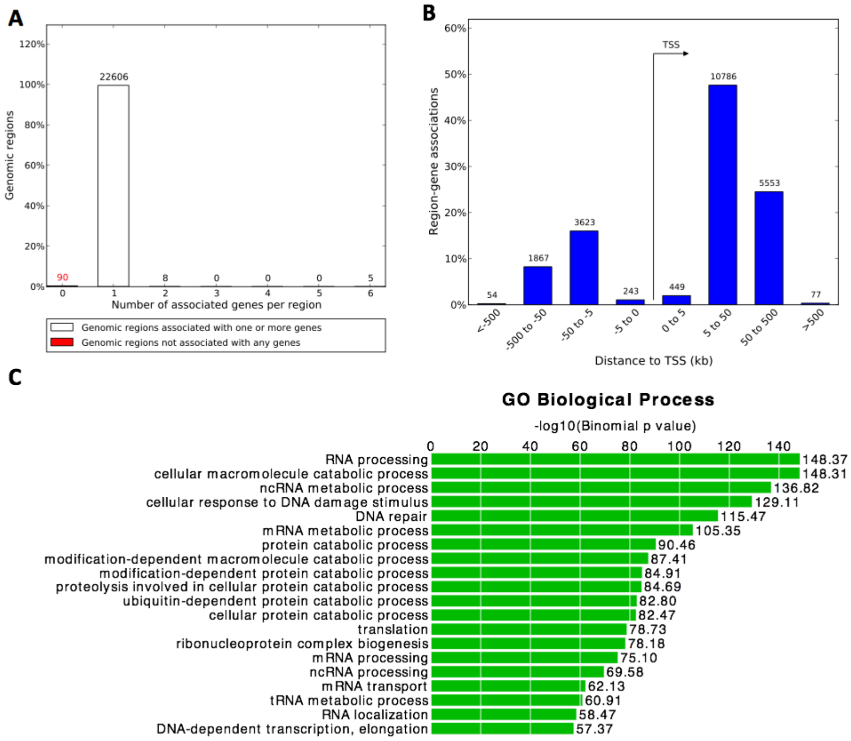
Appendix iii Figure 14 - GREAT gene ontology analysis for CHON 7_TxWk chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 8_TxS



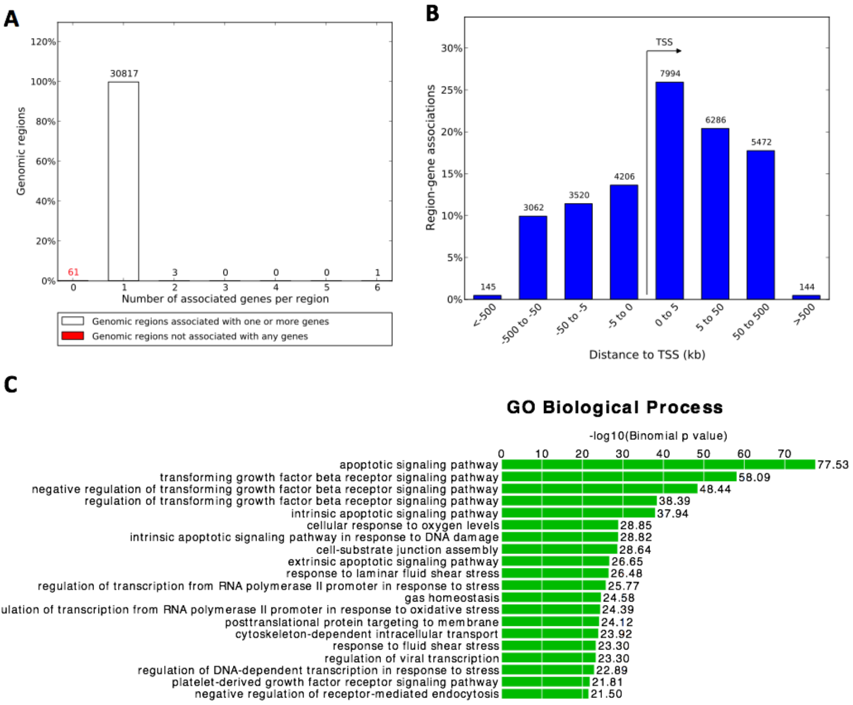
Appendix iii Figure 15 - GREAT gene ontology analysis for hMSC 8_TxS chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 8_TxS



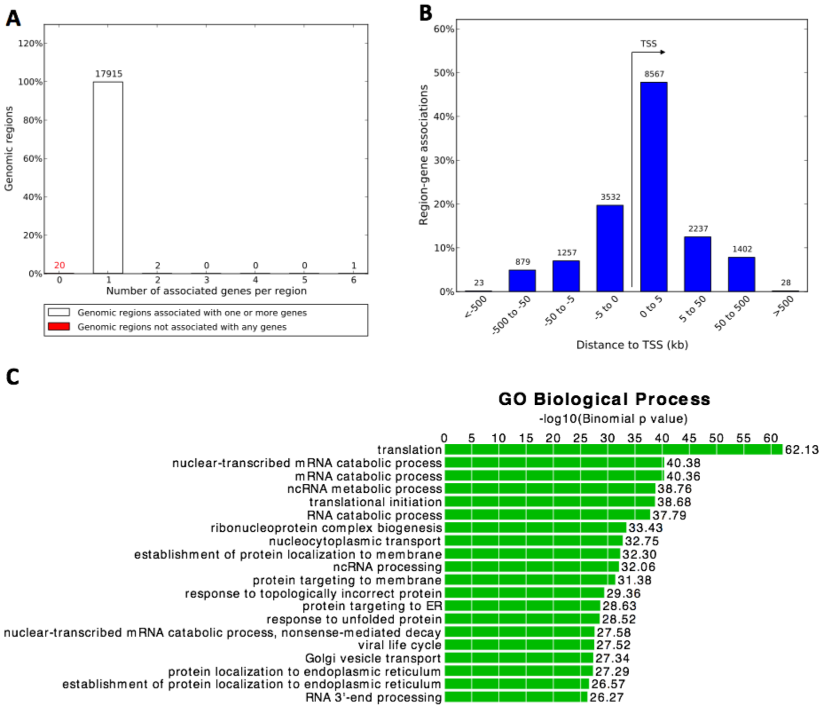
Appendix iii Figure 16 - GREAT gene ontology analysis for CHON 8_TxS chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 9_TxFlnk



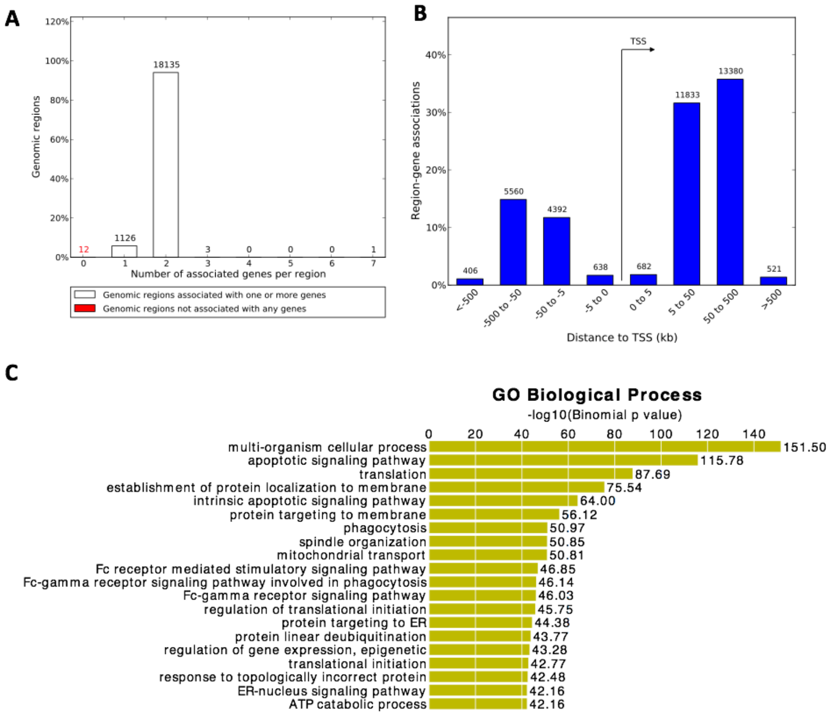
Appendix iii Figure 17 - GREAT gene ontology analysis for hMSC 9_TxFlnk chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 9_TxFlnk



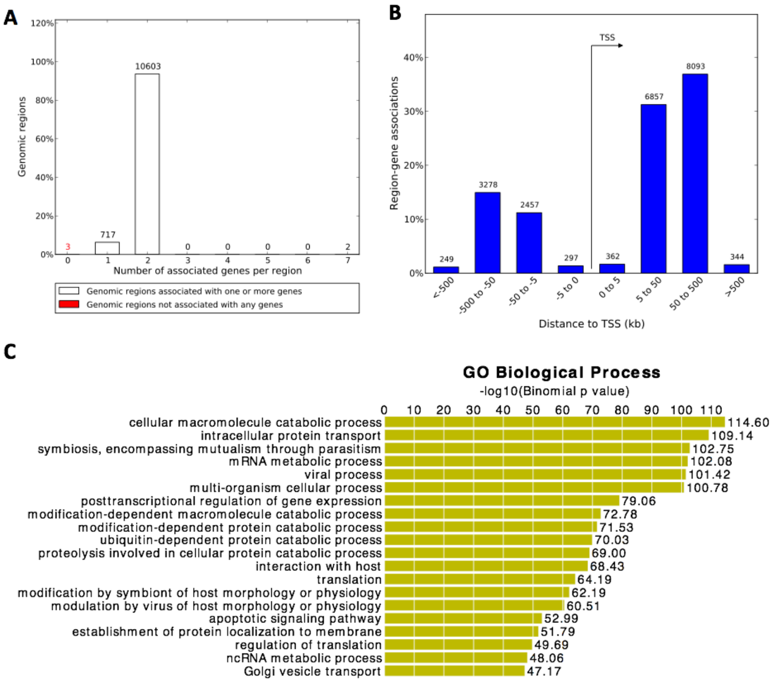
Appendix iii Figure 18 - GREAT gene ontology analysis for CHON 9_TxFlnk chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 10_EnhG1



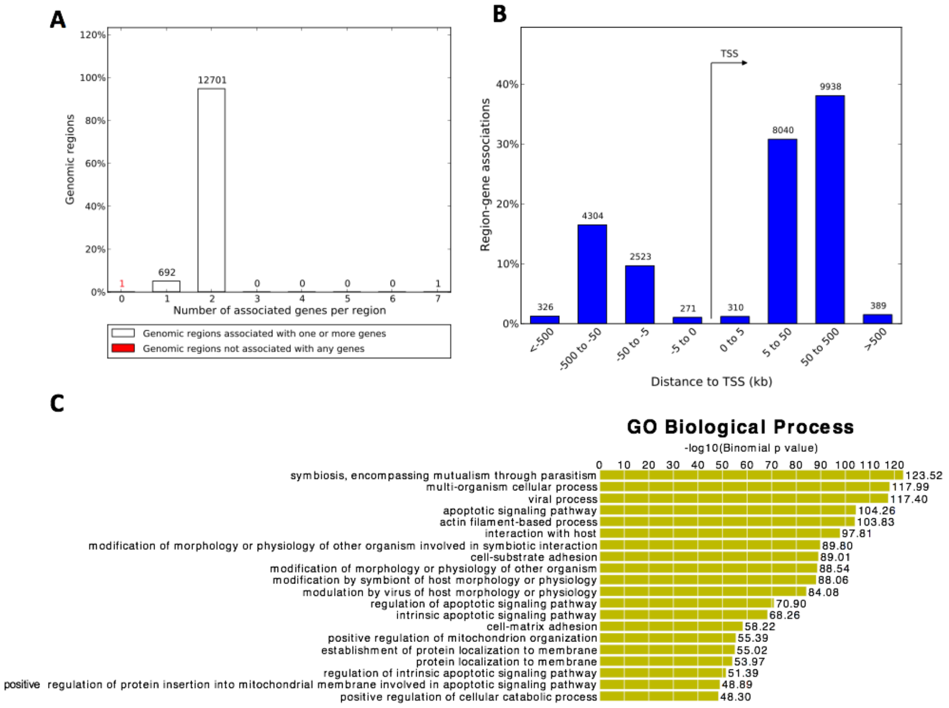
Appendix iii Figure 19 - GREAT gene ontology analysis for hMSC 10_EnhG1 chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 10_EnhG1



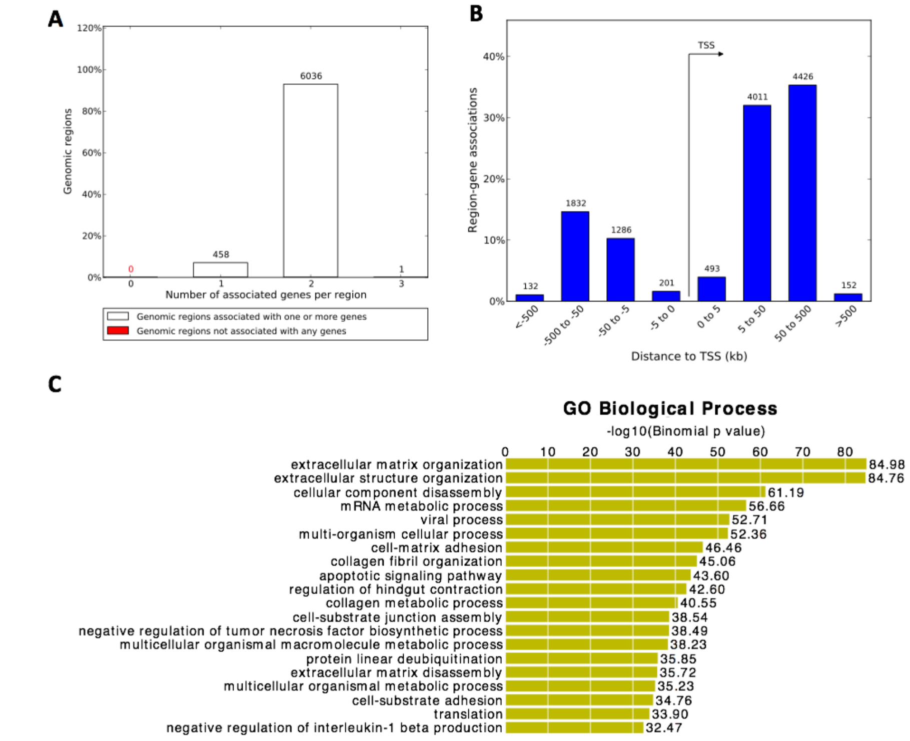
Appendix iii Figure 20 - GREAT gene ontology analysis for CHON 10_EnhG1 chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 11_EnhG2



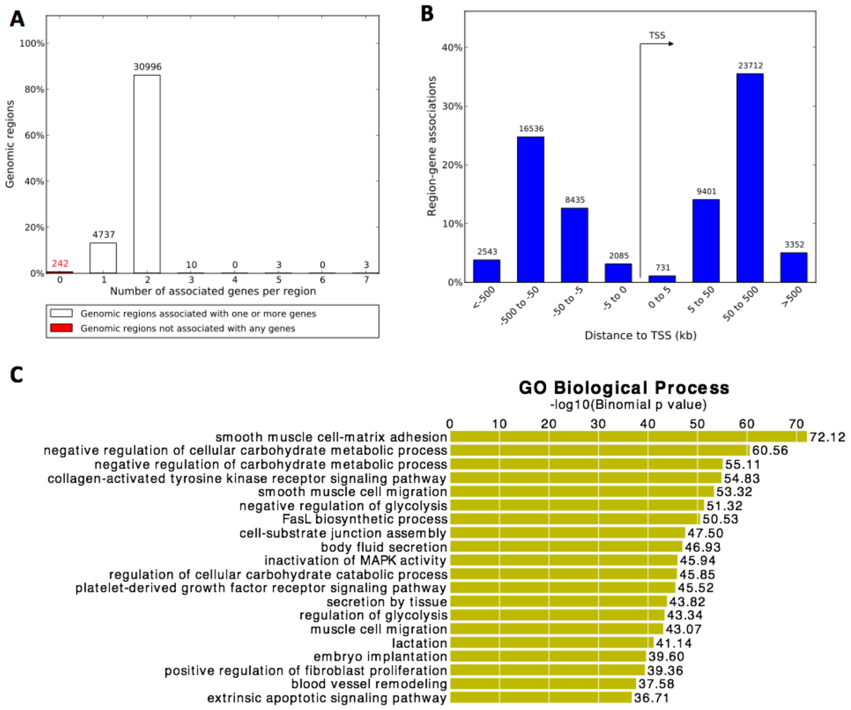
Appendix iii Figure 21 - GREAT gene ontology analysis for hMSC 11_EnhG2 chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 11_EnhG2



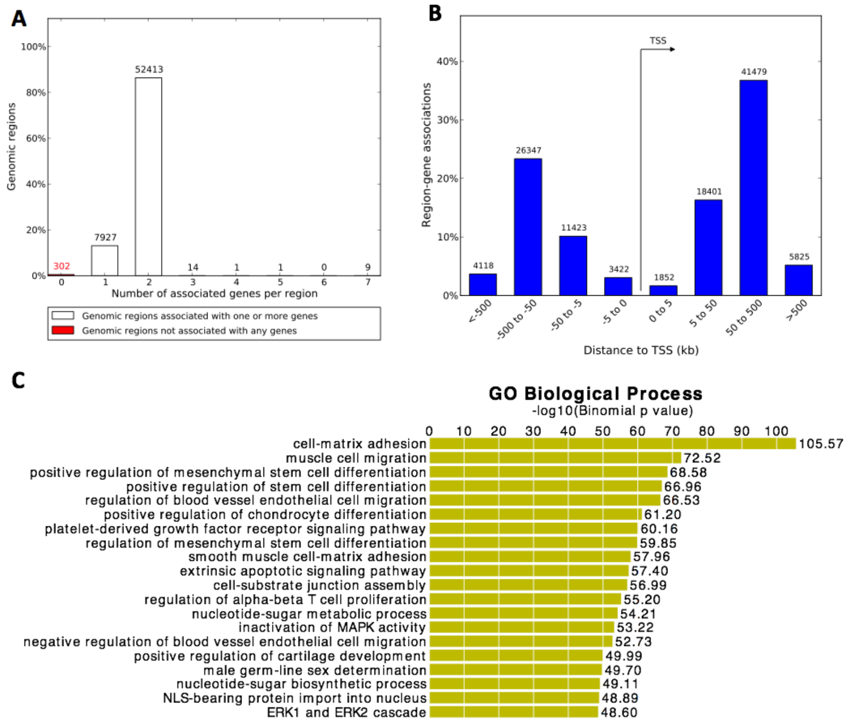
Appendix iii Figure 22 - GREAT gene ontology analysis for hMSC 11_EnhG2 chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 12_EnhA



