The University Of Sheffield.

Computational analysis of enhancer deregulation in Multiple Myeloma

By:

Jaime Álvarez Benayas

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

The University of Sheffield
Faculty of Science
Department of Molecular Biology & Biotechnology

September 30, 2019

Esta tesis está dedicada a mi familia y amigos

por todo su apoyo y comprensión y a

todos los pacientes de cáncer

y sus familias…

♥

Table of contents

Acknowledgements

This “trip” has been a long one and I am grateful to so many people, please apologies in advance if I forget someone. Primero de todo agradecer a mis padres por todo lo que me han enseñado y todos los esfuerzos e inversión que han hecho en mí (particularmente en el ámbito académico) que me han permitido poder optar y aprovechar esta oportunidad. De manera general, a mi familia y a mis amigos cercanos por entender mi motivación para realizar este proyecto, todas las concesiones, entendimiento, comprensión, apoyo y horas al teléfono: decir que esto no hubiera sido posible sin vosotros es quedarme corto. Más difícil que hacer este proyecto ha sido estar lejos de vosotros.

A mi madre por ser mi amiga más incondicional y la persona más fuerte que conozco y, a pesar de lo difícil que ha sido esto para ti, por entender y apoyarme. Por mostrarme hasta dónde es capaz de llegar el corazón humano incluso estando en el límite físico y que "el ser humano es capaz de hacer mucho más a veces por otros que por uno mismo". Negaré que lo he escrito, pero en los momentos duros he tirado de la memoria de tu ejemplo. A mi padre por todo el apoyo, la paciencia en enseñarme a tener serenidad y confianza para afrontar los retos académicos y los momentos en Madrid que me han dado la suficiente energía para este largo camino. A mi hermano por todas esas idas y venidas al aeropuerto y por cuidar de mamá en mi ausencia. A mis abuelos, más recientemente a mi abuela porque los momentos (especialmente los últimos) que pasé contigo, me dieron la motivación, las razones y el empuje para hacer esto.

A José por ayudarme a lanzarme y afrontar esta aventura cuando todavía no había empezado, por ser un amigo de esos que ya no quedan, por llevar 10 años siempre ahí incondicionalmente en las buenas y en las malas, sin importar la distancia ni las circunstancias, por tantos momentos pegados al teléfono y en Madrid que tanto oxígeno me han dado. Mucho tiempo antes de empezar este proyecto me dijiste que estarías "contra viento y marea" y así ha sido siempre.

A Mauro por ser un amigo incondicional, por tanta ayuda prestada, buenos momentos, apoyo y charlas tan interesantes tanto en Sheffield como en Madrid.

A Estefy porque conocer a un doctorando en esta fase de su proyecto y a distancia es un reto en sí, por toda la paciencia, comprensión, entendimiento, cariño, apoyo, ánimo y concesiones. También a toda la familia por ser tan comprehensivos conmigo y poner a mi disposición tantos medios para poder completar este trabajo durante mi estancia.

To Meshari, for being my brother in Sheffield and having my back regardless of what happened, for his unconditional support, like having a family in the place where you are staying.

A Joa porque no podría calcular cuánto ayudaste y me has inspirado con ejemplo a intentar ser mejor persona.

Ian, if I had to explain in detail, how much you have contributed in helping me to develop this project and in general, I could very easily write another thesis and would run out of positive words… Thank you for looking after me when stress and home-sickness was getting to me, all the insights and guidance, picking me up in hard times, offering unconditional help when the going was tough, all the concessions, believing in me sometimes even more than myself and generally helping me throughout the project. Your passion for what you do is very contagious and working under your supervision has been one of the greatest honours, privileges and best experiences in my whole life. As you told me before I began my PhD, for 4 years I have always found your door open. This is probably not the first time I tell you but I cannot imagine a better supervisor, leader and role model: please do not change.

To all the lab mates: Jacob, Vladimir, Justin, Magda, Cristina, Ivo, Nadia, Joe, Matt Howes, Matt Parker and all the summer students I have had the privilege to work with. Thank you to all for all the good moments, the support, the insights and for sharing this experience. Particularly Jacob for all the insights and reminders of hand-ins which saved my life and all the support and insights during my PhD. Also, to Umberto for all the insights regarding the PhD and life in general.

Our collaborators Valentina, Tassos, Nikos, Alexia and the rest of the crew who are great professionals, for trusting in us, helping and always encouraging work of the best quality.

To all the patients who are real heroes for trusting in us sharing their data, without your data, there wouldn't be any work to do.

Álvaro and Maria José por prestarme estancia en su casa en la conferencia de Barcelona y por hacerme sentir como en casa.

Alejandro, for receiving me and easing tremendously my arrival to the UK mentally and logistically, for making his house my house on all my trips to London: whether work, leisure or flying and making me feel like at home in another country.

A Nuria por haberme ayudado a salir adelante en tantos malos tragos.

A Luz por muchas cosas, principalmente por hacerme sentir como en casa en un país lejano, pero principalmente porque como te dije tu principal misión en el doctorado era asegurarte de que yo me mantuviera cuerdo y contra todo pronóstico, lo conseguiste :)

A capitán Sergio y Dani por muchos buenos momentos que me han hecho sentir como en casa y por ayudarme cuando necesitaba una mano, compadres.

To Jocelyn for all the support and understanding, the concessions, the motivation, having my back when I was so far away from home, I cannot imagine standing here today without you.

To Alex for facilitating my PhD studies even before I got to Sheffield through library visits and continuous encouragement during all the PhD.

Nuria, muchas gracias por toda la ayuda psicológica desinteresada que me dio mucho aire y los consejos que me fueron muy útiles durante la tesis.

A María del Toro por todas las veces que me ha ayudado con temas técnicos del doctorado y el apoyo.

To the Sheffield crew: Mirjam, Apurbaa and the rest for all the good times, interesting conversations that made me feel like at home.

To Lianne one of the first people in Sheffield who made me feel connected and "at home", always with warmth and good vibe.

To Abbas for many things to write here, particularly for helping me with this adventure before it began and for the support.

To Ghita for many things, for setting the bar very high and regarding this thesis for sending me that machine learning course which opened my mind to the potential behind it starting all this.

To Prof. Yaser S. Abu-Mostafa for being a passionate and awesome teacher, opening my mind to the idea of machine learning and to the basic idea that with time and effort everything is possible.

To Tereza because seeing is believing and when you have a target to aim, everything becomes much easier.

To Mathieu because you are such an inspiration at such a young age, I can't imagine anyone meeting you and not realizing limits are just mental.

To Montsina for offering so much help and support in this thesis.

To the best teacher I ever had: Scott Massara. You planted the seed for passion and enthusiasm for science and learning at a very early age and showed me that with effort, anything is possible.

To all the people who have ever asked me what I do and when I replied they gave me that "look" or words wishing me the greatest of successes.

Abstract

Gene regulation is a complex process, which dictates how the body reacts to different situations through gene expression. Enhancers are sequences of a few hundred base pairs, involved in the regulation of transcription. This work focuses on enhancer activity changes during the cancer multiple myeloma, an incurable malignancy of the plasma cells: B-cells, which are long-lived, produce immunoglobin and provide protection against antigens that activated them. In this thesis, data from different assays is combined for multiple myeloma and plasma cells samples to determine cancer-specific and subgroup-specific enhancers and these are correlated with target genes based on activity of both actors. I find hundreds of enhancers linked to expression of nearby genes, with a large fraction of these being specific to MAF translocated tumors. Changes in *de-novo* open chromatin distant to the promoter of a gene are more predictive of gene expression than opening of the promoter. Also, combination of chromatin accessibility data and gene expression data is better at distinguishing cancer subtypes than either alone. Many of the regulated genes are known to be important in multiple myeloma, and this study provides a potential mechanism for their deregulation. In addition, I identify novel genes of interest. These enhancers show motif enrichment for transcription factors expressed in plasma cells as opposed to cancer specific factors. In particular, a large, MAF binding, open chromatin region is identified that correlates with the expression of the oncogene CCND2, and distinguishes mutually exclusive sets of samples expressing CCND2 or CCND1, going some way to explaining the known CCND dichotomy. This work lays the foundations of *in vivo* and *de novo* Myeloma vs. PC and MM subgroup specific enhancer – promoter interactions essential for the oncogenic state. Given that currently Myeloma is an incurable cancer, this should be of significant relevance for diagnosis, prognosis and treatment.