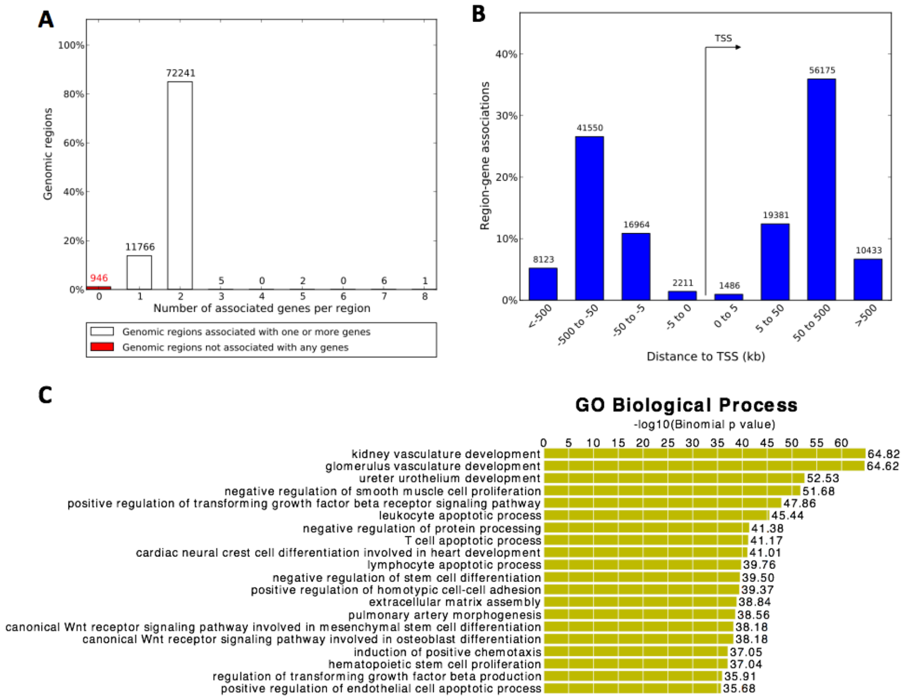
Appendix iii Figure 23 - GREAT gene ontology analysis for hMSC 12_EnhA chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 12_EnhA



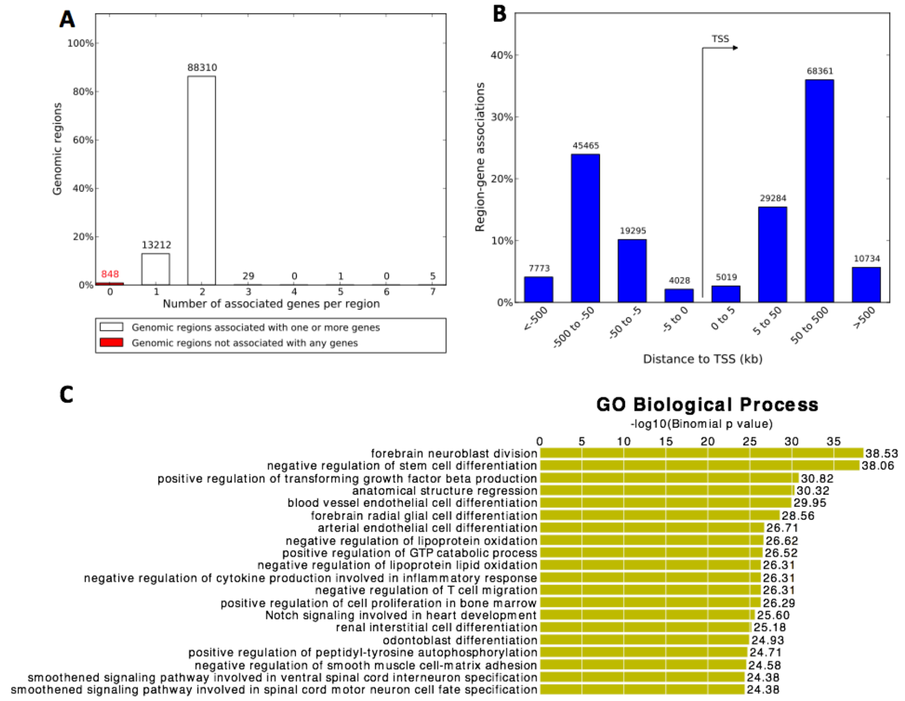
Appendix iii Figure 24 - GREAT gene ontology analysis for CHON 12_EnhA chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 14_EnhP



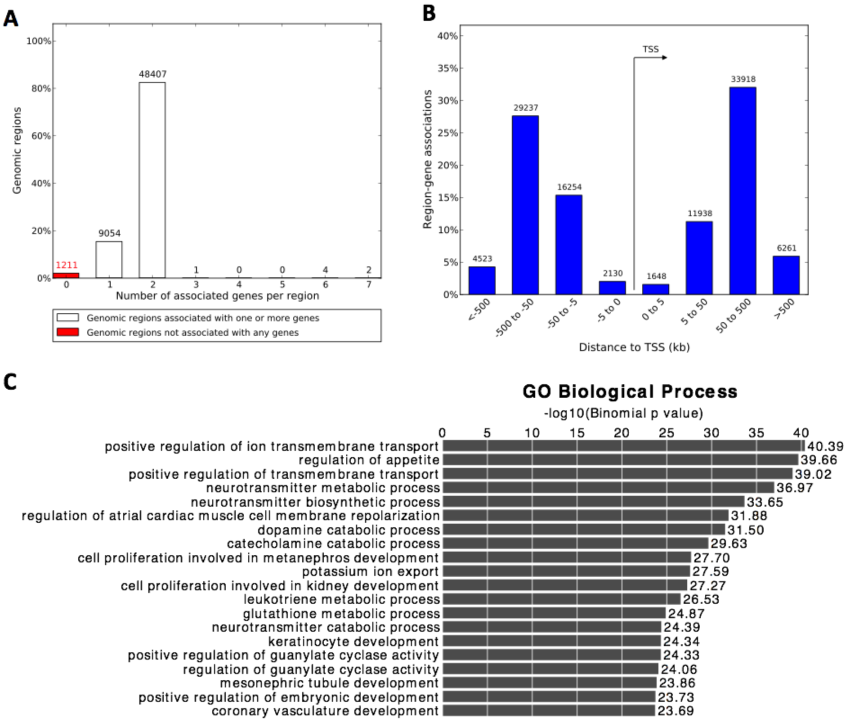
Appendix iii Figure 25 - GREAT gene ontology analysis for hMSC 14_EnhP chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 14_EnhP



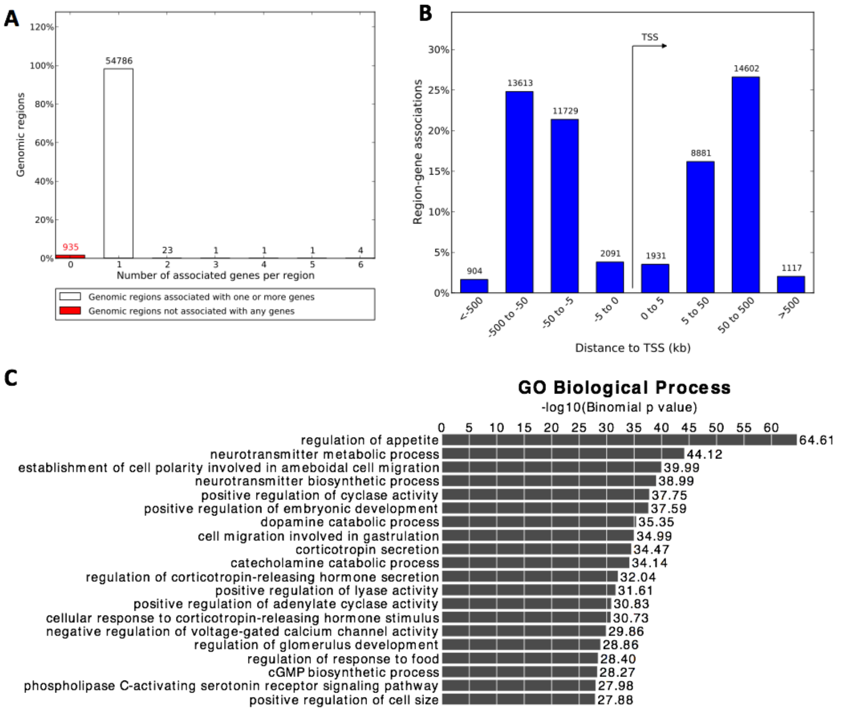
Appendix iii Figure 26 - GREAT gene ontology analysis for hMSC 14_EnhP chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 15_Repr



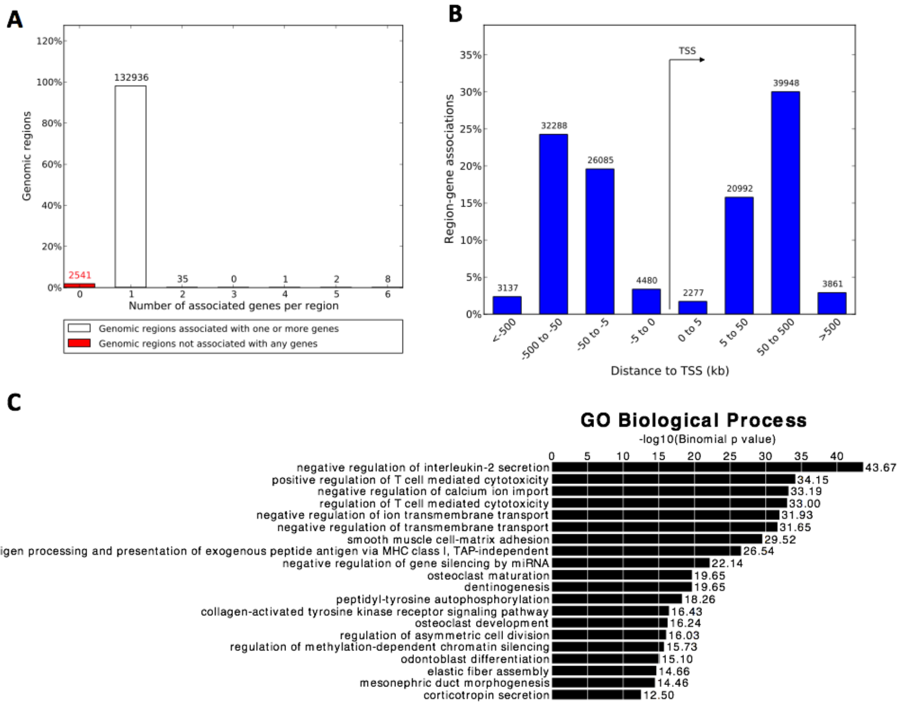
Appendix iii Figure 27 - GREAT gene ontology analysis for hMSC 15_Repr chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

CHON 15_Repr



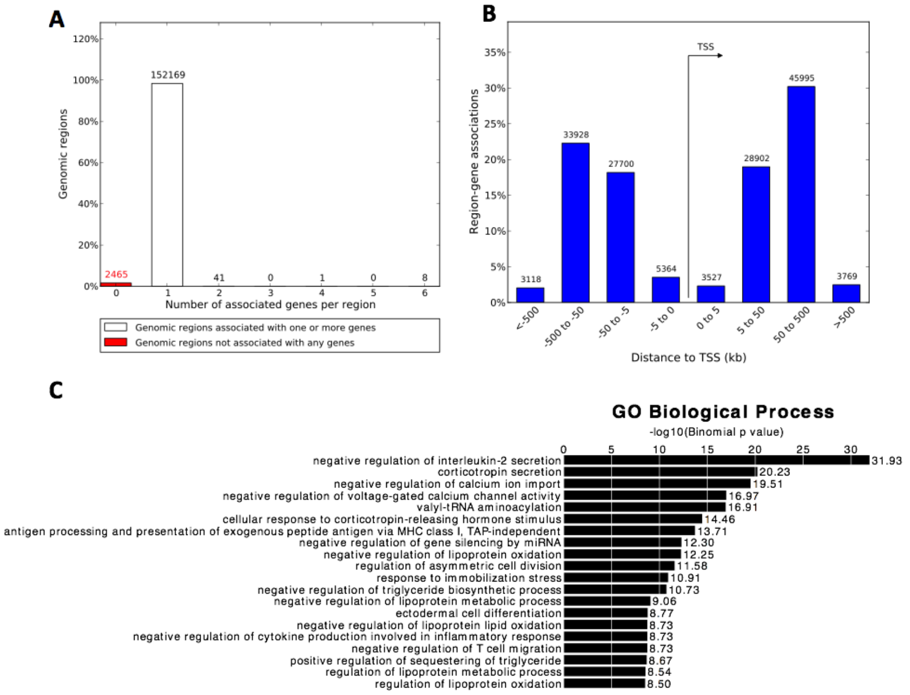
Appendix iii Figure 28 - GREAT gene ontology analysis for CHON 15_Repr chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

hMSC 16_Quies

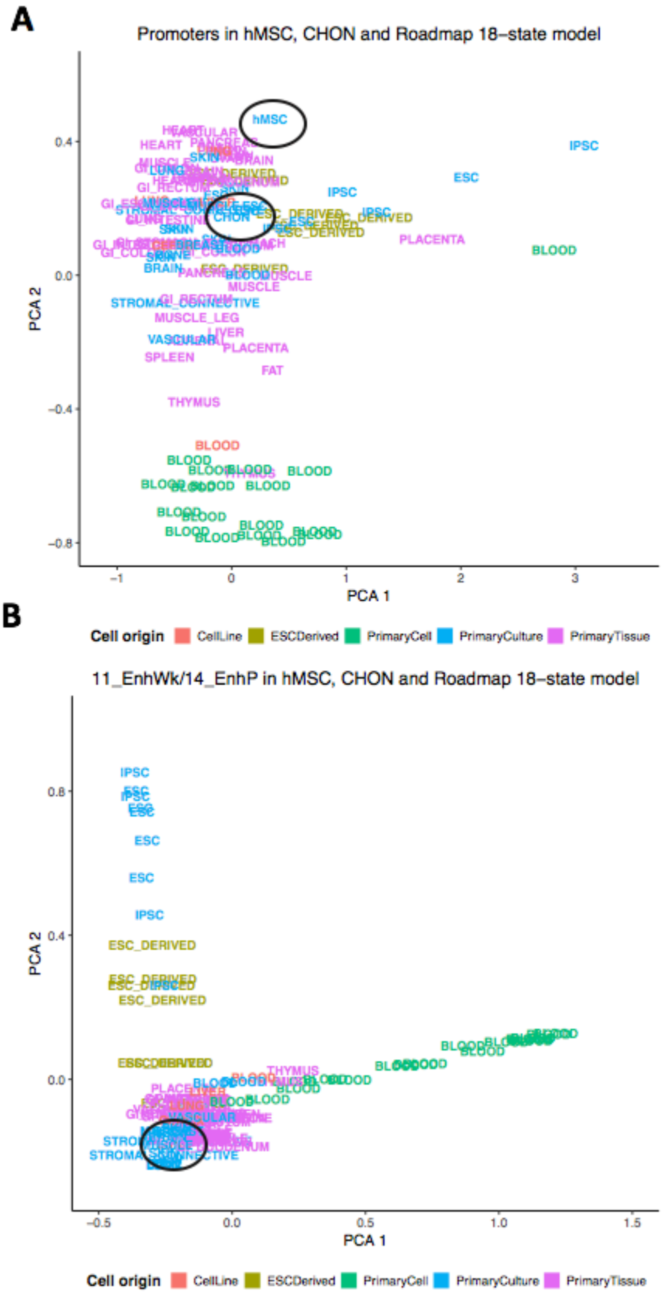


Appendix iii Figure 29 - GREAT gene ontology analysis for hMSC 16_Quies chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms

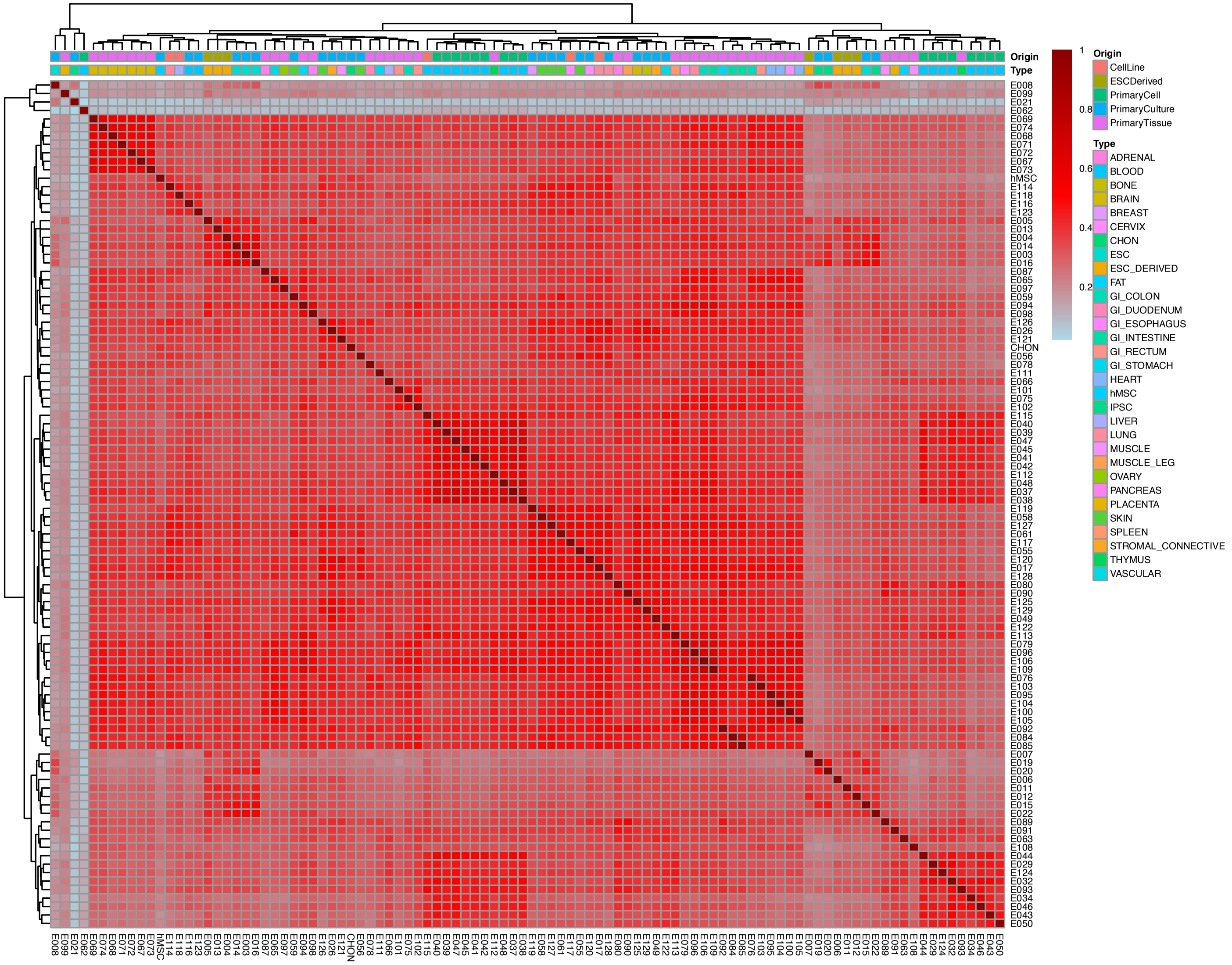
CHON 16_Quies



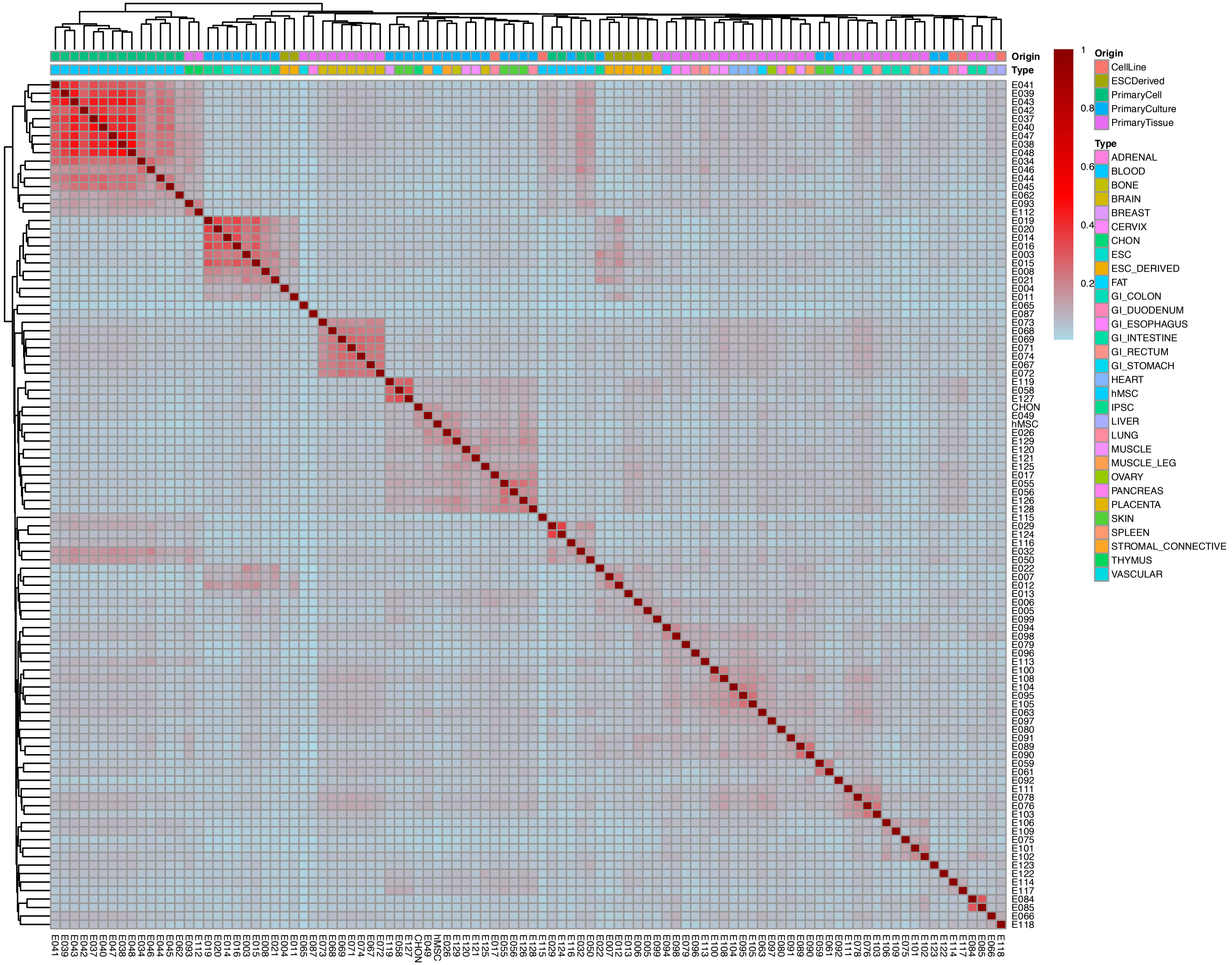
Appendix iii Figure 30 - GREAT gene ontology analysis for CHON 16_Quies chromatin state. (A) number of associated genes, (B) distance to TSS, (C) biological process GO terms



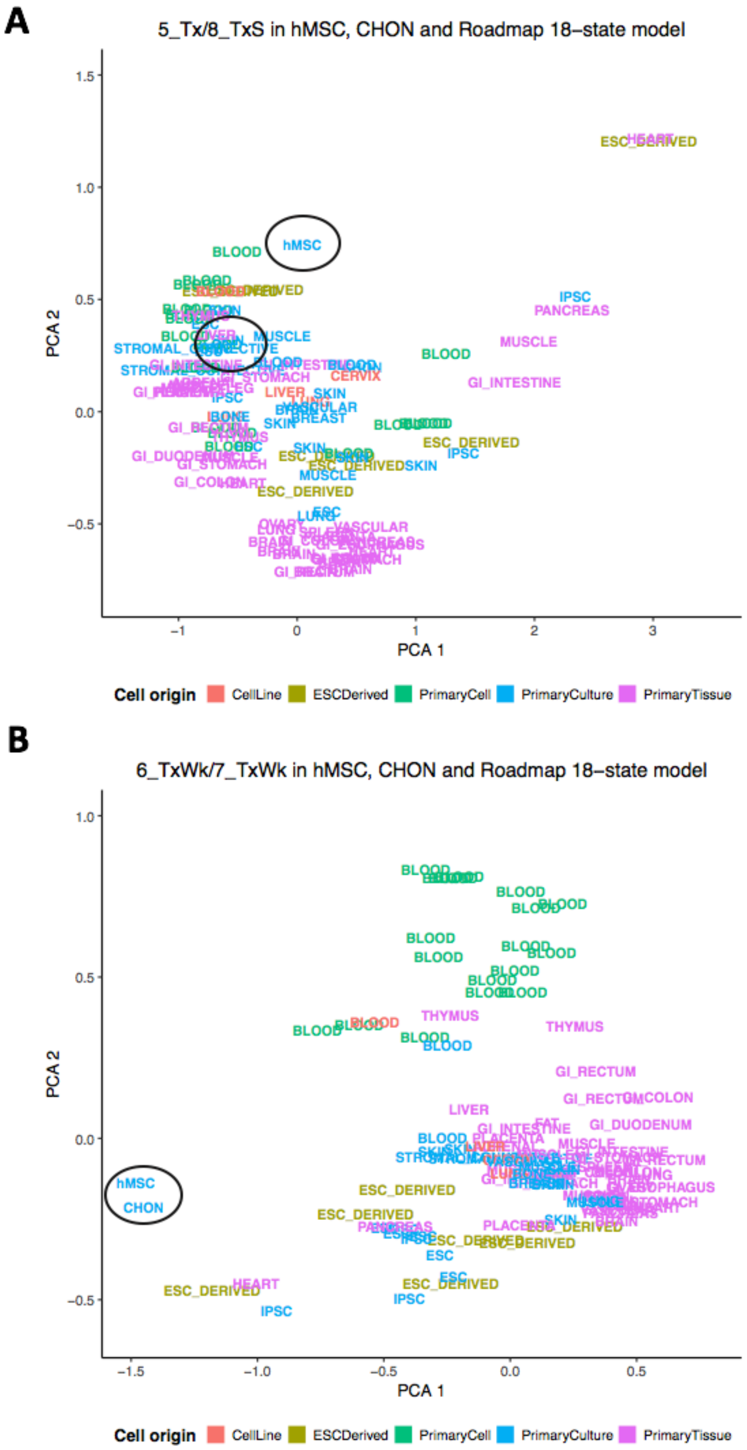
Appendix iii Figure 31 – Principal component analysis of equivalent (A) active promoter states (1_TssA vs 2_TssS) and (B) weak enhancer/poised enhancer (11_EnhWk vs 14_EnhP) states between Roadmap 18 state model and chondrogenesis 16 state model. hMSC and differentiated chondrocytes are circled on the plot. PCA plot was generated using ggplot2 in RStudio.



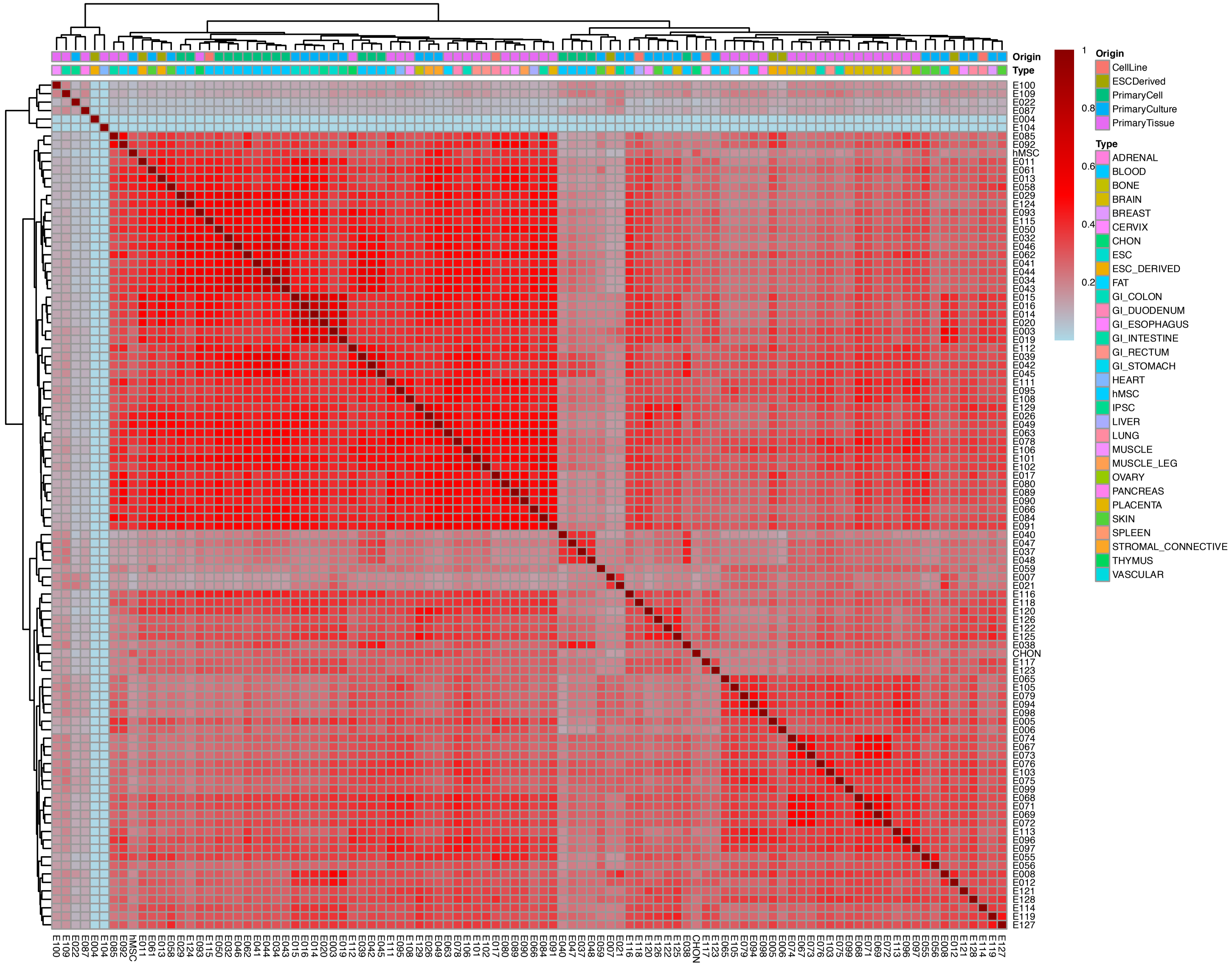
Appendix iii Figure 32 - Heatmap and dendrogram of Jaccard index values for Roadmap 1_TssA compared to chondrogenesis 2_TssS across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.



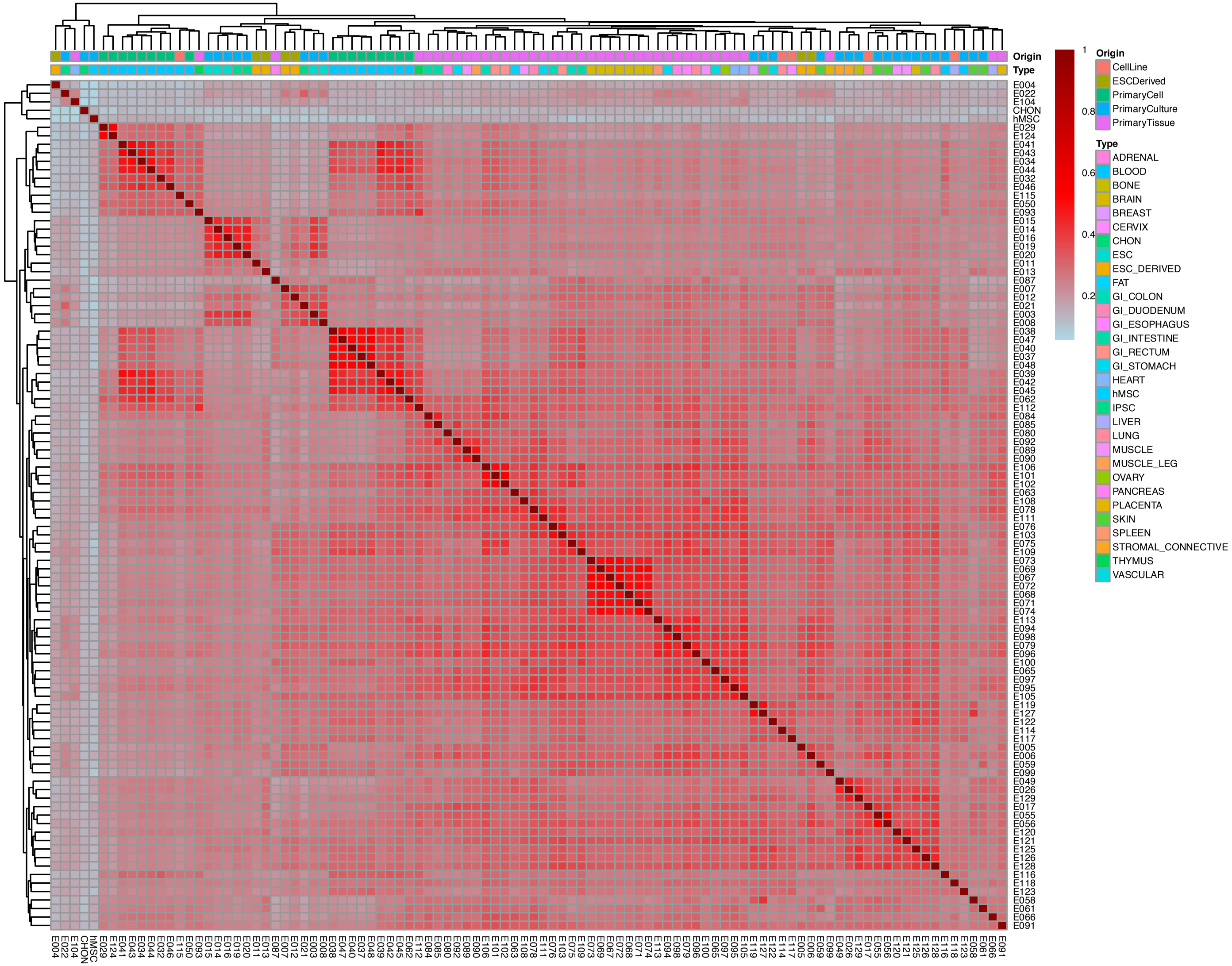
Appendix iii Figure 33 - *Heatmap* and dendrogram of Jaccard index values for Roadmap 11_EnhWk compared to chondrogenesis 14_EnhP across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.