List of Figures

List of Tables

Acronyms and Abbreviations

Acronym	Definition
3C	Chromosome Conformation Capture
AID	Activation-Induced Cytidine Deaminase
AML	Acute Myeloid Leukaemia
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	Adenosine triphosphate
BET	Bromodomain and Extraterminal Domain
BM	Bone Marrow
BMP	Bone Morphogenetic Protein
bp	Base pair
BP	Biological Process
BTK	Bruton's Tyrosine Kinase
CAGE	Cap Analysis of Gene Expression
CCCTC	Transcriptional repressor CTCF also known as 11-zinc finger protein
ChIP-seq	Chromatin Immunoprecipitation sequencing
CL	Cell Line
CLL	Chronic Lymphocytic Leukemia
CNV	Copy Number Variations
CRISPRi	CRISPR interference
DA	Differentially Accessible
DAMM	Differentially Accessible MM
DAPI (staining)	4’,6-diamidino-2-phenylindole fluorescent stain
DASMM	Differentially Accessible Subgroup MM
DCIS	Ductal Carcinoma In Situ
DE	Differentially expressed
DEMM	Differentially Expressed MM
DESMM	Differentially Expressed Subgroup MM
DNase	Deoxyribonuclease
eQTL	Expression quantitative trait loci
eRNA	Enhancer RNA
ESC	Embryonic Stem Cells
FACS	Fluorescence Activated Cell Sorting
FAIRE-seq	Formaldehyde-Assisted Isolation of Regulatory Elements sequencing
FDR	False Discovery Rate
GO	Gene Ontology
GRO-seq	Global Nuclear Run-On sequencing
H3ac	Acetylation of histone H3
H3K27ac	Acetylation of histone H3 lysine 27
H3K4me1	Histone H3 lysine 4 monomethylation
H3K4me3	Histone H3 lysine 4 trimethylation
HD	Hyperdiploid
HDACi	Histone Deacetylase inhibitors
IDC	Invasive Ductal Carcinoma

IgH	Immunoglobulin Heavy Chain
Kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
KIRP	Kidney Renal Papillary
LAD	Lamin-Associated Domain
LF	Latent Factor
lnc-RNA	Long non-coding RNA
log_2foldchange	Log (base 2) fold change
logFC	Log Fold Change
LOH	Loss of Heterozygosity
LRT	Log Ratio Test
Mb	Megabase
mESC	Mouse Embryonic Stem Cells
MF	Molecular Function
MGUS	Monoclonal Gammopathy of Uncertain Significance
miRNA	Micro RNA
MM	Multiple Myeloma
MMPC	Multiple Myeloma and Plasma Cell
MMR	DNA Mismatch Repair
MNC	Mononuclear Cells
MOFA	Multi Omics Factor Analysis
mRNA	Messenger RNA
mSTARR-seq	Methylation Self-Transcribing Active Regulatory Region sequencing
ncRNA	Non-coding RNA
ND	Normal Donor
OE	Overexpressed
OEMM	Over Expressed MM
OESMM	Over Expressed Subgroup MM
PC	Plasma Cell (used interchangeably with ND)
PCA	Principal Component Analysis
PCL	Plasma Cell Leukemia
PCR	Polymerase Chain Reaction
PIC	Pre-Initiation Complex
Pol II	RNA Polymerase II
PR	Perichromatic Region
QC	Quality control
QTL	Quantitative Trait Locus
rLog	Regularized Log
RNA-seq	RNA sequencing
ROSE	Ranking Of Super Enhancer
RR	Regulatory Region
RSS	Recombination Signal Sequences
SCCOT	Squamous Cell Carcinoma of the Oral Tongue
SHM	Somatic Hypermutation
SMM	Subgroup MM
snoRNA	Small nucleolar RNA

SNP	Single Nucleotide Polymorphism
SNV	Single Nucleotide Variant
SP	Larger Speckles
STARR	Self-Transcribing Active Regulatory Region
TAD	Topologically Associating Domain
T-ALL	T-cell acute lymphoblastic leukemia
TBP	TATA-binding protein
TCR	T Cell Receptor
TF	Transcription Factor
TPM	Transcripts per Million
TSS	Transcription Start Site
UTR	Untranslated Region
VDJ	V-D-J block recombination (PC formation)
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing
ZRS	Zone of polarizing activity Regulatory Sequence

1. Chapter 1

1.1. Chapter 1: Introduction

Gene regulation is a complex multistep process, which dictates how the body reacts to intrinsic and extrinsic signals through gene expression. Variation in gene expression from one individual to another ultimately gives rise to phenotypic variability, which in some cases can mean acquiring of complex diseases such as cancer. Understanding this mechanism is therefore critical for the characterization and development of targeted treatment of such complex diseases.

Gene expression can be subdivided into its most fundamental steps: DNA is transcribed into RNA, the RNA is processed, exported out of the nucleus into the cytoplasm and RNA is translated into protein. Within these processes, there are various ways in which the resulting protein can be altered, for example, at the post transcriptional level, there is capping, splicing and addition of a Poly(A) tail, among other events performed on the transcript. The way DNA is transcribed by RNA polymerases and how different factors affect this process has been heavily studied and is still being fully uncovered.

Enhancers are sequences of a few hundred base pairs that play an important role in the regulation of transcription for the majority of genes by combining their activity with insulators or silencers (Nizovtseva et al., 2017). When activated through the binding of TFs they aid in the binding of the polymerase to the promoter, thus augmenting transcription.

This work focuses on enhancer activity changes during disease since changes in regulatory regions are thought to be vital towards development of common diseases in some cases such as prostate cancer or Chron disease (summarized in Manolio et al., 2008). Despite the fact that the changes in coding genes affect protein structure and function, enhancers play a key role in gene regulation.

1.2. Properties of enhancers

Enhancers can be recognized by a series of properties related to their functional effects (Shlyueva et al., 2014). Below, each of these properties is briefly reviewed.

1.2.1. High chromatin accessibility

Nucleosomes are a collection of eight histone proteins with DNA coiled around them. It has been noted that they affect the binding of TFs to enhancer sequences since the TFs require

DNA cleared from nucleosomes to bind the DNA motif (reviewed in Shlyueva et al., 2014). Regulatory elements such as promoters, enhancers or silencers (capable of binding transcriptional regulators to repress transcription) therefore, when active tend to lie in open chromatin regions. Chromatin accessibility as measured by different assays (Prediction of enhancers, section 1.4) is not an on/off switch, but rather a continuum, although categorization into open and close chromatin is typical based on signal thresholds (Boyle et al., 2008). It is also important to note that while active enhancers lie in open chromatin, not all open chromatin regions correspond to enhancers.

Chromatin is made accessible at enhancers by means of a class of TFs known as pioneer factors, which can access chromatin while other TFs cannot. These bind to closed enhancers, displacing nucleosomes, and engage either chromatin remodeling complexes or histone modifying enzymes, further decompacting the chromatin (Clapier and Cairns, 2009). For example, Jacobs *et al.*, found that epithelial enhancers were bound by the TF GNRH1, which lead to an opening of the chromatin and an activation of the enhancers (Jacobs et al., 2018). It remains to be elucidated whether after binding of pioneer factors to the enhancers, polymerase recruitment at the enhancer precedes opening of chromatin or not (Zaret and Mango, 2016).

1.2.2. TF binding and sequence conservation

Enhancers can be cell type specific or shared among different cell types (Visel et al., 2010) and different TFs can activate enhancers in different tissues (Yáñez-cuna et al., 2012). Although enhancers bind TFs, TF binding is not always indicative of an enhancer (Li et al., 2008). The relationship between enhancers and TF binding is many to many, thus, in many cases, enhancers can be recognised by the presence of clusters of TF binding sites.

More complexity is added in terms of TF binding to enhancers by combinatorial occupancy: TFs bind to enhancers containing multiple binding sites for other TFs. Sometimes binding motifs can overlap, so only one TF in the motif overlapping domain will bind at a particular point in time (Spitz and Furlong, 2012). Additionally, TFs required for transcription can be cell-type specific (Spitz and Furlong, 2012) and occupy an enhancer on a time constraint, for example, during different phases of differentiation, for instance, occupancy of the TCF3/E2A TF is dynamic (Lin et al., 2011). The TF binding affinity is further affected by additional binding of cofactors to TFs, for example, the *Saccharomyces cerevisiae* TF CBF1 requires the binding of the cofactors MET4 and MET28 to bind the motif (Siggers et al., 2011).

We do not fully understand all the mechanisms of how TFs aid in transcription, it has been proposed that one of them is in the aiding of chromatin looping (see “DNA looping”) between enhancers and promoters (Deng et al., 2012), but it is likely that instead DNA looping promotes TF-dependent transcription. Finally, it is common but not necessary (Blow et al., 2011) that enhancers conserve their sequence across species (Kheradpour et al., 2007; Visel et al., 2010).

1.2.3. Histone modifications and general DNA methylation

Histones can contain certain post-translational modifications which can affect damage and repair processes, chromatin packaging and, importantly for this work, transcription. These modifications include gain and loss of acetylation, methylation, phosphorylation, deamination and ubiquitination. Nucleosomes surrounding active enhancers regions tend to contain certain modifications such as acetylation of histone H3 lysine 27 (H3K27ac) and histone H3 lysine 4 monomethylation (H3K4me1) (Bulger and Groudine, 2011). More recently it has been determined that H3K4me1 is not a requirement for enhancer driven transcription in mESC (Dorighi et al., 2017) and in *Drosophila melanogaster* (Rickels et al., 2017). Additional findings determine that highly active enhancers are marked by H3K4me3 and not H3K4me1 in flies and mESCs (Henriques et al., 2018).

In general, it was thought that DNA methylation at enhancers was associated with inactivation of the enhancer and repression of target gene expression, while hypo methylation at enhancer regions is associated with enhancer activity (Long et al., 2017; Qu et al., 2017; Wiench et al., 2011). In a study on cancer, it was determined that there is an inverse relationship between methylation and chromatin accessibility at enhancers (Corces et al., 2018). Cases have been reported where, during differentiation, a rapid nucleosome gain at enhancers is followed by DNA methylation which is thought to make the repressed enhancer state stable (You et al., 2011). More recently, through mSTARR-seq (Methylation Self-Transcribing Active Regulatory Region sequencing), it was determined that despite the fact that methylation regulated enhancers tend to have a negative correlation between methylation and target gene expression, only the minority of regulatory regions with enhancer activity (15% of regions found) have methylation-dependent gene regulatory action (Lea et al., 2017). Methylation at the promoter is also an important factor to consider in gene expression, recently it has been determined that methylation at the promoter is not sufficient in general to repress gene transcription (Ford et al., 2017), although further statistical analysis of the data used has proven otherwise (Korthauer and Irizarry, 2018).

1.2.4. Distance and orientation between enhancers and promoters

Studies tend to associate enhancer elements to the closest promoter, however, it has been found that this is not always the case (Fu et al., 2018). In fact, interactions involving DNA looping (see DNA looping, section 1.5.4) have only 22% of the potential enhancers targeting the closest active gene (Sanyal et al., 2012). An enhancer sequence influencing a promoter could be very far away in linear DNA genomic distance. For example, the zone of polarizing activity Regulatory Sequence (ZRS) region regulates *SHH* expression, but lies within an intron of another gene 1 Mb away (Lettice et al., 2003).

Despite this, the majority of enhancer – promoter interactions tend to lie within a certain range. A recent paper, using chromatin accessibility assays, Quantitative Trait Locus (QTL) and a Bayesian approach between pairs of elements separated by 500Kb, established that more than 60% of causal interactions (for which there are enrichment in enhancer – promoter interactions), occur within 20Kb (Kumasaka et al., 2018). As noted by the authors of the study, the conclusions that can be extracted must be taken with caution, since enhancers not containing a genetic variant thought to alter their function are not included. In another study, from all promoters interacting with non-promoter regions, about 90% of these interactions occurred within 1Mb (Javierre et al., 2016).

Also, a DNaseI hypersensitive site assay in conjunction with annotation estimated that 95% of DNaseI I sites (of which enhancers will comprise a fraction of) are further than 2.5 kb from a Transcription Start Site and exhibit cell-specific patterns of activity (Thurman et al., 2012). Enhancers can be oriented upstream or downstream from the promoter and expand across other genes (Gorkin et al., 2014). There is also some evidence of inter-chromosomal interactions (Lomvardas et al., 2006).

1.2.5. Extension, autonomy and combinatorial effect

In general, enhancers are regarded as autonomous and maintain their influence independently of the surrounding sequences (except for insulators such as CCCTC-binding factors, CTCF sites, explained later). However, it has been shown that there is not always a clear-cut boundary to the region that is essential for its enhancer action (known as a core minimal enhancer) on a gene (Spitz and Furlong, 2012). This minimal sequence can require additional flanking sequences for greater range of adaptability to different environmental conditions (Ludwig et al., 2011).

Different enhancers can have complex interactions on the target promoter, for example having an additive effect on the transcription levels or acting as overlapping modules (reviewed in

Spitz and Furlong, 2012). A particular case are “shadow” enhancers: DNA sequences which are thought to perform redundant functions to another enhancer for the same genes. They are especially important during development for the robustness they confer for example against enhancer deletions (Cannavò et al., 2016).

1.2.6. Enhancer transcription

Recently it has been discovered that RNA polymerase II is found at some enhancer regions and transcribes them into relatively unstable non-coding RNAs (ncRNAs), enhancer RNA (eRNA). These eRNAs are thought to contribute in the transcription of the target gene (Ding et al., 2018). Several mechanisms, are proposed, one of which is enhancer-promoter looping formation (Liiu et al., 2018) as it is explained in further detail in DNA looping, section 1.5.4.

1.2.7. Super enhancers

Super enhancers are a theoretical concept describing a particularly set of enhancers of large genomic extent and containing clusters of enhancers (reviewed in Sengupta et al., 2017). The most important characteristics are that enhancers within super enhancers contain an order of magnitude higher enrichment of H3K27ac, H3K4me1, Mediator complex (MED1), BRD4 and cell-type specific TFs compared with normal enhancers. Super enhancers also result in a target gene expression much higher than regular enhancers target gene expression. The target gene expression is more sensitive to any disruption of the TFs or Mediator proteins associated with super enhancers compared with disruption in regular enhancers TFs or Mediators. The target genes of super enhancers tend to be associated with cell-identity processes.

1.3. Gene Transcription

During gene transcription (Figure 1-1), briefly, first binding of the general TF TFIID (through the TATA-binding protein: TBP) to the TATA box containing promoter occurs (Sainsbury et al., 2015). Then, assembly of the necessary components (pre-initiation complex: PIC) in multiple phases occurs by the addition of other general TFs such as TFIIA or TFIIB and Polymerase II (Sainsbury et al., 2015). After creating a DNA opening between both strands, initiation of transcription begins and after transcribing between 20 and 120 nucleotides from the Transcription Start Site (TSS), the polymerase pauses. To allow the release of the polymerase, the action of positive elongation factors are required into productive elongation. The main steps in transcription are outlined in Figure 1-1.