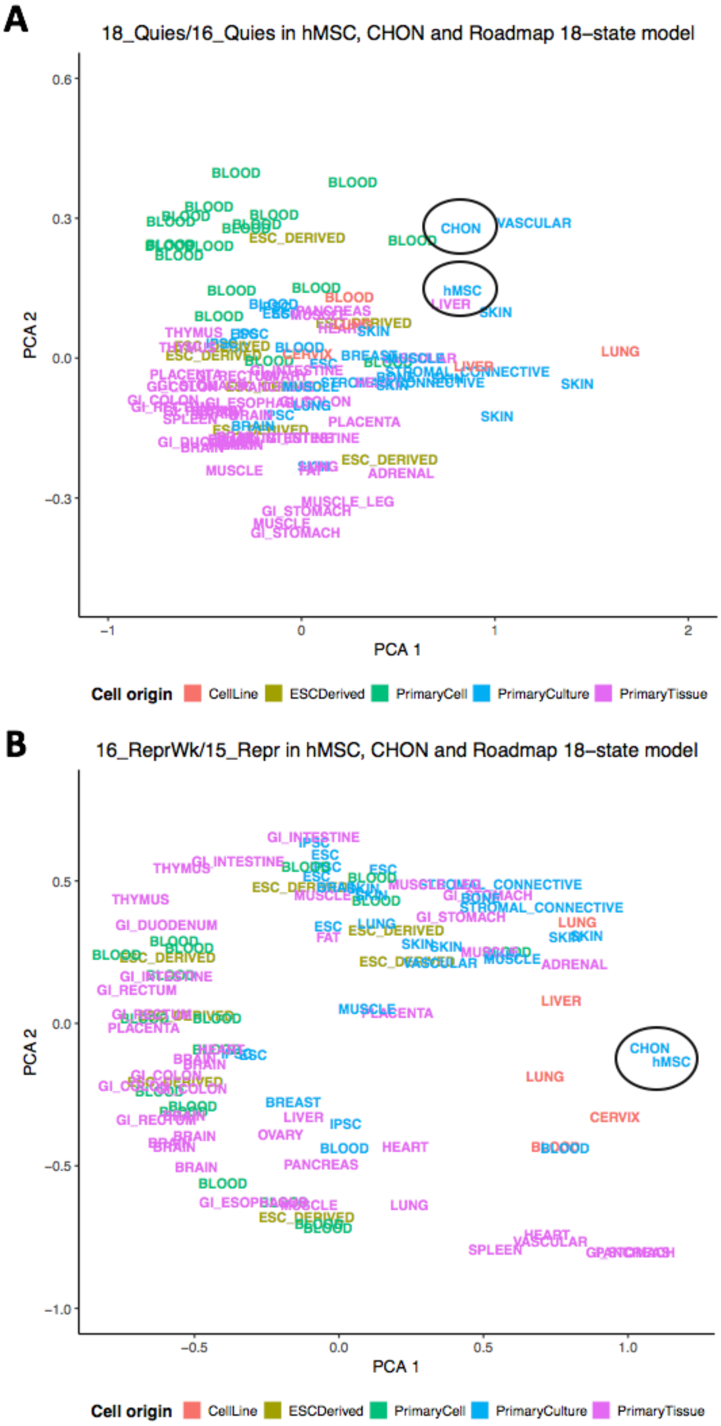
Appendix iii Figure 34 - Principal component analysis of equivalent strong transcription states (A) and weak transcription states (B) between Roadmap 18 state model and chondrogenesis 16 state model. hMSC and differentiated chondrocytes are circled on the plot. PCA plot was generated using ggplot2 in RStudio.



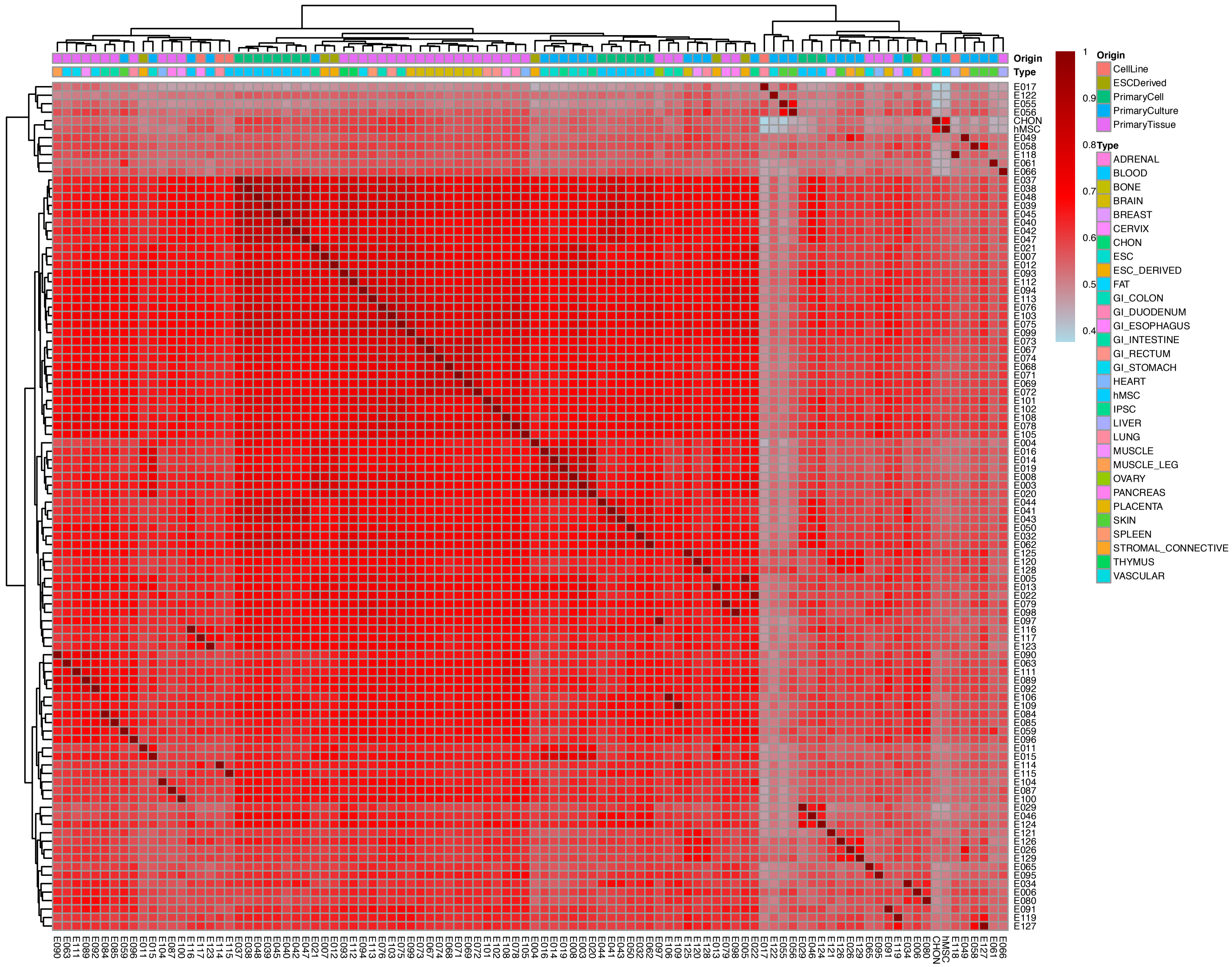
Appendix iii Figure 35 - Heatmap and dendrogram of Jaccard index values for Roadmap 5_Tx compared to chondrogenesis 8_TxS across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.



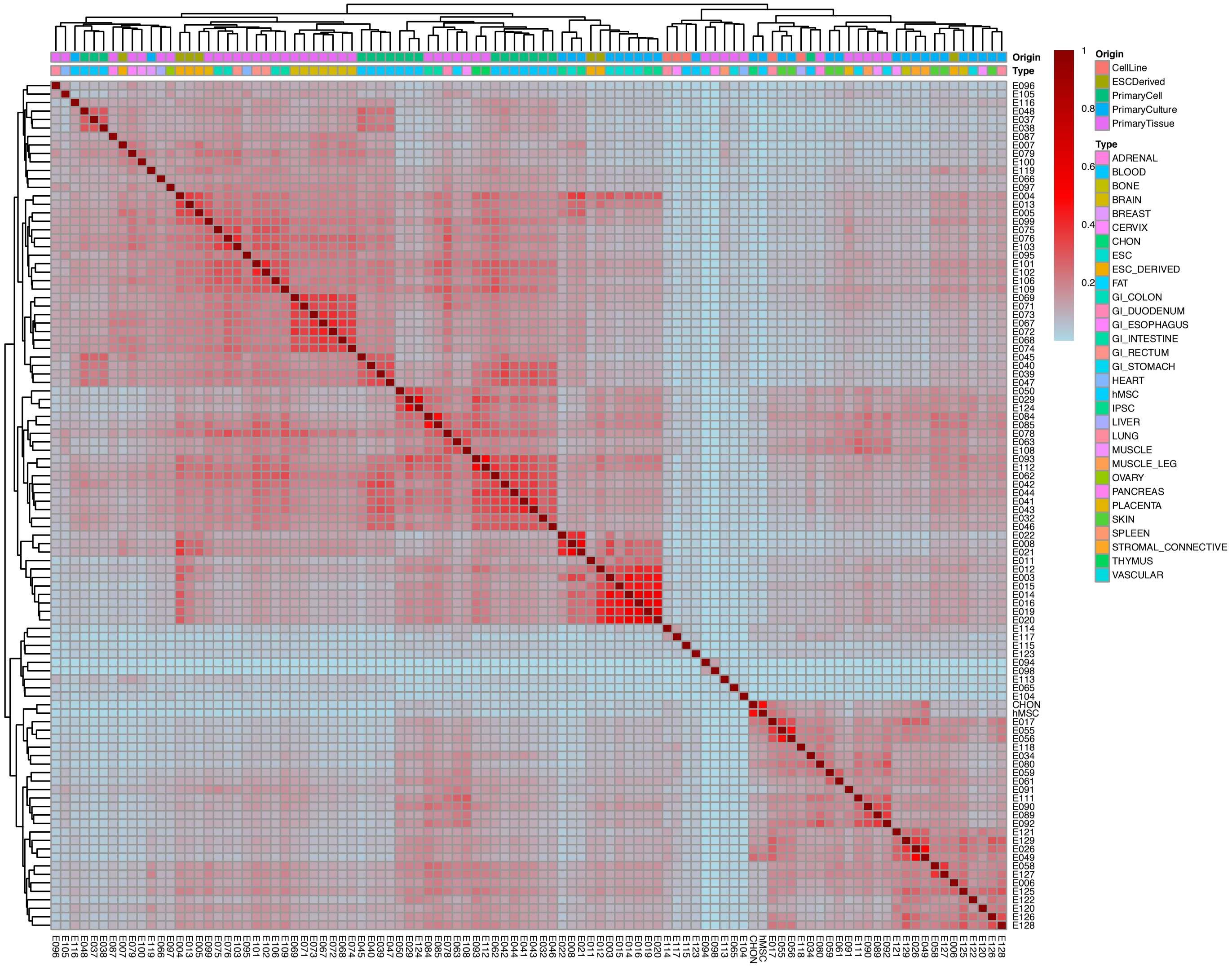
Appendix iii Figure 36 - *Heatmap* and dendrogram of Jaccard index values for Roadmap 6_TxWk compared to chondrogenesis 7_TxWk across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.



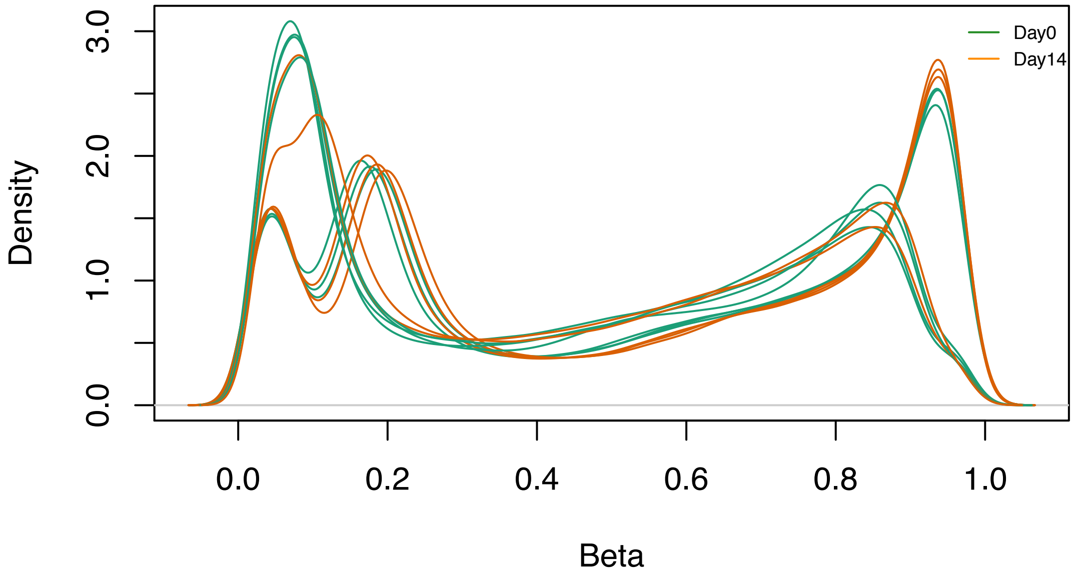
Appendix iii Figure 37 – Principal component analysis of equivalent quiescent states (A) and repressed states (B) between Roadmap 18 state model and chondrogenesis 16 state model. hMSC and differentiated chondrocytes are circled on the plot. PCA plot was generated using ggplot2 in RStudio.



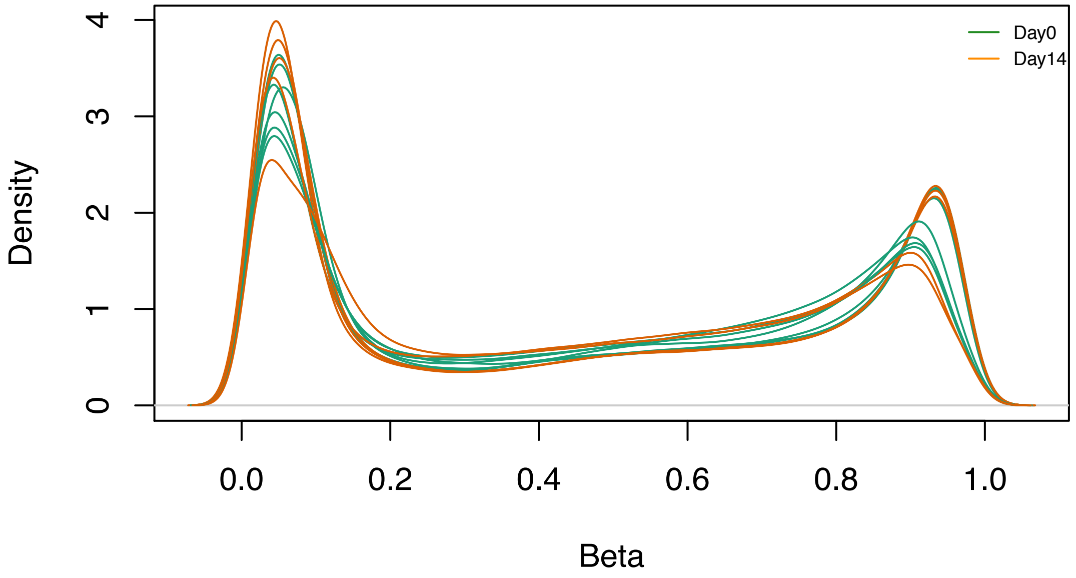
Appendix iii Figure 38 - *Heatmap* and dendrogram of Jaccard index values for Roadmap 18_Quies compared to chondrogenesis 16_Quies across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.



Appendix iii Figure 39 - Heatmap and dendrogram of Jaccard index values for Roadmap 16_ReprPC compared to chondrogenesis 15_Repr across all 98 cell types in the Roadmap extended 18 chromatin state model and hMSCs and differentiated chondrocytes from our in vitro model of chondrogenesis.