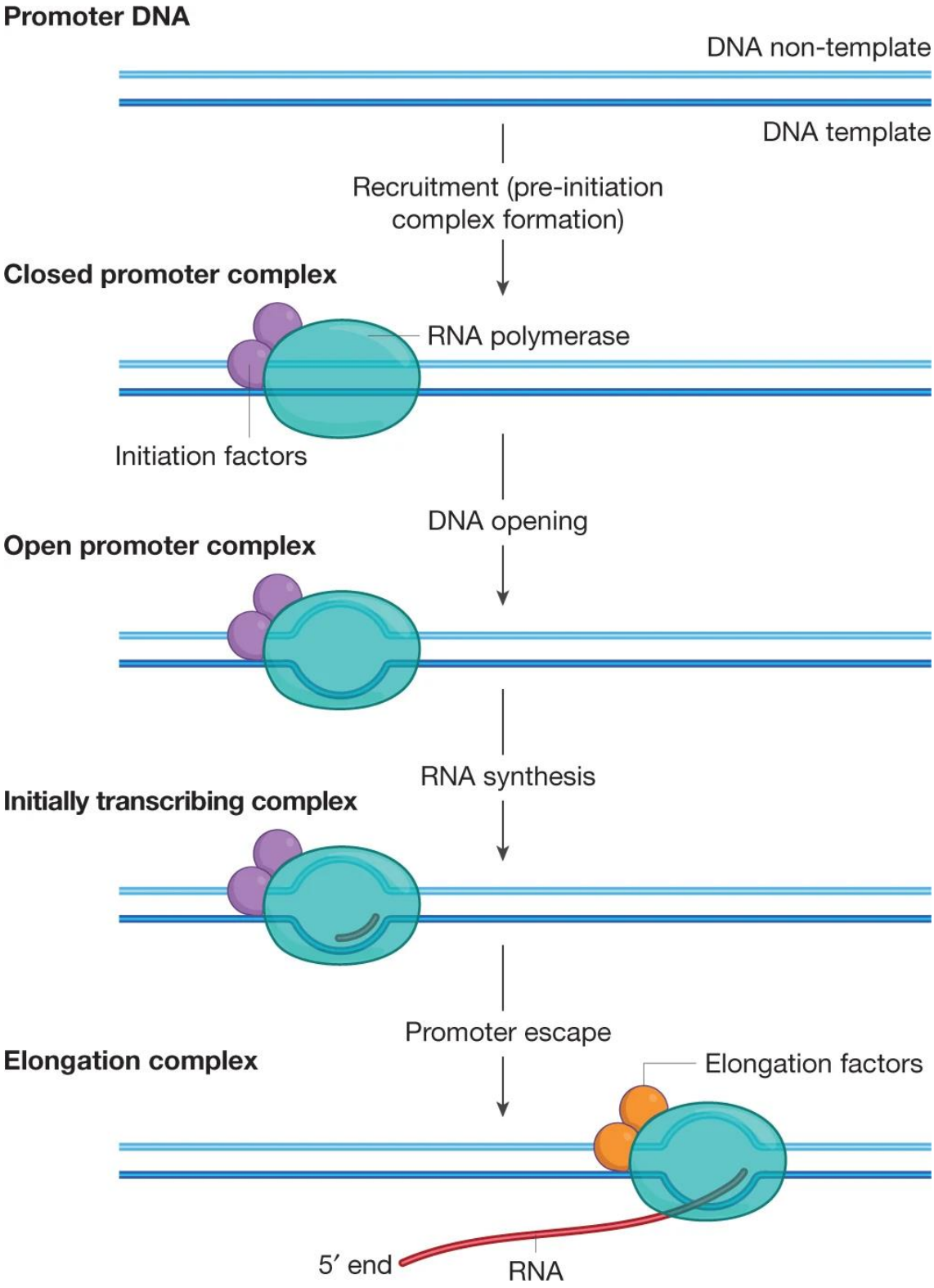


Figure 1-1: Simplified schematic of the transcription process. From (Cramer, 2019).

General TFs (purple bubbles) recognize and bind promoter DNA. The PIC is formed (blue bubble) including these general TFs and Polymerase II enzyme. DNA opening occurs in the transcription bubble changing the closed promoter to the open promoter complex state. The DNA template strand (dark blue line) is then transcribed into RNA

(red) by the Polymerase II for 20 and 120 nucleotides, after which, it pauses. Finally, positive elongation factors (orange bubbles) bind to the paused Polymerase (elongation complex) to release the Polymerase from the paused state and enable productive elongation.

The process of gene transcription and how it is affected by the 3D organization of the genome is under heavy study, at present, there are various theories about how transcription is organised. The first is the concept of "transcription factory" which refers to the idea that components forming the transcription machinery, are found in high concentrations in particular areas within the nucleus where genes are actively transcribed (Andrew J Fritz et al., 2019; Papantonis and Cook, 2013). One of the first ways in which this was observed was with HeLa cells, where nascent transcripts were allowed to extend transcripts in biotin and then biotin-RNA immuno-labeled particles binding to them were observed by microscopy, it is thought that 90% or more of transcripts are produced in transcription factories (Iborra et al., 1996). Multiple genes, in cases thought to come from different chromosomes, can be actively transcribed in each of these regions (Osborne et al., 2004). There is evidence, which shows that if transcription factories exist, they may not be cell type specific, since the association between the same active domains are very overlapping in different cell contexts (Bickmore, 2013).

Another area under study regarding this work is the stepwise assembly of components that takes place at the promoter through the physical contact of the actors, enabled by the particular 3D organization of the genome. Enhancers regulate transcription in various ways, for example, encouraging the assembly of RNA polymerase by binding proteins called activators, which directly or through recruitment of other complexes modify histones, remodel chromatin clearing the promoter sequence from nucleosomes or, in the case of long-range gene-specific enhancers, have cofactors called mediator proteins binding to them (reviewed in García-González et al., 2016). This step enables the correct assembly of the TFs required for transcription on the promoter (collectively known as general TFs). Each TF has different roles, for example TFIID, recognizes the specific targeted promoter (Sainsbury et al., 2015) and mediator complexes help to stabilize chromatin for TFs to bind to DNA (Eyboulet et al., 2015). Once this is completed, RNA polymerase combines with the general TFs to form the pre-initiation complex (PIC) and transcription can begin.

eRNA transcripts can also regulate gene transcription through facilitation of the RNA polymerase going from pausing to productive elongation when transcribing a gene. The

mechanism is thought to occur through eRNA independent chromatin looping (explained in DNA looping, section 1.5.4) enabling transcribed eRNA to inactivate negative elongation factors limiting the polymerase (Schaukowitch et al., 2014). It has also been found that eRNA can bind to co-activator CREBBP and EP300 (the latter transcribed by the *P300* gene) which results in histone acetylation of active enhancers and corresponding gene expression of nearby genes (Bose et al., 2017). While eRNA production often correlates with enhancer activity (Henriques et al., 2018), this is not always the case (Mikhaylichenko et al., 2018).

1.4. Prediction of enhancers

Various methods have been implemented to determine the location of putative enhancers (or regions having the potential of being enhancers) in the genome, these can be broadly divided into indirect methods and direct methods. Indirect methods are based on exploiting different characteristics of enhancers explained in Properties of enhancers, section 1.2, such as nucleosome-depleted accessible chromatin, enhancer histone modifications or clusters of TF binding sites among others. Direct methods include assays that give information about enhancer function, for example: enhancer bashing (Venken and Bellen, 2014), self-transcribing active regulatory region (STARR) sequencing (Arnold et al., 2013) or CAS9-mediated *in situ* saturating mutagenesis (Canver et al., 2015).

It is possible to elucidate enhancers just by observing regions enriched for TF binding sites. This method however does not provide information about whether the binding site is occupied and the enhancer is active. Also, since enhancers are cell-type specific but the DNA sequence is not, this method does not yield tissue specific enhancers, which are activated by cell-specific TFs.

One property of active enhancers used is the fact that they lie in nucleosome depleted open chromatin areas. This is the theory behind chromatin accessibility assays such as Deoxyribonuclease I: DNase I (Thurman et al., 2012) or Assay for Transposase-Accessible Chromatin using sequencing: ATAC-seq (Buenrostro et al., 2013) which work by adding an enzyme to the DNA that cleaves at sites of high chromatin accessibility (free of proteins such as nucleosomes). The details of ATAC-seq can be viewed in Figure 1-2. The fragments obtained from this process are then sequenced and the fragment ends reflecting the cleaved sites are assumed to lie in accessible chromatin. ATAC-seq tends to be the preferred method currently since it requires lower input DNA and the complexity of the experimental procedure is low.

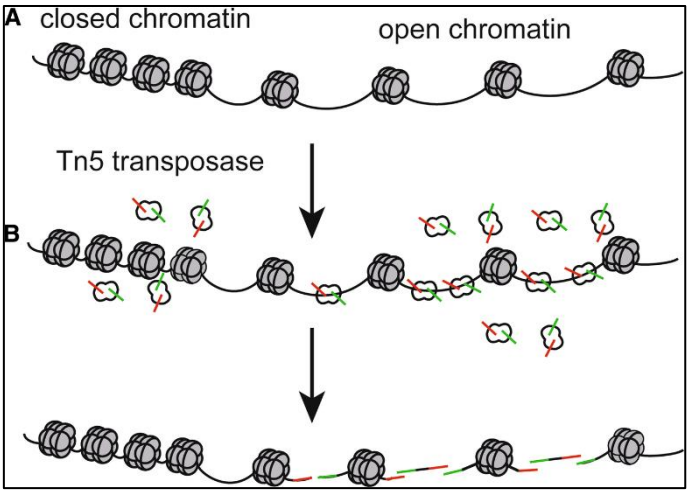


Figure 1-2: ATAC-seq assay. From (Sun et al., 2019)

ATAC-seq uses the Tn5 transposase, which cleaves the DNA at accessible, nucleosome depleted chromatin areas. The transposase tags these regions by inserting sequencing adapters (in red and green). The tagged fragments can then be purified, amplified using PCR and sequenced to obtain the location of the regions bioinformatically (not shown).

A third assay that separates the genome into DNA regions bounded by protein and nucleosome depleted is called Formaldehyde-Assisted Isolation of Regulatory Elements (FAIRE-seq: Giresi et al., 2007). This assay works by cross-linking cells, which fixes DNA to proteins such as nucleosomes, the cross-linked DNA, is then divided into smaller pieces by sonication. Finally, it is separated by phenol-chloroform extraction, which creates two layers and separates the DNA based on its properties: an organic layer, which enriches for DNA containing nucleosomes and an aqueous one with DNA free of nucleosomes. The DNA contained in the aqueous layer is then purified and sequenced and nucleosome-depleted, open chromatin regions are obtained.