Appendix iii Figure 40 – DNA methylation distribution of samples prior to normalisation



Appendix iii Figure 41 – DNA methylation distribution of samples after functional normalisation in minfi

Appendix iii Table 4 - Empirical distribution of DNA methylation changes (delta beta) between hMSC and differentiated chondrocytes in chondrocyte chromatin states

CHON state	Min	1st quarter	Median	Mean	3rd quarter	Max
1_TssA	-0.6188	-0.009713	-0.002017	-0.00581	0.001862	0.2052
2_TssS	-0.6332	-0.005231	-0.00146	-0.00392	0.0009095	0.1869
3_TssFlnk	-0.7435	-0.0147	-0.003249	-0.008072	0.002796	0.3083
4_TssFlnkU	-0.4716	-0.02324	-0.00204	-0.01404	0.01102	0.08802
5_TssFlnkD	-0.5295	-0.01052	0.0007305	-0.00604	0.01325	0.3794
6_TssBiv	-0.2584	-0.01371	-0.003269	-0.006599	0.001716	0.1608
7_TxWk	-0.4573	-0.01347	-0.0005072	-0.003695	0.009153	0.1624
8_TxS	-0.5169	-0.006201	0.001349	0.001007	0.01134	0.1689
9_TxFlnk	-0.841	-0.02058	-0.004874	-0.01731	0.001844	0.2256
10_EnhG1	0.2286	-0.007144	0.001484	0.0008697	0.01286	0.1384
11_EnhG2	-0.672	-0.01518	0.000193	-0.01198	0.01114	0.129
12_EnhA	-0.7849	-0.01518	-0.001214	-0.009232	0.008849	0.2007
13_EnhS	-0.9286	-0.0473	-0.009413	-0.04137	0.005245	0.3959
14_EnhP	-0.8519	-0.0284	-0.004783	-0.02006	0.008739	0.4673
15_Repr	-0.3769	-0.02206	-0.006423	-0.009258	0.005397	0.1459
16_Quies	-0.7053	-0.02331	-0.005419	-0.01031	0.005948	0.2825

Appendix iii Table 5 - Empirical distribution of DNA methylation changes (delta beta) between hMSC and differentiated osteoblasts in Roadmap E129 chromatin states

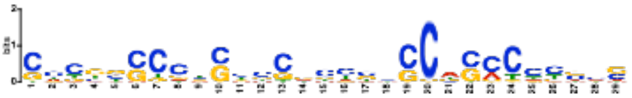
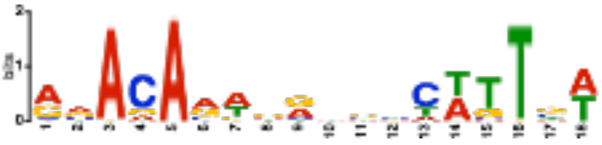
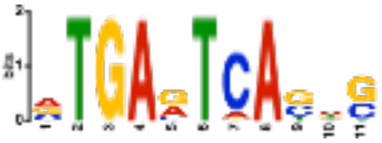
E129 state	Min	1st quarter	Median	Mean	3rd quarter	Max
1_TssA	-0.43340	-0.00217	-0.00023	-0.00053	0.00180	0.18130
2_TssFlnk	-0.23010	-0.00353	-0.00043	-0.00090	0.00260	0.17870
3_TssFlnkU	-0.65220	-0.00696	-0.00045	-0.00349	0.00470	0.18090
4_TssFlnkD	-0.44120	-0.00700	-0.00015	-0.00064	0.00689	0.17880
5_Tx	-0.37300	-0.01193	-0.00191	-0.00367	0.00521	0.67330
6_TxWk	-0.47540	-0.01277	-0.00193	-0.00324	0.00696	0.21010
7_EnhG1	-0.54080	-0.01579	-0.00300	-0.00591	0.00705	0.12750
8_EnhG2	-0.42520	-0.01901	-0.00406	-0.00915	0.00595	0.13160
9_EnhA1	-0.70750	-0.02582	-0.00551	-0.01781	0.00517	0.12870
10_EnhA2	-0.44080	-0.01663	-0.00289	-0.00741	0.00675	0.15830
11_EnhWk	-0.73650	-0.01936	-0.00309	-0.00840	0.00959	0.17260
12_ZNF_Rpts	-0.10350	-0.01132	-0.00041	0.00043	0.01230	0.10070
13_Het	-0.11280	-0.00739	0.00434	0.00548	0.01921	0.14560
14_TssBiv	-0.15800	-0.00320	-0.00006	0.00102	0.00451	0.14630
15_EnhBiv	-0.24280	-0.00557	0.00023	0.00126	0.00788	0.22200
16_ReprPC	-0.24840	-0.00807	0.00012	0.00084	0.00992	0.17620
17_ReprPCWk	-0.30630	-0.01084	0.00079	0.00116	0.01402	0.27990
18_Quies	-0.46420	-0.01191	0.00027	0.00024	0.01337	0.29740

Appendix iii Table 6 - Empirical distribution of DNA methylation changes (delta beta) between hMSC and differentiated adipocytes in Roadmap E063 chromatin states

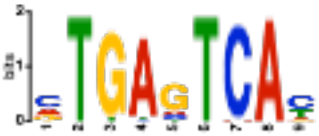
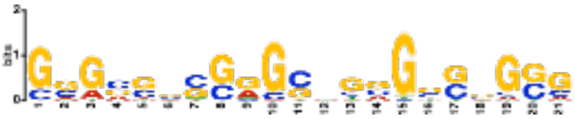
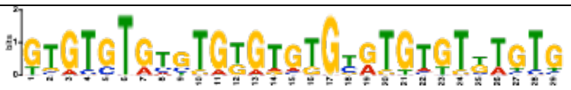
E063 state	Min	1st quarter	Median	Mean	3rd quarter	Max
1_TssA	-0.1597	-0.00109	0.0003925	0.001164	0.00253	0.1719
2_TssFlnk	-0.1155	-0.001295	0.0004909	0.002082	0.003486	0.146
3_TssFlnkU	-0.1739	-0.001702	0.0009707	0.002168	0.00527	0.136
4_TssFlnkD	-0.141	-0.002043	0.001076	0.002604	0.006783	0.1807
5_Tx	-0.1621	-0.007228	0.0001078	0.0002887	0.00677	0.3799
6_TxWk	-0.3052	-0.006806	0.001209	0.002437	0.01082	0.2985
7_EnhG1	-0.162	-0.007198	0.000525	0.001325	0.009332	0.1248
8_EnhG2	-0.08794	-0.006185	0.001206	0.001991	0.0105	0.1474
9_EnhA1	-0.148	-0.005908	0.002052	0.003185	0.01245	0.1539
10_EnhA2	-0.1294	-0.008043	0.001141	0.002266	0.01223	0.137
11_EnhWk	-0.2227	-0.005551	0.002087	0.003602	0.01262	0.1536
12_ZNF_Rpts	-0.1179	-0.004248	0.004799	0.005729	0.01597	0.1203
13_Het	-0.1189	-0.003935	0.006	0.007768	0.01882	0.12
14_TssBiv	-0.1252	-0.001627	0.001065	0.003016	0.006569	0.1152
15_EnhBiv	-0.1066	-0.002173	0.001792	0.003839	0.009527	0.1621
16_ReprPC	-0.219	-0.003583	0.003034	0.004866	0.01348	0.1744
17_ReprPCWk	-0.1533	-0.004556	0.004344	0.005932	0.01638	0.2593
18_Quies	-0.1792	-0.005752	0.003936	0.005736	0.01691	0.1603

Appendix IV

Appendix IV Table 1 – All matches to the top 3 motifs found by MEME in the top 500 significant SOX9 peaks

Database	No.	Logo	No. matches	Matches
MEME	1		16	MA0528.1 (ZNF263), MA0162.2 (EGR1), MA0516.1 (SP2), UP00021_1 (Zfp281_primary), MA0469.1 (E2F3), ZNF740_full, MA0079.3 (SP1), ZNF740_DBD, MA0014.2 (PAX5), Zfp740_DBD, UP00022_1 (Zfp740_primary), MA0470.1 (E2F4), UP00043_2 (Bcl6b_secondary), MA0073.1 (RREB1), UP00002_1 (Sp4_primary), SP3_DBD
MEME	2		7	MA0143.3 (Sox2), MA0442.1 (SOX10), MA0514.1 (Sox3), MA0515.1 (Sox6), MA0078.1 (Sox17), MA0077.1 (SOX9), SOX9_DBD
MEME	3		9	MA0476.1 (FOS), MA0099.2 (JUN:FOS), MA0491.1 (JUND), JDP2_full_1, Jdp2_DBD_1, JDP2_DBD_1, MA0490.1 (JUNB), MA0478.1 (FOSL2), MA0477.1 (FOSL1)

Appendix IV Table 2 – All matches to the top 3 motifs found by MEME in the top 500 significant JUN peaks

Database	No.	Logo	No. matches	Matches
MEME	1		16	MA0478.1 (FOSL2), MA0490.1 (JUNB), MA0477.1 (FOSL1), MA0501.1 (NFE2::MAF), MA0150.2 (Nfe2l2), MA0591.1 (Bach1::Mafk), MA0491.1 (JUND), MA0489.1 (JUN), MA0099.2 (JUN::FOS), MA0476.1 (FOS), JDP2_full_1, Jdp2_DBD_1, UP00103_2 (Jundm2_secondary), JDP2_DBD_1, NFE2_DBD, MA0462.1 (BATF::JUN)
MEME	2		18	MA0146.2 (Zfx), MA0516.1 (SP2), UP00021_1 (Zfp281_primary), MA0162.2 (EGR1), MA0469.1 (E2F3), MA0528.1 (ZNF263), MA0079.3 (SP1), UP00002_1 (Sp4_primary), UP00022_1 (Zfp740_primary), MA0163.1 (PLAG1), MA0024.2 (E2F1), UP00102_1 (Zic1_primary), ZNF740_full, KLF16_DBD, MA0471.1 (E2F6), ZNF740_DBD, MA0470.1 (E2F4), MA0073.1 (RREB1)
MEME	3		2	MA0472.1 (EGR2), MA0073.1 (RREB1)

References

Adam, R.C. et al., 2015. Pioneer factors govern super-enhancer dynamics in stem cell plasticity and lineage choice. *Nature*, 521(7552), pp.366–370.

Agacayak, S. et al., 2012. Effects of mesenchymal stem cells in critical size bone defect. *European review for medical and pharmacological sciences*, 16(5), pp.679–86.

Akiyama, H., 2008. Control of chondrogenesis by the transcription factor Sox9. *Modern Rheumatology*, 18(3), pp.213–219.

Akiyama, H. et al., 2002. The transcription factor Sox9 has essential roles in successive steps of the chondrocyte differentiation pathway and is required for expression of Sox5 and Sox6. *Genes & development*, 16(21), pp.2813–28.

Akiyama, H. et al., 2004. Interactions between Sox9 and β-catenin control chondrocyte differentiation. *Genes & development*, 18, pp.1072–1087.

Alford, A.I., Kozloff, K.M. & Hankenson, K.D., 2015. Extracellular matrix networks in bone remodeling. *The International Journal of Biochemistry & Cell Biology*, 65, pp.20–31.

Allen, M. et al., 2016. Origin of the U87MG glioma cell line: Good news and bad news. *Science Translational Medicine*, 8(354).

Altun, G., Loring, J.F. & Laurent, L.C., 2010. DNA methylation in embryonic stem cells. *Journal of Cellular Biochemistry*, 109(1), pp.1–6.

Alzahrani, F. et al., 2015. LOXL3, encoding lysyl oxidase-like 3, is mutated in a family with autosomal recessive Stickler syndrome. *Human Genetics*, 134(4), pp.451–453.

Aran, D., Sabato, S. & Hellman, A., 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. *Genome Biology*, 14(3), p.R21.

Archin, N.M. et al., 2009. Expression of latent HIV induced by the potent HDAC inhibitor suberoylanilide hydroxamic acid. *AIDS research and human retroviruses*, 25(2), pp.207–12.

Aryee, M.J. et al., 2014. Minfi: a flexible and comprehensive Bioconductor package for the analysis of Infinium DNA methylation microarrays. *Bioinformatics*, 30(10), pp.1363–1369.

Arzenani, M.K. et al., 2011. Genomic DNA hypomethylation by histone deacetylase inhibition implicates DNMT1 nuclear dynamics. *Molecular and cellular biology*, 31(19), pp.4119–28.

Ayyappan, V. et al., 2015. Genome-Wide Profiling of Histone Modifications (H3K9me2 and H4K12ac) and Gene Expression in Rust (Uromyces appendiculatus) Inoculated Common Bean (Phaseolus vulgaris L.) R. Mantovani, ed. *PLOS ONE*, 10(7), p.e0132176.

Bahar Halpern, K., Vana, T. & Walker, M.D., 2014. Paradoxical Role of DNA Methylation in Activation of FoxA2 Gene Expression during Endoderm Development. *Journal of Biological Chemistry*, 289(34), pp.23882–23892.

Bannister, A.J. & Kouzarides, T., 2011. Regulation of chromatin by histone modifications. *Cell research*, 21(3), pp.381–95.

Barter, M.J. et al., 2014. Long non-coding rnas in osteoarthritis and chondrogenesis. *Osteoarthritis and Cartilage*, 22, p.S139.

Barter, M.J. et al., 2015. Genome-Wide MicroRNA and Gene Analysis of Mesenchymal Stem Cell Chondrogenesis Identifies an Essential Role and Multiple Targets for miR-140-5p. *STEM CELLS*, 33(11), pp.3266–3280.

Beck, B. et al., 2013. Unravelling cancer stem cell potential. *Nature Reviews Cancer*, 13(10), pp.727–738.

Becker, J.S., Nicetto, D. & Zaret, K.S., 2016. H3K9me3-Dependent Heterochromatin: Barrier to Cell Fate Changes. *Trends in genetics : TIG*, 32(1), pp.29–41.

Bell, R.E. et al., 2016. Enhancer methylation dynamics contribute to cancer plasticity and patient mortality. *Genome Research*, 26(5), pp.601–611.

Berger, S.L. et al., 2009. An operational definition of epigenetics. *Genes & development*, 23(7), pp.781–3.

Bernard, P. et al., 2003. Dimerization of SOX9 is required for chondrogenesis, but not for sex determination. *Human molecular genetics*, 12(14), pp.1755–65. Available at: http://www.ncbi.nlm.nih.gov/pubmed/12837698 [Accessed July 13, 2017].

Bernstein, B.E. et al., 2006. A bivalent chromatin structure marks key developmental genes in embryonic stem cells. *Cell*, 125(2), pp.315–26.

Bertheloot, D. & Latz, E., 2017. HMGB1, IL-1?, IL-33 and S100 proteins: dual-function alarmins. *Cellular & Molecular Immunology*, 14(1), pp.43–64.

Bethel, M. et al., 2013. The changing balance between osteoblastogenesis and adipogenesis in aging and its impact on hematopoiesis. *Current osteoporosis reports*, 11(2), pp.99–106.

Bhosale, A.M. & Richardson, J.B., 2008. Articular cartilage: structure, injuries and review of management. *British medical bulletin*, 87, pp.77–95.

Bi, W. et al., 2001. Haploinsufficiency of Sox9 results in defective cartilage primordia and premature skeletal mineralization. *Proceedings of the National Academy of Sciences of the United States of America*, 98(12), pp.6698–703.

Bird, A., 2007. Perceptions of epigenetics. *Nature*, 447(7143), pp.396–398.

Blache, P. et al., 2004. SOX9 is an intestine crypt transcription factor, is regulated by the Wnt pathway, and represses the CDX2 and MUC2 genes. *The Journal of cell biology*, 166(1), pp.37–47.

Blackledge, N.P. et al., 2014. Variant PRC1 complex-dependent H2A ubiquitylation drives PRC2 recruitment and polycomb domain formation. *Cell*, 157(6), pp.1445–59.

Blanpain, C. et al., 2007. Epithelial Stem Cells: Turning over New Leaves. *Cell*, 128(3), pp.445–458.

Blattler, A. et al., 2014. Global loss of DNA methylation uncovers intronic enhancers in genes showing expression changes. *Genome Biology*, 15(9), p.469.

Boros, J. et al., 2014. Polycomb repressive complex 2 and H3K27me3 cooperate with H3K9 methylation to maintain heterochromatin protein 1α at chromatin. *Molecular and cellular biology*, 34(19), pp.3662–74.

Boyle, K.B. et al., 2009. The transcription factors Egr1 and Egr2 have opposing influences on adipocyte differentiation. *Cell Death and Differentiation*, 16(5), pp.782–789.

Brandeis, M. et al., 1994. Spl elements protect a CpG island from de novo methylation. *Nature*, 371(6496), pp.435–438.

Breschi, A., Gingeras, T.R. & Guigó, R., 2017. Comparative transcriptomics in human and mouse. *Nature Reviews Genetics*, 18(7), pp.425–440.

Brown, P.T., Squire, M.W. & Li, W.-J., 2014. Characterization and evaluation of mesenchymal stem cells derived from human embryonic stem cells and bone marrow. *Cell and tissue research*, 358(1), pp.149–64.

Bruna, F. et al., 2016. Regenerative Potential of Mesenchymal Stromal Cells: Age-Related Changes. *Stem Cells International*, 2016, pp.1–15.

Brunet, L.J. et al., 1998. Noggin, cartilage morphogenesis, and joint formation in the mammalian skeleton. *Science (New York, N.Y.)*, 280, pp.1455–1457.

Butler, D.L., Juncosa, N. & Dressler, M.R., 2004. Functional efficacy of tendon repair processes. *Annual review of biomedical engineering*, 6, pp.303–29.

Calo, E. & Wysocka, J., 2013. Modification of Enhancer Chromatin: What, How, and Why? *Molecular Cell*, 49(5), pp.825–837.

Cedar, H. & Bergman, Y., 2009. Linking DNA methylation and histone modification: patterns and paradigms. *Nature Reviews Genetics*, 10(5), pp.295–304.

Chaboissier, M.-C. et al., 2004. Functional analysis of Sox8 and Sox9 during sex determination in the mouse. *Development*, 131(9):1891-901

Chambers, D. et al., 2002. An enhancer complex confers both high-level and cell-specific expression of the human type X collagen gene. *FEBS Letters*, 531(3), pp.505–508.

Chatterjee, A. et al., 2012. Comparison of alignment software for genome-wide bisulphite sequence data. *Nucleic Acids Research*, 40(10), pp.e79–e79.

Chen, Y. et al., 2012. Systematic evaluation of factors influencing ChIP-seq fidelity. *Nature methods*, 9(6), pp.609–14.

Chinwalla, A.T. et al., 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature*, 420(6915), pp.520–562.

Cho, K.W.Y., 2012. Enhancers. *Wiley interdisciplinary reviews. Developmental biology*, 1(4), pp.469–78.

Cico, A., Andrieu-Soler, C. & Soler, E., 2016. Enhancers and their dynamics during hematopoietic differentiation and emerging strategies for therapeutic action. *FEBS Letters*, 590(22), pp.4084–4104.

Cloos, P.A.C. et al., 2008. Erasing the methyl mark: histone demethylases at the center of cellular differentiation and disease. *Genes & development*, 22(9), pp.1115–40.

Cohen, J., 1992. A power primer. *Psychological bulletin*, 112(1), pp.155–9.

Creyghton, M.P. et al., 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences*, 107(50), pp.21931–21936.

Creyghton, M.P. et al., 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proceedings of the National Academy of Sciences of the United States of America*, 107(50), pp.21931–6.

Cui, K. et al., 2009. Chromatin signatures in multipotent human hematopoietic stem cells indicate the fate of bivalent genes during differentiation. *Cell stem cell*, 4(1), pp.80–93.

Culley, K.L. et al., 2013. Class I histone deacetylase inhibition modulates metalloproteinase expression and blocks cytokine-induced cartilage degradation. *Arthritis and rheumatism*, 65(7), pp.1822–30.

de Andrés, M.C. et al., 2013. Loss of methylation in CpG sites in the NF-κB enhancer elements of inducible nitric oxide synthase is responsible for gene induction in human articular chondrocytes. *Arthritis and rheumatism*, 65(3), pp.732–42.

de Boer, B.A. et al., 2014. OccuPeak: ChIP-Seq Peak Calling Based on Internal Background Modelling T. Langmann, ed. *PLoS ONE*, 9(6), p.e99844.

De Gobbi, M. et al., 2011. Generation of bivalent chromatin domains during cell fate decisions. *Epigenetics & Chromatin*, 4(1), p.9.

Derek K Lim, A.H. et al., 2010. SAC review DNA methylation: a form of epigenetic control of gene expression. *SAC review*, 1212(1), pp.37–4237.

Detich, N. et al., 2003. The methyl donor S-Adenosylmethionine inhibits active demethylation of DNA: a candidate novel mechanism for the pharmacological effects of S-Adenosylmethionine. *The Journal of biological chemistry*, 278(23), pp.20812–20.

Diaz, A. et al., 2012. Normalization, bias correction, and peak calling for ChIP-seq. *Statistical applications in genetics and molecular biology*, 11(3), p.Article 9.

Dinant, C., Houtsmuller, A.B. & Vermeulen, W., 2008. Chromatin structure and DNA damage repair. *Epigenetics & Chromatin*, 1(1), p.9.

Dixon, J.R. et al., 2015. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539), pp.331–6.

Dolgin, E., 2016. Venerable brain-cancer cell line faces identity crisis. *Nature*, 537(7619), pp.149–150.

Dominguez, A.A., Lim, W.A. & Qi, L.S., 2016. Beyond editing: repurposing CRISPR-Cas9 for precision genome regulation and interrogation. *Nature reviews. Molecular cell biology*, 17(1), pp.5–15.

Dominici, M. et al., 2006. Minimal criteria for defining multipotent mesenchymal stromal cells. The International Society for Cellular Therapy position statement. *Cytotherapy*, 8(4), pp.315–317.

Du, N. et al., 2014. EGR2 is critical for peripheral naïve T-cell differentiation and the T-cell response to influenza. *Proceedings of the National Academy of Sciences of the United States of America*, 111(46), pp.16484–9.

Dudziec, E. et al., 2012. Integrated Epigenome Profiling of Repressive Histone Modifications, DNA Methylation and Gene Expression in Normal and Malignant Urothelial Cells M. Campbell, ed. *PLoS ONE*, 7(3), p.e32750.

Dukler, N. et al., 2016. Is a super-enhancer greater than the sum of its parts? *Nature Genetics*, 49(1), pp.2–3.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Dunham, I. et al., 2012. An integrated encyclopedia of DNA elements in the human genome. *Nature*, 489(7414), pp.57–74.

Dvir-Ginzberg, M. et al., 2008. Regulation of cartilage-specific gene expression in human chondrocytes by SirT1 and nicotinamide phosphoribosyltransferase. *The Journal of biological chemistry*, 283(52), pp.36300–10.

Eaves, C.J., 2015. Hematopoietic stem cells: concepts, definitions, and the new reality. *Blood*, 125(17):2605-13

Eberharter, A. & Becker, P.B., 2002. Histone acetylation: a switch between repressive and permissive chromatin. Second in review series on chromatin dynamics. *EMBO reports*, 3(3), pp.224–9.

Ebrahim, S., 2012. Epigenetics: the next big thing: Figure 1. *International Journal of Epidemiology*, 41(1), pp.1–3.

Ehrlich, K.C. et al., 2016. DNA Hypomethylation in Intragenic and Intergenic Enhancer Chromatin of Muscle-Specific Genes Usually Correlates with their Expression. *The Yale journal of biology and medicine*, 89(4), pp.441–455.

El-Jawhari, J.J. et al., 2016. Collagen-containing scaffolds enhance attachment and proliferation of non-cultured bone marrow multipotential stromal cells. *Journal of Orthopaedic Research*, 34(4), pp.597–606.

Elsafadi, M. et al., 2016. Transgelin is a TGF?-inducible gene that regulates osteoblastic and adipogenic differentiation of human skeletal stem cells through actin cytoskeleston organization. *Cell Death and Disease*, 7(8), p.e2321.

Ernst, J. & Kellis, M., 2010. Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, 28(8), pp.817–825.

Ernst, J. & Kellis, M., 2012. ChromHMM: automating chromatin-state discovery and characterization. *Nature Methods*, 9(3), pp.215–216.

Eyre, D.R., Weis, M.A. & Wu, J.-J., 2008. Advances in collagen cross-link analysis. *Methods (San Diego, Calif.)*, 45(1), pp.65–74.

Felisbino, M.B. et al., 2013. Chromatin remodeling induced by histone deacetylase inhibitors (HDACis) in HeLa, NIH 3T3 and HepG2 cells. *Epigenetics & Chromatin*, 6(Suppl 1), p.P17.

Feng, J. et al., 2012. Identifying ChIP-seq enrichment using MACS. *Nature protocols*, 7(9), pp.1728–40.

Ferrari, S. et al., 2004. Chromatin domain boundaries delimited by a histone-binding protein in yeast. *Journal of Biological Chemistry*, 279(53), pp.55520–55530.

Flensburg, C. et al., 2014. A comparison of control samples for ChIP-seq of histone modifications. *Frontiers in genetics*, 5, p.329.

Fortin, J.-P. et al., 2014. Functional normalization of 450k methylation array data improves replication in large cancer studies. *Genome biology*, 15(12), p.503.

Frazier, S.B. et al., 2008. The Quantification of Glycosaminoglycans: A Comparison of HPLC, Carbazole, and Alcian Blue Methods. *Open Glycoscience*, 1(1), pp.31–39.

Furey, T.S., 2012. ChIP–seq and beyond: new and improved methodologies to detect and characterize protein–DNA interactions. *Nature Reviews Genetics*, 13(12), pp.840–852.

Furlan-Magaril, M., Rincón-Arano, H. & Recillas-Targa, F., 2009. Sequential Chromatin Immunoprecipitation Protocol: ChIP-reChIP. In *Methods in molecular biology (Clifton, N.J.)*. pp. 253–266.

Furumatsu, T. & Asahara, H., 2010. Histone acetylation influences the activity of Sox9-related transcriptional complex. *Acta medica Okayama*, 64(6), pp.351–7.

Furumatsu, T. & Ozaki, T., 2010. Epigenetic regulation in chondrogenesis. *Acta medica Okayama*, 64(3), pp.155–61.

Galloway, M.T., Lalley, A.L. & Shearn, J.T., 2013. The role of mechanical loading in tendon development, maintenance, injury, and repair. *The Journal of bone and joint surgery. American volume*, 95(17), pp.1620–8.

Gebauer, M. et al., 2005. Comparison of the chondrosarcoma cell line SW1353 with primary human adult articular chondrocytes with regard to their gene expression profile and reactivity to IL-1β. *Osteoarthritis and Cartilage*, 13(8), pp.697–708.

Geiman, T.M. & Robertson, K.D., 2002. Chromatin remodeling, histone modifications, and DNA methylation - How does it all fit together? *Journal of Cellular Biochemistry*, 87(2), pp.117–125.

Ghaoui, L. El, Viallon, V. & Rabbani, T., 2010. Safe Feature Elimination for the LASSO and Sparse Supervised Learning Problems.

Gharibi, B. & Hughes, F.J., 2012. Effects of medium supplements on proliferation, differentiation potential, and in vitro expansion of mesenchymal stem cells. *Stem cells translational medicine*, 1(11), pp.771–82.

Giannoukakis, N. et al., 1993. Parental genomic imprinting of the human IGF2 gene. *Nature Genetics*, 4(1), pp.98–101.

Gil, J.A. et al., 2017. Challenges of Fracture Management for Adults With Osteogenesis Imperfecta. *Orthopedics*, 40(1), pp.e17–e22.

Goldberg, A. et al., 2017. The use of mesenchymal stem cells for cartilage repair and regeneration: a systematic review. *Journal of orthopaedic surgery and research*, 12(1), p.39.

Goldring, M.B., 2012. Chondrogenesis, chondrocyte differentiation, and articular cartilage metabolism in health and osteoarthritis. *Therapeutic Advances in Musculoskeletal Disease*, 4, pp.269–285.

Goldring, M.B. et al., 2005. Human chondrocyte cultures as models of cartilage-specific gene regulation. *Methods in molecular medicine*, 107, pp.69–95.

Gomez-Cabrero, D. et al., 2014. Data integration in the era of omics: current and future challenges. *BMC systems biology*, 8 Suppl 2(Suppl 2), p.I1.

Gómez-Picos, P. & Eames, B.F., 2015. On the evolutionary relationship between chondrocytes and osteoblasts. *Frontiers in genetics*, 6, p.297.

Gomoll, A.H. & Minas, T., 2014. The quality of healing: Articular cartilage. *Wound Repair and Regeneration*, 22(S1), pp.30–38.

Goodwin, S., McPherson, J.D. & McCombie, W.R., 2016. Coming of age: ten years of next-generation sequencing technologies. *Nature Reviews Genetics*, 17(6), pp.333–351.

Gregg, C. et al., 2010. High-Resolution Analysis of Parent-of-Origin Allelic Expression in the Mouse Brain. *Science*, 329(5992), pp.643–648.

Gryder, B.E., Sodji, Q.H. & Oyelere, A.K., 2012. Targeted cancer therapy: giving histone deacetylase inhibitors all they need to succeed. *Future medicinal chemistry*, 4(4), pp.505–24.

Guenther, M.G. et al., 2007. A chromatin landmark and transcription initiation at most promoters in human cells. *Cell*, 130(1), pp.77–88.

Guerquin, M.-J. et al., 2013. Transcription factor EGR1 directs tendon differentiation and promotes tendon repair. *Journal of Clinical Investigation*, 123(8), pp.3564–3576.

Guerrero, L. et al., 2010. Secondary enhancers synergise with primary enhancers to guarantee fine-tuned muscle gene expression. *Developmental Biology*, 337(1), pp.16–28.

Hagen, J.B., 2000. The origins of bioinformatics. *Nature Reviews Genetics*, 1(3), pp.231–236.

Halfmann, R. & Lindquist, S., 2010. Epigenetics in the Extreme: Prions and the Inheritance of Environmentally Acquired Traits. *Science*, 330(6004), pp.629–632.

Hall, B.K. & Miyake, T., 2000. All for one and one for all: condensations and the initiation of skeletal development. *BioEssays : news and reviews in molecular, cellular and developmental biology*, 22, pp.138–147.

Haniffa, M.A. et al., 2009. Mesenchymal stem cells: the fibroblasts' new clothes? *Haematologica*, 94(2), pp.258–63.

Haque, N., Kasim, N.H.A. & Rahman, M.T., 2015. Optimization of pre-transplantation conditions to enhance the efficacy of mesenchymal stem cells. *International journal of biological sciences*, 11(3), pp.324–34.

Hashimshony, T. et al., 2003. The role of DNA methylation in setting up chromatin structure during development. *Nature Genetics*, 34(2), pp.187–192.

Hata, K., 2015. Epigenetic regulation of chondrocyte differentiation. *Japanese Dental Science Review*, 51(4), pp.105–113.

Hata, K. et al., 2013. Arid5b facilitates chondrogenesis by recruiting the histone demethylase Phf2 to Sox9-regulated genes. *Nature Communications*, 4, p.2850.

Hay, D. et al., 2016. Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, 48(8), pp.895–903.

Hayman, D.M. et al., 2006. The Effects of Isolation on Chondrocyte Gene Expression. *Tissue Engineering*, 12(9), pp.2573–2581.

He, B. et al., 2014. Global view of enhancer-promoter interactome in human cells. *Proceedings of the National Academy of Sciences of the United States of America*, 111(21), pp.E2191-9.

He, X. et al., 2016. AP-1 family members act with Sox9 to promote chondrocyte hypertrophy. *Development*, 143(16):3012-23

Heinz, S. et al., 2015. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3), pp.144–154.

Heinz, S. et al., 2015. The selection and function of cell type-specific enhancers. *Nature Reviews Molecular Cell Biology*, 16(3), pp.144–154.

Hemberg, M. & Kreiman, G., 2011. Conservation of transcription factor binding events predicts gene expression across species. *Nucleic acids research*, 39(16), pp.7092–102.

Herlofsen, S.R. et al., 2013. Genome-wide map of quantified epigenetic changes during in vitro chondrogenic differentiation of primary human mesenchymal stem cells. *BMC Genomics*, 14(1), p.105.

Hilton, I.B. et al., 2015. Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, 33(5), pp.510–517.

Hnisz, D. et al., 2013. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4), pp.934–47.

Hodkinson, B.P. & Grice, E.A., 2015. Next-Generation Sequencing: A Review of Technologies and Tools for Wound Microbiome Research. *Advances in wound care*, 4(1), pp.50–58.

Hoffmann, A. et al., 2006. Neotendon formation induced by manipulation of the Smad8 signalling pathway in mesenchymal stem cells. *Journal of Clinical Investigation*, 116(4), pp.940–952.

Hogeweg, P., 2011. The roots of bioinformatics in theoretical biology. *PLoS computational biology*, 7(3), p.e1002021.

Hon, G.C., Hawkins, R.D. & Ren, B., 2009. Predictive chromatin signatures in the mammalian genome. *Human molecular genetics*, 18(R2), pp.R195-201.

Houde, A.-A. et al., 2014. Cross-tissue comparisons of leptin and adiponectin: DNA methylation profiles. *Adipocyte*, 3(2), pp.132–40.

Huang, C.-Y.C., Reuben, P.M. & Cheung, H.S., 2005. Temporal Expression Patterns and Corresponding Protein Inductions of Early Responsive Genes in Rabbit Bone Marrow-Derived Mesenchymal Stem Cells Under Cyclic Compressive Loading. *Stem Cells*, 23(8), pp.1113–1121.

Huang, L. et al., 2015. Proteomic analysis of porcine mesenchymal stem cells derived from bone marrow and umbilical cord: implication of the proteins involved in the higher migration capability of bone marrow mesenchymal stem cells. *Stem cell research & therapy*, 6(1), p.77.

Huang, W. et al., 2007. Signaling and transcriptional regulation in osteoblast commitment and differentiation. *Frontiers in bioscience : a journal and virtual library*, 12, pp.3068–92.

Huh, Y.H., Ryu, J.-H. & Chun, J.-S., 2007. Regulation of type II collagen expression by histone deacetylase in articular chondrocytes. *The Journal of biological chemistry*, 282(23), pp.17123–31.

Huynh, N.P.T. et al., 2017. Emerging roles for long noncoding RNAs in skeletal biology and disease. *Connective tissue research*, 58(1), pp.116–141.

Iftikhar, M. et al., 2011. Lysyl oxidase-like-2 (LOXL2) is a major isoform in chondrocytes and is critically required for differentiation. *The Journal of biological chemistry*, 286(2), pp.909–18.

Illingworth, R.S. & Bird, A.P., 2009. CpG islands - "A rough guide." *FEBS Letters*, 583(11), pp.1713–1720.

Iturbide, A., García de Herreros, A. & Peiró, S., 2015. A new role for LOX and LOXL2 proteins in transcription regulation. *FEBS Journal*, 282(9), pp.1768–1773.

Jenuwein, T., 2001. Translating the Histone Code. *Science*, 293(5532), pp.1074–1080.

Jo, A. et al., 2014. The versatile functions of Sox9 in development, stem cells, and human diseases. *Genes & Diseases*, 1(2), pp.149–161.

Johnstone, B. et al., 1998. In VitroChondrogenesis of Bone Marrow-Derived Mesenchymal Progenitor Cells. *Experimental Cell Research*, 238(1), pp.265–272.

Juan, D. et al., 2013. Late-replicating CNVs as a source of new genes. *Biology open*, 2(12), pp.1402–11.

Jung, Y.L. et al., 2014. Impact of sequencing depth in ChIP-seq experiments. *Nucleic acids research*, 42(9), p.e74.

Kadaja, M. et al., 2014. SOX9: a stem cell transcriptional regulator of secreted niche signaling factors. *Genes & development*, 28(4), pp.328–41.

Kadoch, C., Copeland, R.A. & Keilhack, H., 2016. PRC2 and SWI/SNF Chromatin Remodeling Complexes in Health and Disease. *Biochemistry*, 55(11), pp.1600–1614.

Kanehisa, M. & Bork, P., 2003. Bioinformatics in the post-sequence era. *Nature Genetics*, 33(3s), pp.305–310.

Kanhere, A. & Bansal, M., 2005. Structural properties of promoters: similarities and differences between prokaryotes and eukaryotes. *Nucleic acids research*, 33(10), pp.3165–75.

Karin, M., 1995. The regulation of AP-1 activity by mitogen-activated protein kinases. *The Journal of biological chemistry*, 270(28), pp.16483–6.

Karmodiya, K. et al., 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. *BMC genomics*, 13(1), p.424.

Katan-Khaykovich, Y. & Struhl, K., 2005. Heterochromatin formation involves changes in histone modifications over multiple cell generations. *The EMBO journal*, 24(12), pp.2138–49.

Kawakami, Y., Rodriguez-León, J. & Belmonte, J.C.I., 2006. The role of TGFβs and Sox9 during limb chondrogenesis. *Current Opinion in Cell Biology*, 18(6), pp.723–729.

Kent, W.J. et al., 2002. The human genome browser at UCSC. *Genome research*, 12(6), pp.996–1006.

Khan, A. & Zhang, X., 2016. dbSUPER: a database of super-enhancers in mouse and human genome. *Nucleic Acids Research*, 44(D1), pp.D164–D171.

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

KIANI, C. et al., 2002. Structure and function of aggrecan. *Cell Research*, 12(1), pp.19–32.

Kidder, B.L., Hu, G. & Zhao, K., 2011. ChIP-Seq: technical considerations for obtaining high-quality data. *Nature Immunology*, 12(10), pp.918–922.

Kizer, K.O. et al., 2005. A novel domain in Set2 mediates RNA polymerase II interaction and couples histone H3 K36 methylation with transcript elongation. *Molecular and cellular biology*, 25(8), pp.3305–16.

Klein-Nulend, J., Bacabac, R.G. & Bakker, A.D., 2012. Mechanical loading and how it affects bone cells: the role of the osteocyte cytoskeleton in maintaining our skeleton. *European cells & materials*, 24, pp.278–91.

Kleinjan, D.J. & van Heyningen, V., 1998. Position effect in human genetic disease. *Human molecular genetics*, 7(10), pp.1611–8.

Kolaparthy, L.K. et al., 2015. Adipose Tissue - Adequate, Accessible Regenerative Material. *International journal of stem cells*, 8(2), pp.121–7.