Chromatin accessibility methods have been extensively used in enhancer identification, particularly in cancer development (Davie et al., 2015). It is important to take into account that chromatin accessibility signal can be subject to bias due to factors such as copy number amplifications of regions for example and thus it is necessary to factor it in if the information is available whenever possible (Corces et al., 2018).

Chromatin accessibility assays can also help in determining enhancers indirectly through a computational technique called TF footprinting (Rendeiro et al., 2016). TF footprinting can determine binding sites in the genome for a range of TFs. Observing the chromatin accessibility information on and around these sites, it can conclude that the corresponding TF is bound to the DNA. This is reflected by a closed chromatin region on the sites (protecting the DNA from

cleavage at these regions) while having the surrounding regions in open chromatin. This approach, also gives an idea of how the enhancers are being regulated through TFs. Additionally, through TF footprinting of chromatin accessibility data, it is possible to relate significant binding strength and the ability to open chromatin around a particular TF with the gene expression of that TF (Corces et al., 2018), marking potential enhancers. An important caveat of this approach is that the motifs associated to TF binding can sometimes be very short and low-complexity which can result in many false positive binding sites.

Chromatin Immunoprecipitation (ChIP) experiments target a protein of interest such as a TF thought to be collaborating in transcription by direct enhancer binding or a cofactor like EP300 collaborating in enhancer – promoter interactions (Visel et al., 2010). Binding of general cofactors, such as EP300, have the advantage that their involvement in transcription is one step closer towards producing the transcript than just individual binding of TFs and therefore can render more accurate predictions (Visel et al., 2010).

An experiment targeting an individual protein provides the areas of DNA where the protein is bound, by combining multiple assays targeting different TFs, clusters of different TFs binding sites become a proxy for enhancer regions (Nizovtseva et al., 2017). Other targets of ChIP can be RNA polymerase or a histone modification characteristic of enhancers.

Histone modifications also serve as a means of predicting enhancers, as explained earlier, simply targeting H3K27ac through ChIP-seq can yield enhancer regions (Zhu et al., 2013). A more complex histone code can be used to segment the genome more precisely into its constituent functional units: genic enhancers, enhancers, promoters, TSS, exons, repressed chromatin, etc… For this, a combination of histone modifications including the enhancer marks mentioned and others such as trimethylated histone H3 at lysine 36 (H3K36me3) or histone H3 lysine 4 trimethylation (H3K4me3) can be used. These marks are then used in combination with bioinformatics algorithms such as chromatin segmentation through a multivariate hidden Markov model: ChromHMM (Ernst and Kellis, 2017) as it was done with the ENCODE Project (Hoffman et al., 2013). A functional label is produced for each DNA segment of a fixed length based on the signal for the different marks in that region and the nearby region's functional labels. It has to be noted that some histone modifications when found, point at a region of 1kb or more, which can make difficult pinpointing the precise location of an enhancer within that region (Heinig et al., 2015).

Additionally, various methods can be used to target eRNA and the underlying enhancer function. Global Nuclear Run-On sequencing (GRO-seq) (Lis, 2015) was an assay first designed

to capture nascent RNA transcripts in real time, preventing for example, missing the quantification of RNA transcripts by degradation. The protocol isolates nuclei in ice-cold temperature, brominated (tagged) nucleotides are incorporated and transcription is then resumed by exposing the cells to higher temperature with compounds preventing transcription re-initiation. Only transcripts with tagged nucleotides that began to be transcribed prior to isolation are selected through Immunoprecipitation and then sequenced. This protocol can also be used to capture eRNA nascent transcripts, which can be differentiated from RNA transcripts, by using genomic annotations.

Cap Analysis of Gene Expression (CAGE) (Kodzius et al., 2006) was also initially designed to obtain a snapshot of the 5' of nascent RNA transcripts but it is also used for eRNA mapping. The nascent transcripts 5' is capped after the production of about 25 nucleotides by enzymes and are tagged, extracted and sequenced.

Any combination of data from the described approaches can be used in the task of determining enhancers, for example by overlapping open chromatin regions obtained from ATAC-seq with H3K27ac enrichment areas to determine regulatory regions during erythro-megakaryopoiesis (Heuston et al., 2018) or to recapitulate *Drosophila melanogaster* dorso-ventral networks (Koenecke et al., 2016). ATAC-seq in combination with chromatin segmentation of ChIP-seq of histone modifications has also been used for this duty finding an enrichment of chromatin accessibility in regions marked as enhancer and promoters by chromatin segmentation (Corces et al., 2018). Adding to the use of chromatin state data, sequence conservation can also be included (Van Duijvenboden et al., 2015). Additionally, purely *in silico* approaches have also been used, for example, some studies have tried to accomplish the task solely through creating models of TF motif binding sites clusters but found that reinforcing the model with additional chromatin state data improved the accuracy of the predictions (Fang et al., 2016). Combining different sources of data has yielded some success (Erwin et al., 2014) but further studies have concluded that the predictions based on different combinations of data types and methods generate sets of putative enhancer regions which are not significantly coinciding (Kleftogiannis et al., 2015). Since it is currently challenging to establish a "ground truth" for enhancers, it is difficult to know if the low concordance in results across combinations of data types and methods could be due to a high rate of false positive calls or the fact that each prediction yields a particular subset of all the enhancers.

The mentioned approaches provide a way of obtaining candidate putative enhancer regions to test. One technique that allows for testing of enhancer function is the Self-transcribing active

regulatory region (STARR) sequencing (Arnold et al., 2013). STARR uses reporter assays; these involve shearing the DNA to obtain different fragments to be tested for enhancer functionality. The different fragments are cloned downstream of a minimal promoter such that, after transfecting the different constructs into cells of interest, *bona fide* enhancers will transcribe themselves. The RNA produced is then sequenced, analysed and linked back to the corresponding enhancer construct. A higher RNA count indicates the construct has a higher enhancer activity. Using STARR to survey enhancer candidates in complex genomes without prior narrowing down of the number of potential regions can be a challenging task in terms of generating high-coverage transfecting libraries (Muerdter et al., 2015). Other limitations of STARR include the fact that is not performed in an endogenous context (artificial constructs are created), the plasmid may not contain histone modifications regulating the enhancer (Muerdter et al., 2018) or the fact that specific enhancers are cell type specific and this assay may not yield that specificity.

Another method to test enhancer function is through enhancer bashing (Venken and Bellen, 2014) which creates artificial constructs of candidate enhancer regions near a neutral promoter. Since the position of the enhancer can influence the transcription, the enhancer regions tested should be placed on the same location.

CAS9-mediated in situ saturating mutagenesis (Canver et al., 2015) can also be used to test enhancer function by creating copies of the wild type cells and altering the genome, removing candidate enhancer regions and observing the results in terms of gene expression of the target gene.

Another aspect to take into account when inferring enhancer related knowledge from the available data is whether the data corresponds to multiple or single cells. Data coming from multiple cells has to be interpreted as an average of all the cells, this average may or may not be representative of the individual cells.

1.5. Nuclear organization of the cell

To understand how elements within the cell's DNA interact with each other it is important to understand the layers of packaging involved (Figure 1-3). Around two meters of DNA coiled around histones is packed into a space, which is only six micrometers in diameter, this gives an idea of the complexity of the mechanisms involved. With technologies allowing increasing resolution within the cell nucleus, various structures have been observed in a hierarchical

manner. Each of them also contributing in gene regulation. It must be noted that with recent high resolution Chromosome Conformation Capture approaches, it is beginning to emerge that the forces governing each layer of organization can work synergistically or antagonistically (Rada-Iglesias et al., 2018).