Kolasinska-Zwierz, P. et al., 2009. Differential chromatin marking of introns and expressed exons by H3K36me3. *Nature Genetics*, 41(3), pp.376–381.

Kolf, C.M., Cho, E. & Tuan, R.S., 2007. Mesenchymal stromal cells. Biology of adult mesenchymal stem cells: regulation of niche, self-renewal and differentiation. *Arthritis research & therapy*, 9(1), p.204.

Koopman, P., 2001. Sry, Sox9 and mammalian sex determination. *EXS*, (91), pp.25–56.

Korkmaz, G. et al., 2016. Functional genetic screens for enhancer elements in the human genome using CRISPR-Cas9. *Nature Biotechnology*, 34(2), pp.192–198.

Kornblum, H.I., 2007. Introduction to Neural Stem Cells. *Stroke*, 38(2).

Krebsbach, P.H. et al., 1996. Identification of a minimum enhancer sequence for the type II collagen gene reveals several core sequence motifs in common with the link protein gene. *The Journal of biological chemistry*, 271(8), pp.4298–303.

Kretlow, J.D. et al., 2008. Donor age and cell passage affects differentiation potential of murine bone marrow-derived stem cells. *BMC cell biology*, 9, p.60.

Kriaucionis, S. & Heintz, N., 2009. The Nuclear DNA Base 5-Hydroxymethylcytosine Is Present in Purkinje Neurons and the Brain. *Science*, 324(5929), pp.929–930.

Kundaje, A. et al., 2015. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539), pp.317–330.

Kuras, L., Borggrefe, T. & Kornberg, R.D., 2003. Association of the Mediator complex with enhancers of active genes. *Proceedings of the National Academy of Sciences of the United States of America*, 100(24), pp.13887–91.

Kvist, A.J. et al., 2006. Chondroitin sulfate perlecan enhances collagen fibril formation. Implications for perlecan chondrodysplasias. *The Journal of biological chemistry*, 281(44), pp.33127–39.

Lai, W.-T., Krishnappa, V. & Phinney, D.G., 2011. Fibroblast growth factor 2 (Fgf2) inhibits differentiation of mesenchymal stem cells by inducing Twist2 and Spry4, blocking extracellular regulated kinase activation, and altering Fgf receptor expression levels. *Stem cells (Dayton, Ohio)*, 29(7), pp.1102–11.

Lam, M.T.Y. et al., 2014. Enhancer RNAs and regulated transcriptional programs. *Trends in biochemical sciences*, 39(4), pp.170–82.

Lander, E.S. et al., 2001. Initial sequencing and analysis of the human genome. *Nature*, 409(6822), pp.860–921.

Landt, S.G. et al., 2012. ChIP-seq guidelines and practices of the ENCODE and modENCODE consortia. *Genome research*, 22(9), pp.1813–31.

Langmead, B. & Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. *Nature methods*, 9(4), pp.357–9.

Lee, J.-E. & Kim, Y., 2006. A tissue-specific variant of the human lysyl oxidase-like protein 3 (LOXL3) functions as an amine oxidase with substrate specificity. *The Journal of biological chemistry*, 281(49), pp.37282–90.

Leek, J.T. et al., 2012. The sva package for removing batch effects and other unwanted variation in high-throughput experiments. *Bioinformatics (Oxford, England)*, 28(6), pp.882–3.

Lefebvre, V.R. et al., 1997. SOX9 Is a Potent Activator of the Chondrocyte-Specific Enhancer of the Pro ⎵ 1(II) Collagen Gene. , 17(4), pp.2336–2346.

Lelièvre, E. et al., 2001. Signaling Pathways Recruited by the Cardiotrophin-like Cytokine/Cytokine-like Factor-1 Composite Cytokine. *Journal of Biological Chemistry*, 276(25), pp.22476–22484.

Lengner, C.J. et al., 2005. Nkx3.2-mediated repression of Runx2 promotes chondrogenic differentiation. *The Journal of biological chemistry*, 280(16), pp.15872–9.

Leontiou, C.A. et al., 2015. Bisulfite Conversion of DNA: Performance Comparison of Different Kits and Methylation Quantitation of Epigenetic Biomarkers that Have the Potential to Be Used in Non-Invasive Prenatal Testing O. El-Maarri, ed. *PLOS ONE*, 10(8), p.e0135058.

Leung, V.Y.L. et al., 2011. SOX9 Governs Differentiation Stage-Specific Gene Expression in Growth Plate Chondrocytes via Direct Concomitant Transactivation and Repression V. van Heyningen, ed. *PLoS Genetics*, 7(11), p.e1002356.

Li, C. et al., 2015. Comparative analysis of human mesenchymal stem cells from bone marrow and adipose tissue under xeno-free conditions for cell therapy. *Stem Cell Research & Therapy*, 6(1), p.55.

Li, H. et al., 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, 25(16), pp.2078–2079.

Li, Q. et al., 2011. Measuring reproducibility of high-throughput experiments. *The Annals of Applied Statistics*, 5(3), pp.1752–1779.

Li, W., Notani, D. & Rosenfeld, M.G., 2016. Enhancers as non-coding RNA transcription units: recent insights and future perspectives. *Nature Reviews Genetics*, 17(4), pp.207–223.

Lim, D.H.K. & Maher, E.R., 2010. Genomic Imprinting Syndromes and Cancer. In *Advances in genetics*. pp. 145–175.

Liu, C.-F. & Lefebvre, V., 2015. The transcription factors SOX9 and SOX5/SOX6 cooperate genome-wide through super-enhancers to drive chondrogenesis. *Nucleic Acids Research*, 43(17), pp.8183–8203.

Liu, R. & States, D.J., 2002. Consensus promoter identification in the human genome utilizing expressed gene markers and gene modeling. *Genome research*, 12(3), pp.462–9.

Lo, B. & Parham, L., 2009. Ethical issues in stem cell research. *Endocrine reviews*, 30(3), pp.204–13.

Loeser, R.F., Collins, J.A. & Diekman, B.O., 2016. Ageing and the pathogenesis of osteoarthritis. *Nature Reviews Rheumatology*, 12(7), pp.412–420.

Lorda-Diez, C.I. et al., 2011. Defining the earliest transcriptional steps of chondrogenic progenitor specification during the formation of the digits in the embryonic limb. *PLoS ONE*, 6(9):e24546

Lovén, J. et al., 2013. Selective inhibition of tumor oncogenes by disruption of super-enhancers. *Cell*, 153(2), pp.320–34.

LoVerso, P.R. & Cui, F., 2015. A Computational Pipeline for Cross-Species Analysis of RNA-seq Data Using R and Bioconductor. *Bioinformatics and biology insights*, 9, pp.165–74.

Lövkvist, C. et al., 2016. DNA methylation in human epigenomes depends on local topology of CpG sites. *Nucleic Acids Research*, 44(11), pp.5123–5132.

Ma, S. et al., 2014. Immunobiology of mesenchymal stem cells. *Cell Death and Differentiation*, 21(2), pp.216–225.

Mackie, E.J. et al., 2008. Endochondral ossification: How cartilage is converted into bone in the developing skeleton. *International Journal of Biochemistry and Cell Biology*, 40, pp.46–62.

Macleod, D. et al., 1994. Sp1 sites in the mouse aprt gene promoter are required to prevent methylation of the CpG island. *Genes & development*, 8(19), pp.2282–92.

Mäki, J.M. et al., 2002. Inactivation of the lysyl oxidase gene Lox leads to aortic aneurysms, cardiovascular dysfunction, and perinatal death in mice. *Circulation*, 106(19), pp.2503–9.

Maldonado, M. & Nam, J., 2013. The role of changes in extracellular matrix of cartilage in the presence of inflammation on the pathology of osteoarthritis. *BioMed research international*, 2013, p.284873.

Maleki, M. et al., 2014. Comparison of mesenchymal stem cell markers in multiple human adult stem cells. *International journal of stem cells*, 7(2), pp.118–26.

Marks, P.A., Richon, V.M. & Rifkind, R.A., 2000. Histone Deacetylase Inhibitors: Inducers of Differentiation or Apoptosis of Transformed Cells. *JNCI Journal of the National Cancer Institute*, 92(15), pp.1210–1216.

Marsman, J. & Horsfield, J.A., 2012. Long distance relationships: Enhancer–promoter communication and dynamic gene transcription. *Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms*, 1819(11), pp.1217–1227.

McEwen, K.R. et al., 2013. The impact of culture on epigenetic properties of pluripotent stem cells and pre-implantation embryos. *Biochemical Society Transactions*, 41(3), pp.711–719.

McLean, C.Y. et al., 2010. GREAT improves functional interpretation of cis-regulatory regions. *Nature Biotechnology*, 28(5), pp.495–501.

McLeay, R.C. et al., 2012. Genome-wide in silico prediction of gene expression. *Bioinformatics*, 28(21), pp.2789–2796.

Meng, X. et al., 2014. Stem cells in a three-dimensional scaffold environment. *SpringerPlus*, 3, p.80.

Meyer, J. et al., 1997. Mutational analysis of the SOX9 gene in campomelic dysplasia and autosomal sex reversal: lack of genotype/phenotype correlations. *Human Molecular Genetics*, 6(1), pp.91–98.

Mignone, F. et al., 2002. Untranslated regions of mRNAs. *Genome biology*, 3(3), p.REVIEWS0004.

Mikkelsen, T.S. et al., 2007. Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. *Nature*, 448(7153), pp.553–60.

Miyaki, S. et al., 2010. MicroRNA-140 plays dual roles in both cartilage development and homeostasis. *Genes & development*, 24(11), pp.1173–85.

Mohrs, M. et al., 2001. Deletion of a coordinate regulator of type 2 cytokine expression in mice. *Nature Immunology*, 2(9), pp.842–847.

Mokry, M. et al., 2010. Efficient Double Fragmentation ChIP-seq Provides Nucleotide Resolution Protein-DNA Binding Profiles B. Jürg, ed. *PLoS ONE*, 5(11), p.e15092.

Moorthy, S.D. et al., 2017. Enhancers and super-enhancers have an equivalent regulatory role in embryonic stem cells through regulation of single or multiple genes. *Genome research*, 27(2), pp.246–258.

Mori-Akiyama, Y. et al., Sox9 is required for determination of the chondrogenic cell lineage in the cranial neural crest.

Muers, M., 2011. Functional genomics: The modENCODE guide to the genome. *Nature Reviews Genetics*, 12(2), pp.80–80.

Murdoch, A.D. et al., 2007. Chondrogenic Differentiation of Human Bone Marrow Stem Cells in Transwell Cultures: Generation of Scaffold-Free Cartilage. *Stem Cells*, 25(11), pp.2786–2796.

Namba, R.S. et al., 1998. Spontaneous repair of superficial defects in articular cartilage in a fetal lamb model. *The Journal of bone and joint surgery. American volume*, 80, pp.4–10.

Niederriter, A.R. et al., 2015. Super Enhancers in Cancers, Complex Disease, and Developmental Disorders. *Genes*, 6(4), pp.1183–200.

Noer, A., Lindeman, L.C. & Collas, P., 2009. Histone H3 Modifications Associated With Differentiation and Long-Term Culture of Mesenchymal Adipose Stem Cells. *Stem Cells and Development*, 18(5), pp.725–736.

Nombela-Arrieta, C., Ritz, J. & Silberstein, L.E., 2011. The elusive nature and function of mesenchymal stem cells. *Nature Reviews. Molecular Cell Biology*, 12(2), p.126.

Oh, S.-J. et al., 2017. Human U87 glioblastoma cells with stemness features display enhanced sensitivity to natural killer cell cytotoxicity through altered expression of NKG2D ligand. *Cancer cell international*, 17, p.22.

Ohba, S. et al., 2015. Distinct Transcriptional Programs Underlie Sox9 Regulation of the Mammalian Chondrocyte. *Cell reports*, 12(2), pp.229–43.

Ong, C.-T. & Corces, V.G., 2011. Enhancer function: new insights into the regulation of tissue-specific gene expression. *Nature reviews. Genetics*, 12(4), pp.283–93.

Orlando, D. et al., 2014. Quantitative ChIP-Seq Normalization Reveals Global Modulation of the Epigenome. *Cell Reports*, 9(3), pp.1163–1170.

Paassilta, P. et al., 1999. COL9A3: A Third Locus for Multiple Epiphyseal Dysplasia. *The American Journal of Human Genetics*, 64(4), pp.1036–1044.

Pacifici, M. et al., 1990. Hypertrophic chondrocytes. The terminal stage of differentiation in the chondrogenic cell lineage? *Annals of the New York Academy of Sciences*, 599, pp.45–57.

Palazzo, A.F. & Lee, E.S., 2015. Non-coding RNA: what is functional and what is junk? *Frontiers in genetics*, 6, p.2.

Park, J.S. et al., 2015. Engineering mesenchymal stem cells for regenerative medicine and drug delivery. *Methods*, 84, pp.3–16.

Park, J. et al., 2015. Dual pathways to endochondral osteoblasts: a novel chondrocyte-derived osteoprogenitor cell identified in hypertrophic cartilage. *Biology Open*.

Parker, S.C.J. et al., 2013. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences of the United States of America*, 110(44), pp.17921–6.

Patel, S. et al., 2015. Integrating Epigenetic Modulators into NanoScript for Enhanced Chondrogenesis of Stem Cells. *Journal of the American Chemical Society*, 137(14), pp.4598–4601.

Patro, R., Duggal, G. & Kingsford, C., 2015. Salmon: Accurate, Versatile and Ultrafast Quantification from RNA-seq Data using Lightweight-Alignment. *bioRxiv*.

Payne, K.A., Didiano, D.M. & Chu, C.R., 2010. Donor sex and age influence the chondrogenic potential of human femoral bone marrow stem cells. *Osteoarthritis and cartilage*, 18(5), pp.705–13.

Pchelintsev, N.A. et al., 2016. Critical Parameters for Efficient Sonication and Improved Chromatin Immunoprecipitation of High Molecular Weight Proteins M. Wu, ed. *PLOS ONE*, 11(1), p.e0148023.

Peffers, M.J. et al., 2016. Age-related changes in mesenchymal stem cells identified using a multi-omics approach. *European cells & materials*, 31, pp.136–59.

Pennacchio, L.A. et al., 2013. Enhancers: five essential questions. *Nature reviews. Genetics*, 14(4), pp.288–95.

Pertea, M., 2012. The human transcriptome: An unfinished story. *Genes*, 3(3), pp.344–360.

Pertea, M. et al., 2016. Transcript-level expression analysis of RNA-seq experiments with HISAT, StringTie and Ballgown. *Nature Protocols*, 11(9), pp.1650–1667.

Peschansky, V.J. & Wahlestedt, C., 2014. Non-coding RNAs as direct and indirect modulators of epigenetic regulation. *Epigenetics*, 9(1), pp.3–12.

Phipson, B., Maksimovic, J. & Oshlack, A., 2015. missMethyl: an R package for analyzing data from Illumina’s HumanMethylation450 platform. *Bioinformatics*, 32(2), p.btv560.

Phipson, B., Maksimovic, J. & Oshlack, A., 2015. missMethyl: an R package for analyzing data from Illumina’s HumanMethylation450 platform. *Bioinformatics*, 32(2), p.btv560.

Pitsillides, A.A. & Beier, F., 2011. Cartilage biology in osteoarthritis—lessons from developmental biology. *Nature Reviews Rheumatology*, 7(11), pp.654–663.

Plath, K. et al., 2002. *Xist* RNA and the Mechanism of X Chromosome Inactivation. *Annual Review of Genetics*, 36(1), pp.233–278.

Plessy, C. et al., 2005. Enhancer sequence conservation between vertebrates is favoured in developmental regulator genes. *Trends in Genetics*, 21(4), pp.207–210.

Pombo, A. & Dillon, N., 2015. Three-dimensional genome architecture: players and mechanisms. *Nature Reviews Molecular Cell Biology*, 16(4), pp.245–257.

Pott, S. & Lieb, J.D., 2014. What are super-enhancers? *Nature Genetics*, 47(1), pp.8–12.

Quinlan, A.R. & Hall, I.M., 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6), pp.841–2.

Ravasi, T. et al., 2003. Systematic characterization of the zinc-finger-containing proteins in the mouse transcriptome. *Genome research*, 13(6B), pp.1430–42.

Richmond, T.J. et al., 1997. Crystal structure of the nucleosome core particle at 2.8|[thinsp]||[Aring]| resolution. *Nature*, 389(6648), pp.251–260.

Ritchie, M.E. et al., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7), pp.e47–e47.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–26.

Robinson, J.T. et al., 2011. Integrative genomics viewer. *Nature Biotechnology*, 29(1), pp.24–26.

Robinson, M.D. et al., 2010. A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biology*, 11(3), p.R25.

Roelen, B.A.J. & Ten Dijke, P., 2003. Controlling mesenchymal stem cell differentiation by TGF?? family members. *Journal of Orthopaedic Science*, 8(5), pp.740–748.

Romanoski, C.E. et al., 2015. Epigenomics: Roadmap for regulation. *Nature*, 518(7539), pp.314–316.

Rose, N.R. & Klose, R.J., 2014. Understanding the relationship between DNA methylation and histone lysine methylation. *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, 1839(12), pp.1362–1372.

Rushton, M.D. et al., 2014. Characterization of the Cartilage DNA Methylome in Knee and Hip Osteoarthritis. *Arthritis & Rheumatology*, 66(9), pp.2450–2460.

Saksouk, N., Simboeck, E. & Déjardin, J., 2015. Constitutive heterochromatin formation and transcription in mammals. *Epigenetics & Chromatin*, 8(1), p.3.

Santoro, A. et al., 2015. Choosing the right chondrocyte cell line: Focus on nitric oxide. *Journal of Orthopaedic Research*, 33(12), pp.1784–1788.

Sanyal, A. et al., 2012. The long-range interaction landscape of gene promoters. *Nature*, 489(7414), pp.109–13.

Schaft, D. et al., 2003. The histone 3 lysine 36 methyltransferase, SET2, is involved in transcriptional elongation. *Nucleic Acids Research*, 31(10), pp.2475–2482.