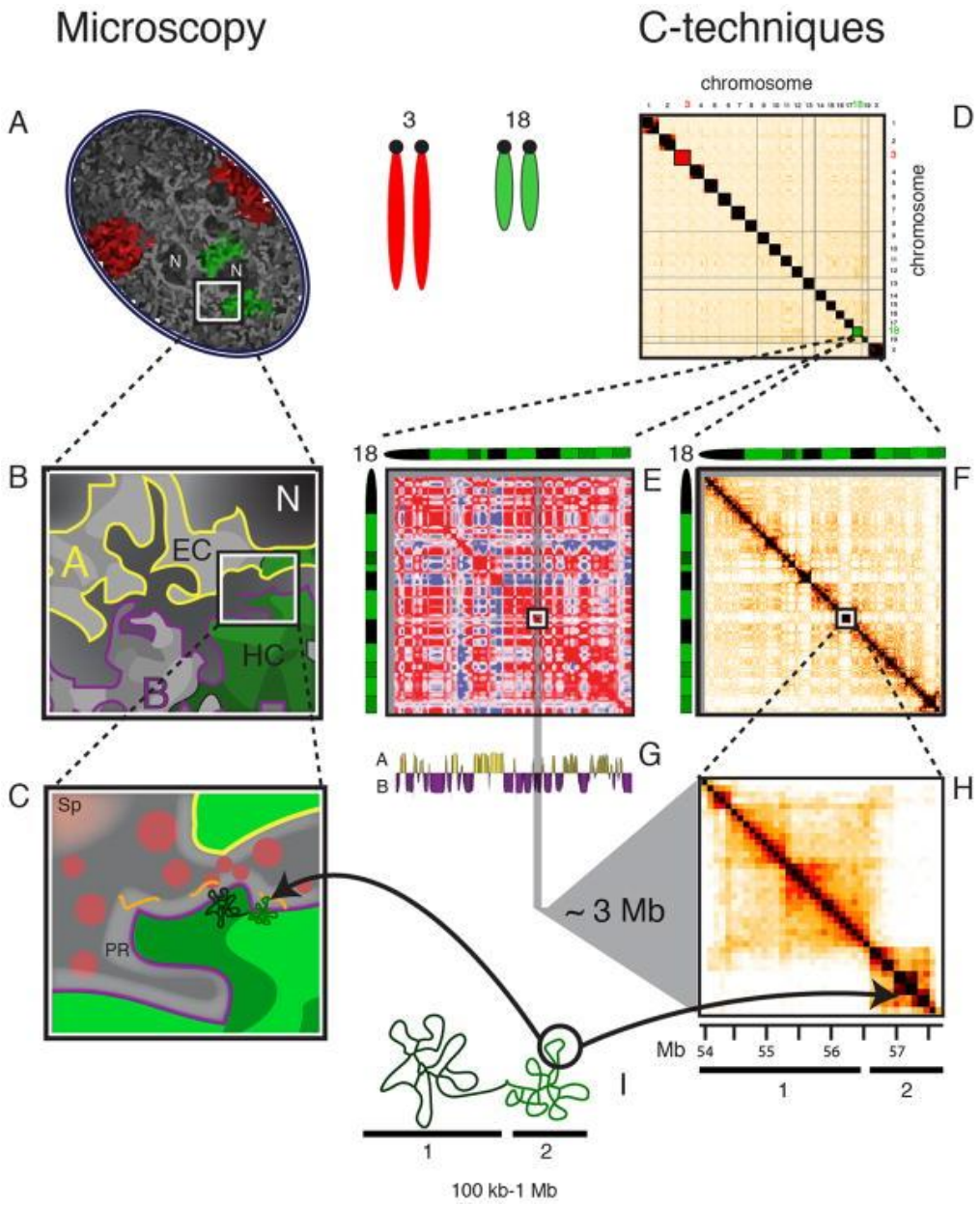


Figure 1-3: Large-Scale Nuclear Organization in Mammals. From (Gibcus and Dekker, 2013)

Hierarchical division of DNA in mammals (highest genomic space to lowest):

(A) Left: Chromosome territories shown for the mouse chromosome 3 (red) and 18 (green) with the nucleoli marked with N shown by DAPI staining. Dark intensity reflecting the level of chromatin condensation (darker is heterochromatin). Right: Comparative size of the chromosomes 3 and 18.

(B): Zoom in from (A) showing A/B compartments within Chromosome Territories, and in the nucleus in general, euchromatic chromatin is spatially separate from heterochromatin, which is inactive (HC).

(C): Zoom in from (B) showing chromosome domains. PR stands for the perichromatic region, which is the outer layer of the chromosome domains and is shown interacting with TFs (pink circles) and RNA (orange lines). SP stands for larger speckles and reside in areas of low chromatin content distant from the PR.

(D): Genome-wide contact heat-map matrix showing interaction density within and between chromosomes, as can be seen, the concentration of interactions occur within the same chromosomes (darker regions in the heat-map).

(E): Subdivision of a chromosome territory in D for chromosome 18 into a Pearson correlation coefficient heat-map. Starting by dividing the chromosome into bins of a certain length, a matrix of contact frequencies between each pair of bins is produced. From this matrix a Pearson correlation coefficient matrix is generated reflecting how pairwise comparisons correlate (correlation here means how each member of the pair varies with respect to the corresponding pair average interactions in conjunction for each of the bins). Red indicates a positive correlation and blue for a negative correlation.

(F): Zoom in from (D) with the same genome fragment as E showing a contact heat-map interaction matrix for chromosome 18 similar to (D). The inset is from a 3 Mb segment corresponding to a B compartment (as can be seen from (G)).

(G): From the information in (E), for every bin, a pattern of correlations is obtained and these patterns are classified clustered together, creating the A/B compartments based on their principal component of variability.

(H): Heat-map zoom of the shown 3 Mb area from (F), showing two TADs based on significant interaction frequency within the two different regions (1 and 2) compared with interactions between the regions.

(I): Looping of chromatin of the two regions (marked 1 and 2) represented microscopically in (C) and in the 3C techniques in (H). There is with a higher frequency than normal of interactions within region 2 (black square in (H)). The model of the loop is trying to match the contact frequencies from Hi-C data (Zhang et al., 2012b).

1.5.1. Chromosome Territories

In order of decreasing genomic extent, this classification begins with Chromosome Territories (Figure 1-3 A): elements within a chromosome (Figure 1-3 E) will interact more frequently with each other (even if they are greatly separated in genomic distance) than with regulatory elements on other chromosomes (Figure 1-3 D) in general. Inter-chromosomal spatial proximity is somewhat higher between chromosomes that are simultaneously similar in both size (only small and large, not medium) (Zhang et al., 2012a) and gene density (Lieberman-Aiden and Berkum, 2009; Sengupta et al., 2008). Due to the large extension of chromosomes, their capacity to relocate within the nucleus is very limited and their connections are highly determined by their initial positioning (Gibcus and Dekker, 2013). Inter chromosomal areas

close to each other in a cell-type specific manner may explain the frequency of certain chromosomal translocations in these tissues (Andrew J. Fritz et al., 2019)

1.5.2. A/B compartments, Lamin-associated domains and Nuclear Pore Complexes

Within one Chromosome Territory, chromosomes are divided into two alternating compartments: A and B (Figure 1-3 B, Figure 1-3 F, Figure 1-3 G), despite genomic distance (Lieberman-Aiden and Berkum, 2009). This clustering is thought to be mediated at least in part by the CTCF insulator-binding protein (Li et al., 2013). The A compartment generally contains transcriptionally active genes, is enriched for open chromatin and is gene-rich, while the B compartment contains inactive and gene-poor chromatin (Lieberman-Aiden and Berkum, 2009).

Different cell types express different genes, this means that A/B compartment classification is cell-specific and cannot be based solely on DNA composition (Lieberman-Aiden and Berkum, 2009). As with Chromosome Territories, the associations within each compartment are variable between different cells and determined by their initial positioning due to their large extension (Gibcus and Dekker, 2013).

A/B compartmentalization is also supported by interactions with the Nuclear Lamina capsule which shapes the Nuclear Envelope, active chromatin (A compartments) occupy the central regions while B compartments are found on the edges near the Nuclear Lamina in what is called lamin-associated domains (LADs). This organization is also maintained throughout cell differentiation (Peric-Hupkes et al., 2010) but is not always constant in all cell types. For example, in the photoreceptors of nocturnal animals or quiescent or senescent human fibroblasts, active chromatin is located in the periphery of the nucleus and heterochromatin in the centre (Bickmore, 2013). Other ways in which the typical chromatin arrangement can be inverted is by a translocation event having translocation partners in different compartments (Bickmore, 2013).

Nuclear Pore Complexes are made up of nucleoporins (Nups) and allow exchange of molecules such as RNA between the nucleus and the cytoplasm. Nups can be localized at the core of Nuclear Pore Complexes or dynamically orbiting the nucleus of cells, where they preferentially bind to active or induced euchromatin regions (Brown et al., 2008). They are thought to have an effect on chromatin remodelling among other actions (Tan-wong et al., 2009) and in some cases thought to maybe establish insulator boundaries between highly active and repressed genes (Brown et al., 2008; Kalverda and Fornerod, 2010).

1.5.3. Chromosomal domains and Topologically Associating Domains

Chromosomal domains (Figure 1-3 C) were observed using super-resolution fluorescence microscopy (Markaki et al., 2011). They are chromatin poor packaged structures present at the border of Chromosome Territories that contain a gene-rich outer layer with enrichment of transcriptional elements such as RNA polymerase II and histone modifications called the perichromatin region (Markaki et al., 2011). They are considered independent from A/B compartmentalization due to their different size (Gibcus and Dekker, 2013).

Within A/B compartments there is another hierarchical layer: Topologically Associating Domains (TADs), (Figure 1-3 H) defined as theoretical territories that have a significant enrichment (about two fold increase) of interactions between elements within them compared to interactions between elements from two different TADs (Dixon et al., 2012). This is important because most specific long-range interactions between enhancers and their regulating promoters occur within TADs (Shen et al., 2012). It is suggested that these enhancer – promoter interactions have higher frequency than other interactions within TADs (using CTCF binding sites as proxy for TADs in Sanyal, Lajoie, Jain, & Dekker, 2012). TADs therefore isolate the elements within them from the rest of the elements: it is thought that a TAD's extent and the location of the elements within a TAD determines to a high degree which interactions between elements within a TAD will occur in a cell (Gibcus and Dekker, 2013). One heavily studied way in which TADs are formed is through DNA looping.

1.5.4. DNA looping

To understand how enhancers regulate genes in a cell type specific manner, it is important understand what principles are acting at the chromatin level. There have been several mechanisms proposed to explain cell type specific enhancer action (reviewed in Johan H. Gibcus and Dekker, 2013). For example one mechanism in neurons, involves an enhancer decompacting its target gene *SHH* locus and increasing the distance between *SHH* promoter and the enhancer (Benabdallah et al., 2017).

The prevalent mechanism, however, thought to be responsible in a significant number of cases (Bulger and Groudine, 2011; Dekker and Mirny, 2016), is that enhancers and promoters interact by being in close three dimensional spatial proximity, enabled by DNA looping (Figure 1-3 I and Figure 1-4). This also helps to explain how enhancers can exert their effects on a network of promoters, within its domain of action and may exclude closest promoters (Fu et al., 2018) and (genes 2 and 3 promoters in dark blue and yellow rectangles respectively in Figure 1-5). In some cases, artificially bringing of the promoter close to the enhancer (thereby

simulating looping) is enough to generate transcription (Deng et al., 2012). Enhancer and promoter proximity seems to be necessary for gene activation in some cases (H. Chen et al., 2018).

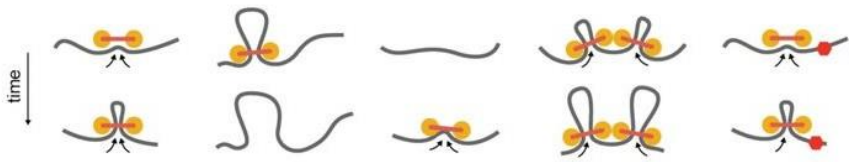


Figure 1-4: Loop extrusion formation. From (Fudenberg et al., 2015)

Loop extrusion model of chromatin domain formation across time (up to down): chromatin represented by grey lines, cohesin forming loops in yellow circles linked by red lines, TAD boundaries in red hexagons. From left to right: Loop extrusion formation of chromatin; Dissociation of the elements keeping the loop in place; Association of the elements to form a loop; Impeding of a growing loop by another formed loop; Impeding of a loop formation extrusion from one side by a TAD boundary.

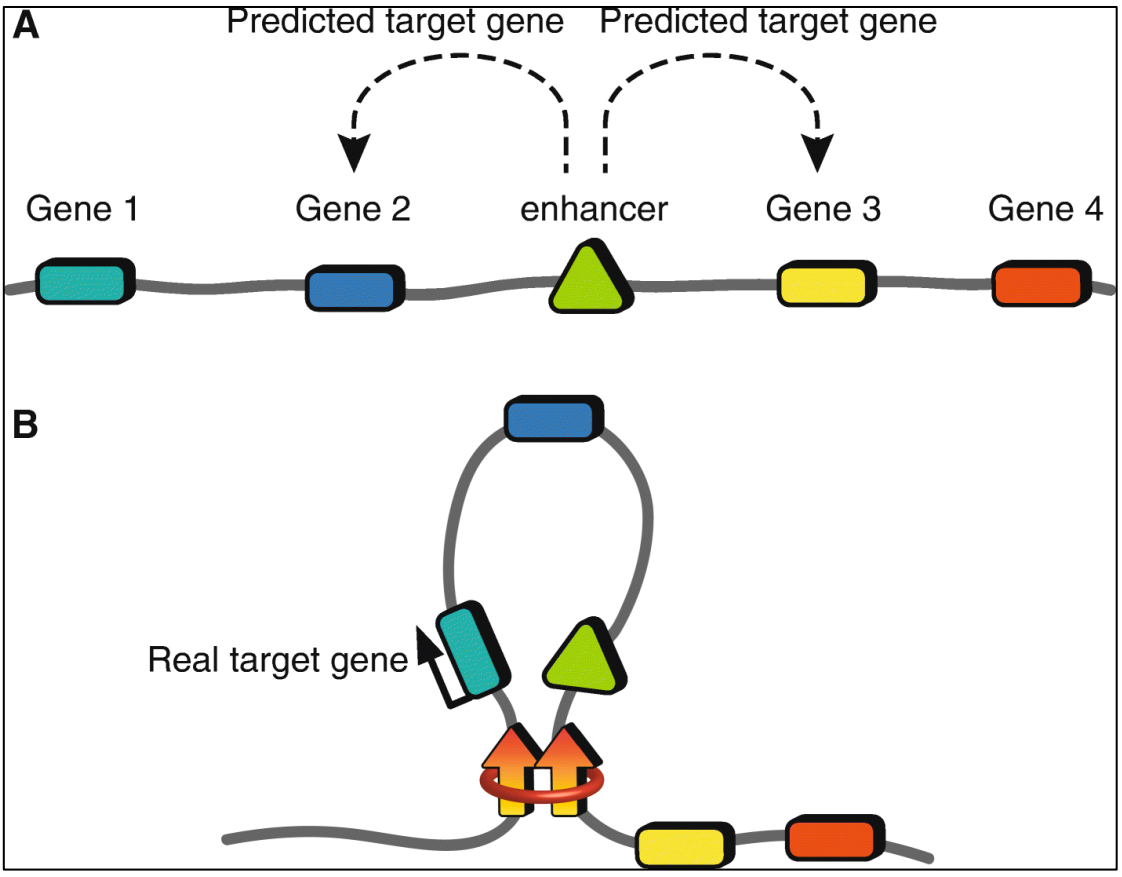


Figure 1-5: Prediction of enhancer - promoter interactions. From (Fu et al., 2018)

(a) *Typically assumed enhancer – promoter association of least genomic distance. The enhancer shown in the green triangle is thought to regulate the genes 2 and 3 promoters (dark blue and yellow rectangles respectively).*

(b) *Real gene regulation occurring through looping where the enhancer (green triangle) is really regulating gene 1 (light blue rectangle) and skipping gene 2 (dark blue rectangle). Additionally, the loop created by CTCF-CTCF (orange arrows) and cohesin (red ring) is isolating the enhancer from the genes 3 and 4 promoters (yellow and red rectangles respectively).*

DNA looping is a mechanism allowing the establishment of regulatory networks within a hierarchy of structures, (reviewed in Johan H. Gibcus and Dekker, 2013 and Figure 1-4, Figure 1-5). It is known that chromatin insulators such as CTCF can prevent enhancer-promoter interactions, stop heterochromatin spreading, and can even stop repression of chromatin by Polycomb response elements when CTCF binding sites are not methylated (Phillips and Corces, 2009; West et al., 2002).

This isolation mechanism can separate elements in cis from each other but are thought to not be able to prevent contacts of elements in trans (elements corresponding to different chromosomes) (Comet et al., 2011). CTCF insulators are both found within and flanking TADs in *Drosophila melanogaster* (Leblanc et al., 2012) which point to the fact that additional elements or mechanisms are needed for insulation. Cohesin complex was found to also be required for the boundary function (Roy et al., 2012; Vietri Rudan et al., 2015; Wendt et al., 2008). Additionally, the CTCF found at TAD boundaries was found to be oriented in opposite directions (de Wit et al., 2015; Gómez-Marín et al., 2015; Vietri Rudan et al., 2015).