Scharstuhl, A. et al., 2007. Chondrogenic Potential of Human Adult Mesenchymal Stem Cells Is Independent of Age or Osteoarthritis Etiology. *Stem Cells*, 25(12), pp.3244–3251.

Schlissel, M., 2004. The spreading influence of chromatin modification. *Nature genetics*, 36(5), pp.438–440.

Schlötzer-Schrehardt, U. & Naumann, G.O.H., 2006. Ocular and Systemic Pseudoexfoliation Syndrome. *American Journal of Ophthalmology*, 141(5), p.921–937.e2.

Schmidl, C. et al., 2009. Lineage-specific DNA methylation in T cells correlates with histone methylation and enhancer activity. *Genome research*, 19(7), pp.1165–74.

Schmitt, A.D., Hu, M. & Ren, B., 2016. Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, 17(12), pp.743–755.

Scott, M.A. et al., 2011. Current methods of adipogenic differentiation of mesenchymal stem cells. *Stem cells and development*, 20(10), pp.1793–804.

Scuteri, A. et al., 2011. Mesenchymal stem cells neuronal differentiation ability: a real perspective for nervous system repair? *Current stem cell research & therapy*, 6(2), pp.82–92.

Seymour, P.A., 2014. Sox9: a master regulator of the pancreatic program. *The review of diabetic studies : RDS*, 11(1), pp.51–83.

Sharp, A.J. et al., 2011. DNA methylation profiles of human active and inactive X chromosomes. *Genome research*, 21(10), pp.1592–600.

Sharpless, N.E. & DePinho, R.A., 2007. How stem cells age and why this makes us grow old. *Nature Reviews Molecular Cell Biology*, 8(9), pp.703–713.

Sheaffer, K.L. et al., 2014. DNA methylation is required for the control of stem cell differentiation in the small intestine. *Genes & Development*, 28(6), pp.652–664.

Shen, J. et al., 2011. Arterial injury promotes medial chondrogenesis in Sm22 knockout mice. *Cardiovascular research*, 90(1), pp.28–37.

Shen, L. et al., 2014. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, 15(1), p.284.

Shi, Y. et al., 2004. Histone Demethylation Mediated by the Nuclear Amine Oxidase Homolog LSD1. *Cell*, 119(7), pp.941–953.

Shukla, S. et al., 2011. CTCF-promoted RNA polymerase II pausing links DNA methylation to splicing. *Nature*, 479(7371), pp.74–79.

Silva, W.A. et al., 2003. The Profile of Gene Expression of Human Marrow Mesenchymal Stem Cells. *Stem Cells*, 21(6), pp.661–669.

Singh, A., Singh, A. & Sen, D., 2016. Mesenchymal stem cells in cardiac regeneration: a detailed progress report of the last 6 years (2010-2015). *Stem cell research & therapy*, 7(1), p.82.

Smith, D.R., 2015. Broadening the definition of a bioinformatician. *Frontiers in genetics*, 6, p.258.

Smith, E.M. et al., 2016. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. *The American Journal of Human Genetics*, 98(1), pp.185–201.

Smith, Z.D. & Meissner, A., 2013. DNA methylation: roles in mammalian development. *Nature Reviews Genetics*, 14(3), pp.204–220.

Solchaga, L.A., Penick, K.J. & Welter, J.F., 2011. Chondrogenic differentiation of bone marrow-derived mesenchymal stem cells: tips and tricks. *Methods in molecular biology (Clifton, N.J.)*, 698, pp.253–78.

Somoza, R.A. et al., 2014. Chondrogenic differentiation of mesenchymal stem cells: challenges and unfulfilled expectations. *Tissue engineering. Part B, Reviews*, 20(6), pp.596–608.

Soneson, C., Love, M.I. & Robinson, M.D., 2015. Differential analyses for RNA-seq: transcript-level estimates improve gene-level inferences. *F1000Research*, 4, p.1521.

Spaapen, F. et al., 2013. The Immediate Early Gene Product EGR1 and Polycomb Group Proteins Interact in Epigenetic Programming during Chondrogenesis Y. Tsuji, ed. *PLoS ONE*, 8(3), p.e58083.

Stamatoyannopoulos, J.A. et al., 2012. An encyclopedia of mouse DNA elements (Mouse ENCODE). *Genome Biology*, 13(8), p.418.

Stark, R., DiffBind: Differential binding analysis of ChIP-Seq peak data. http://bioconductor.org/packages/release/bioc/vignettes/DiffBind/inst/doc/DiffBind.pdf.

Stolt, C.C. et al., 2003. The Sox9 transcription factor determines glial fate choice in the developing spinal cord. *Genes & development*, 17(13), pp.1677–89.

Stunnenberg, H.G. et al., 2016. The International Human Epigenome Consortium: A Blueprint for Scientific Collaboration and Discovery. *Cell*, 167(5), pp.1145–1149.

Sun, H.B., 2010. Mechanical loading, cartilage degradation, and arthritis. *Annals of the New York Academy of Sciences*, 1211, pp.37–50.

Sur, I. & Taipale, J., 2016. The role of enhancers in cancer. *Nature Reviews Cancer*, 16(8), pp.483–493.

Svensson, A. et al., 2017. Identification of two distinct mesenchymal stromal cell populations in human malignant glioma. *Journal of neuro-oncology*, 131(2), pp.245–254.

Tahiliani, M. et al., 2009. Conversion of 5-Methylcytosine to 5-Hydroxymethylcytosine in Mammalian DNA by MLL Partner TET1. *Science*, 324(5929), pp.930–935.

Tarantino, U. et al., 2011. Bone healing and osteoporosis. *Aging clinical and experimental research*, 23(2 Suppl), pp.62–4.

Tew, S.R. et al., 2007. SOX9 transduction of a human chondrocytic cell line identifies novel genes regulated in primary human chondrocytes and in osteoarthritis. *Arthritis research & therapy*, 9(5), p.R107.

Teytelman, L. et al., 2009. Impact of Chromatin Structures on DNA Processing for Genomic Analyses R. Aramayo, ed. *PLoS ONE*, 4(8), p.e6700.

Thesingh, C.W., Groot, C.G. & Wassenaar, A.M., 1991. Transdifferentiation of hypertrophic chondrocytes into osteoblasts in murine fetal metatarsal bones, induced by co-cultured cerebrum. *Bone and mineral*, 12(1), pp.25–40.

Thiel, A. et al., 2017. Osteoblast migration in vertebrate bone. *Biological Reviews of the Cambridge Philosophical Society*, 93(1):350-363

Thorleifsson, G. et al., 2007. Common Sequence Variants in the LOXL1 Gene Confer Susceptibility to Exfoliation Glaucoma. *Science*, 317(5843), pp.1397–1400.

Tian, Y. et al., 2016. MicroRNA-30a promotes chondrogenic differentiation of mesenchymal stem cells through inhibiting Delta-like 4 expression. *Life Sciences*, 148, pp.220–228.

Tiller, G.E. et al., 1995. Dominant mutations in the type II collagen gene, COL2A1, produce spondyloepimetaphyseal dysplasia, Strudwick type. *Nature Genetics*, 11(1), pp.87–89.

Trojer, P. & Reinberg, D., 2007. Facultative heterochromatin: is there a distinctive molecular signature? *Molecular cell*, 28(1), pp.1–13.

Vakoc, C.R. et al., 2006. Profile of Histone Lysine Methylation across Transcribed Mammalian Chromatin. *Molecular and Cellular Biology*, 26(24), pp.9185–9195.

van den Dungen, M.W. et al., 2016. Comprehensive DNA Methylation and Gene Expression Profiling in Differentiating Human Adipocytes. *Journal of Cellular Biochemistry*, 117(12), pp.2707–2718.

van der Kraan, P.M. & van den Berg, W.B., 2012. Chondrocyte hypertrophy and osteoarthritis: role in initiation and progression of cartilage degeneration? *Osteoarthritis and Cartilage*, 20(3), pp.223–232.

Vastenhouw, N.L. & Schier, A.F., 2012. Bivalent histone modifications in early embryogenesis. *Current opinion in cell biology*, 24(3), pp.374–86.

Villar, D. et al., 2015. Enhancer evolution across 20 mammalian species. *Cell*, 160(3), pp.554–66.

Visel, A. et al., 2008. Ultraconservation identifies a small subset of extremely constrained developmental enhancers. *Nature Genetics*, 40(2), pp.158–160.

Voigt, P., Tee, W.-W. & Reinberg, D., 2013. A double take on bivalent promoters. *Genes & development*, 27(12), pp.1318–38.

Völkel, S. et al., 2015. Zinc Finger Independent Genome-Wide Binding of Sp2 Potentiates Recruitment of Histone-Fold Protein Nf-y Distinguishing It from Sp1 and Sp3. *PLoS Genetics*, 20;11(3):e1005102

Von Der Mark, K. et al., 1992. Type x collagen synthesis in human osteoarthritic cartilage. indication of chondrocyte hypertrophy. *Arthritis & Rheumatism*, 35(7), pp.806–811.

Vong, K.I. et al., 2015. Sox9 is critical for suppression of neurogenesis but not initiation of gliogenesis in the cerebellum. *Molecular Brain*, 8(1), p.25.

Wagner, W. et al., 2008. Replicative senescence of mesenchymal stem cells: a continuous and organized process. *PloS one*, 3(5), p.e2213.

Wakabayashi, T. et al., 2010. Fibulin-3 negatively regulates chondrocyte differentiation. *Biochemical and Biophysical Research Communications*, 391(1), pp.1116–1121.

Wang, J.-P. et al., 2011. Trichostatin A inhibits TGF-β1 induced in vitro chondrogenesis of hMSCs through Sp1 suppression. *Differentiation*, 81(2), pp.119–126.

Wang, M. et al., 2017. Advances and Prospects in Stem Cells for Cartilage Regeneration. *Stem Cells International*, 2017, pp.1–16.

Wang, Q.-W., Chen, Z.-L. & Piao, Y.-J., 2005. Mesenchymal stem cells differentiate into tenocytes by bone morphogenetic protein (BMP) 12 gene transfer. *Journal of Bioscience and Bioengineering*, 100(4), pp.418–422.

Wang, S. et al., 2016. Modeling cis-regulation with a compendium of genome-wide histone H3K27ac profiles. *Genome research*, 26(10), pp.1417–1429.

Wang, Y. et al., 2017. The 3D Genome Browser: a web-based browser for visualizing 3D genome organization and long-range chromatin interactions. *bioRxiv*.

Watanabe, H., Yamada, Y. & Kimata, K., 1998. Roles of aggrecan, a large chondroitin sulfate proteoglycan, in cartilage structure and function. *Journal of biochemistry*, 124(4), pp.687–93.

Weber, M. et al., 2007. Distribution, silencing potential and evolutionary impact of promoter DNA methylation in the human genome. *Nature Genetics*, 39(4), pp.457–466.

Westhrin, M. et al., 2015. Osteogenic Differentiation of Human Mesenchymal Stem Cells in Mineralized Alginate Matrices G. Papaccio, ed. *PLOS ONE*, 10(3), p.e0120374.

Weth, O. et al., 2014. CTCF induces histone variant incorporation, erases the H3K27me3 histone mark and opens chromatin. *Nucleic Acids Research*, 42(19), pp.11941–11951.

Whalen, S., Truty, R.M. & Pollard, K.S., 2016. Enhancer–promoter interactions are encoded by complex genomic signatures on looping chromatin. *Nature Genetics*, 48(5), pp.488–496.

Whitney, G.A. et al., 2012. Methods for producing scaffold-free engineered cartilage sheets from auricular and articular chondrocyte cell sources and attachment to porous tantalum. *BioResearch open access*, 1(4), pp.157–65.

Whyte, W.A. et al., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2), pp.307–19.

Williams, E.L., White, K. & Oreffo, R.O.C., 2013. Isolation and Enrichment of Stro-1 Immunoselected Mesenchymal Stem Cells from Adult Human Bone Marrow. In *Methods in molecular biology (Clifton, N.J.)*. pp. 67–73.

Wood, A.M. et al., 2013. A Review on the Management of Hip and Knee Osteoarthritis. *International Journal of Chronic Diseases*, 2013, pp.1–10.

Worster, A.A. et al., 2000. Effect of transforming growth factor beta1 on chondrogenic differentiation of cultured equine mesenchymal stem cells. *American journal of veterinary research*, 61(9), pp.1003–10.

Wu, C., 1997. Chromatin remodeling and the control of gene expression. *The Journal of biological chemistry*, 272(45), pp.28171–4.

Wu, H. et al., 2017. Chromatin dynamics regulate mesenchymal stem cell lineage specification and differentiation to osteogenesis. *Biochimica et biophysica acta*, 1860(4), pp.438–449.

Wunderle, V.M. et al., 1998. Deletion of long-range regulatory elements upstream of SOX9 causes campomelic dysplasia. *Proceedings of the National Academy of Sciences of the United States of America*, 95(18), pp.10649–54.

Xie, M. et al., 2013. DNA hypomethylation within specific transposable element families associates with tissue-specific enhancer landscape. *Nature Genetics*, 45(7), pp.836–841.

Yammani, R.R., 2012. S100 proteins in cartilage: role in arthritis. *Biochimica et biophysica acta*, 1822(4), pp.600–6.

Yan, C. & Boyd, D.D., 2006. Histone H3 acetylation and H3 K4 methylation define distinct chromatin regions permissive for transgene expression. *Molecular and cellular biology*, 26(17), pp.6357–71.

Yang, B. et al., 2011. MicroRNA-145 Regulates Chondrogenic Differentiation of Mesenchymal Stem Cells by Targeting Sox9 R. Linden, ed. *PLoS ONE*, 6(7), p.e21679.

Yang, L. et al., 2014. Hypertrophic chondrocytes can become osteoblasts and osteocytes in endochondral bone formation. *Proceedings of the National Academy of Sciences of the United States of America*, 111(33), pp.12097–102.

Yang, X. et al., 2014. Gene body methylation can alter gene expression and is a therapeutic target in cancer. *Cancer Cell*, 26(4), pp.577–590.

Yang, Y. et al., 2014. LEVERAGING BIOLOGICAL REPLICATES TO IMPROVE ANALYSIS IN CHIP-SEQ EXPERIMENTS. *Computational and Structural Biotechnology Journal*, 9(13), p.e201401002.

Yapp, C. et al., 2016. H3K27me3 demethylases regulate in vitro chondrogenesis and chondrocyte activity in osteoarthritis. *Arthritis research & therapy*, 18(1), p.158.

Yoo, H.J. et al., 2011. Gene expression profile during chondrogenesis in human bone marrow derived mesenchymal stem cells using a cDNA microarray. *Journal of Korean medical science*, 26(7), pp.851–8.

Young, D.A. et al., 2005. Histone deacetylase inhibitors modulate metalloproteinase gene expression in chondrocytes and block cartilage resorption. *Arthritis research & therapy*, 7(3), pp.R503-12.

Young, M.D. et al., 2011. ChIP-seq analysis reveals distinct H3K27me3 profiles that correlate with transcriptional activity. *Nucleic acids research*, 39(17), pp.7415–27.

Young, R.G. et al., 1998. Use of mesenchymal stem cells in a collagen matrix for achilles tendon repair. *Journal of Orthopaedic Research*, 16(4), pp.406–413.

Yu, D.-A., Han, J. & Kim, B.-S., 2012. Stimulation of chondrogenic differentiation of mesenchymal stem cells. *International journal of stem cells*, 5(1), pp.16–22.

Yuan, J. et al., 2009. Histone H3-K56 acetylation is important for genomic stability in mammals. *Cell cycle (Georgetown, Tex.)*, 8(11), pp.1747–53.

Zeggini, E. et al., 2012. Identification of new susceptibility loci for osteoarthritis (arcOGEN): a genome-wide association study. *Lancet*, 380(9844), pp.815–23.

Zhang, J. et al., 2015. Loss of lysyl oxidase-like 3 causes cleft palate and spinal deformity in mice. *Human molecular genetics*, 24(21), pp.6174–85.

Zhang, L. et al., 2010. Chondrogenic differentiation of human mesenchymal stem cells: a comparison between micromass and pellet culture systems. *Biotechnology Letters*, 32(9), pp.1339–1346.

Zhang, Q. et al., 2015. SOX9 is a regulator of ADAMTSs-induced cartilage degeneration at the early stage of human osteoarthritis. *Osteoarthritis and Cartilage*, 23(12), pp.2259–2268.

Zhang, Y. et al., 2013. Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. *Nature*, 504(7479), pp.306–310.

Zhao, H. & Dean, A., 2004. An insulator blocks spreading of histone acetylation and interferes with RNA polymerase II transfer between an enhancer and gene. *Nucleic Acids Research*, 32(16), pp.4903–4914.

Zhao, S. et al., 2014. Comparison of RNA-Seq and microarray in transcriptome profiling of activated T cells. *PloS one*, 9(1), p.e78644.

Zhou, H.Y. et al., 2014. A Sox2 distal enhancer cluster regulates embryonic stem cell differentiation potential. *Genes & development*, 28(24), pp.2699–711.

Zhou, X. et al., 2014. Chondrocytes Transdifferentiate into Osteoblasts in Endochondral Bone during Development, Postnatal Growth and Fracture Healing in Mice M. L. Warman, ed. *PLoS Genetics*, 10(12), p.e1004820.

Zilberman, D. et al., 2008. Histone H2A.Z and DNA methylation are mutually antagonistic chromatin marks. *Nature*, 456(4), pp.125–129.

Filename:	thesis_corrections.docx

Directory: /Users/Kat/Library/Containers/com.microsoft.Word/Data/Documents

Template: /Users/Kat/Library/Group Containers/UBF8T346G9.Office/User Content.locali

Title:

Subject:

Author:	Microsoft Office User

Keywords:

Comments:

Creation Date: 13/03/2018 17:23:00

Change Number: 2

Last Saved On: 13/03/2018 17:23:00

Last Saved By: Microsoft Office User

Total Editing Time:	0 Minutes

Last Printed On: 13/03/2018 17:23:00

As of Last Complete Printing

Number of Pages: 317

Number of Words: 68,728 (approx.)

Number of Characters: 391,755 (approx.)