Two studied ways of disturbing TADs (which can lead to diseases such as cancer) is by disrupting the TAD boundaries themselves or by gene rearrangements such as translocations within TADs (Valton and Dekker, 2016). Changing the orientation of flanking CTCF sites is proven to disrupt the boundaries (de Wit et al., 2015; Sanborn et al., 2015) and patterns of expression (Lupiáñez et al., 2015). All these observations have led to argue that previously used models to explain the assortment of chromatin, such as the equilibrium state for an ordinary condensed polymer or the fractal globule are not consistent with the observed data (Dekker and Mirny, 2016; Sanborn et al., 2015).

A loop extrusion model has been suggested as being a more plausible genome folding model (Goloborodko et al., 2016) through mathematical modelling of Hi-C data. Recently, this was confirmed through microscopy in yeast (Ganji et al., 2018). Here, cohesin forms dynamically increasing loops between cis-regulatory sequences. Cohesin forms a ring structure through

which the DNA slides, the sliding is restricted in both ends by CTCF bound proteins in convergent orientation in the chromatin. Once both the two DNA bound CTCF proteins come in contact, they homodimerize and create a stable complex with cohesin and this limits the loop extrusion, loops and TADs are thought to be dynamic structures (Hansen et al., 2017). It is suggested that multiple loops would form within TADs creating sub-TAD domains and would be limited at TAD boundaries by CTCF, thereby not allowing a loop to overlap two TADs (Fudenberg et al., 2015; Sanborn et al., 2015), see Figure 1-4.

The looping process influencing promoter-enhancer communication is thought to be governed by both general and, to some extent, cell lineage specific principles (Ghavi-Helm et al., 2014; Jin et al., 2013) for example through cell-type specific CTCF interactions despite similar cohesin and CTCF binding in different cell types (Hanssen et al., 2017; Hou et al., 2010). There is also evidence from single cell experiments of cell to cell stochastic variation within the same cell type (Strickfaden et al., 2010) yielding differential chromosome folding, TADs and DNA looping (summarized in Ulianov et al., 2017). This can affect enhancer – promoter interactions (and general contact patterns between any two regions in the genome) at any given time (Parada et al., 2003; Walter et al., 2003) for example, as a result of transcriptional bursting (Nicolas et al., 2017).

During differentiation there is a higher correlation in expression between genes in the same TAD (independent of the distance between them) compared with genes in different TADs (Nora et al., 2012). Moreover, evidence of relative conservation of TAD boundaries across species (Dixon et al., 2012) exists. There is also evidence that different cell types have different looping and gene expression levels (E. M. Smith et al., 2016), which clearly establishes cell type specific enhancer - promoter contacts. It has also been shown that interactions of loci located in different TADs are cell-specific (E. M. Smith et al., 2016). These observations reinforce the idea that TADs orchestrate the expression of genes within of them (Dixon et al., 2012).

1.6. Enhancer-promoter communication: how does a promoter determine its enhancer(s)?

In the Prediction of enhancers, section 1.4, different methods to obtain enhancers have been shown, once an enhancer is discovered, the next task is establishing which enhancer, or enhancers (Angelica and Fong, 2008), affect a given gene promoter (Gheldof et al., 2010). This relation can be many to many (Sanyal et al., 2012; Shen et al., 2012), for example, conserved developmental enhancers can be redundant in function and deleting one reduces but still

maintains some target gene expression (Zlotorynski, 2018). Enhancers and promoters can be separated by a great genomic distance and still interact (Angelica and Fong, 2008). For example, it was determined that around half of the promoter interacting regions (which should be enriched for potential enhancers) were interacting with one target promoter and around 35% were interacting with two to four promoters in human primary blood cell types (Javierre et al., 2016). These results are also consistent with another study performed on different human cancer samples (Corces et al., 2018). Indeed, it was determined that the interactions between a promoter and a non-promoter had a median linear distance between the elements of 331Kb with promoters interacting with regions within 1Mb in about 90% of the cases (Javierre et al., 2016). Another study found the average enhancer – promoter distance was 100Kb (Spitz, 2016). As it has been mentioned, recent studies have found that most enhancers interacting with promoters are within 20Kb (Hait et al., 2018; Kumasaka et al., 2018). Interacting frequency of enhancers with their target genes is also thought to decrease with distance (Corces et al., 2018; Hait et al., 2018).

As it is the case in the task of enhancer discovery, various methods and assays have been developed to elucidate enhancer – promoter interactions.

1.6.1. Correlation between chromatin accessibility and gene expression

The first step tends to be determining a set of enhancer and promoter interactions to test. Given the fact that most enhancer and promoter interactions occur within a certain distance, it is possible to infer these relationships through correlation between chromatin accessibility and gene expression where a large enough number of samples is available. This method is repeatedly found in the literature and the results obtained from this method give strong candidates to be tested. For example with a computationally determined kidney renal papillary cell carcinoma (KIRP) subgrouping a high positive correlation is found between the average chromatin accessibility at open chromatin regions in the neighborhood of the *MECOM* gene and gene expression of this gene (Corces et al., 2018).

It is also possible to determine candidate enhancer – promoter links by looking into the effects of somatic mutations in putative enhancer regions when Whole-Genome Sequencing (WGS) data is available (Corces et al., 2018). Once the mutations have been identified through WGS data, a mutation that allows binding of a new TF that opens chromatin around it will have higher ATAC-seq signal for that allele due to the nature of the assay. If the region is a true enhancer region, this activation of chromatin should be co-incident with a gene expression

increase. Conversely, if the mutation restricts the binding of a pioneer TF, chromatin accessibility will decrease and any gene being regulated by the region should be repressed.

1.6.2. Other commonly used methods

To test the enhancer – promoter predictions, different assays are used. Chromosome Conformation Capture (3C) is a family of assays that target certain viewpoint sequences of DNA (for example promoters) and obtain chromatin elements that are in close spatial proximity (for example enhancers), (Dekker, 2002). While spatial proximity does not imply a functional interaction, a functional interaction is assumed to involve spatial proximity. 3C techniques provide an average of interactions for the population of cells, each cell has a unique chromatin arrangement at any given point in time (reviewed in Gibcus and Dekker, 2013).

CRISPR interference (CRISPRi) is another validation tool which can be used to test enhancer and promoter interactions, it involves CRISPR, which uses a guide RNA to target specific enhancer regions with the catalytically dead CAS9 (dCas9) protein (Qi et al., 2013). The guide RNA can target putative enhancer regions to be tested and the KRAB (Krüppel-associated box) repressor fussed to CAS9 can repress these regions for example through changing the chromatin state to heterochromatin (Corces et al., 2018). If the relationship between the putative enhancer and the gene exists and assuming that this enhancer is necessary for the expression of the gene, the expression of the gene should be significantly reduced. Within this context, a recent method was developed termed crisprQTL mapping, briefly, this method targets multiple candidate enhancers simultaneously in each cell through CRISPRi. This creates unique combinations of deactivated enhancers per cell. Single cell RNA-seq is then used to quantify gene expression and relate it (for example, within each TAD) in each cell with the enhancer combination using an eQTL-like framework (Gasperini et al., 2019).

Other methods such as deleting the enhancers through CRISPR/CAS9 and seeing the target gene expression can be used (Cunningham et al., 2018). Also enhancer bashing (Venken and Bellen, 2014), CAS9-mediated *in situ* saturating mutagenesis (Canver et al., 2015) can also be used and have been mentioned in Prediction of enhancers, section 1.4.

1.7. Enhancer – promoter deregulation in cancer

Cancer is a disease of deregulation from the origin cell type in the different layers of gene regulation. At the accessible chromatin level (of which a significant subset is associated with

regulatory regions), a particular cancer cell's chromatin accessibility tends to be the most similar to the cancer cell type origin tissue than with other tissues, although exceptions can occur. Such is the case with ATAC-seq chromatin accessible regions from cancer types such as lung squamous cell carcinoma, which have a more similar pattern to DNase I peaks of breast tissue than lung tissue (Corces et al., 2018), although it has to be noted that different experiment batches and assays were used in this comparison. Another study also found, that in terms of chromatin accessibility cancer reactivates regions by gain of accessibility which were present in embryonic stem cells (ESCs) and other cell line lineages different from the origin cell tissue (Stergachis et al., 2013).

Multiple other ways describing how enhancers or their function is altered in hematopoietic cancers has been documented and is summarized (Bhagwat et al., 2018) and Table 1-1. At the enhancer – promoter interaction level, some deregulation examples include commissioning and activation of new enhancers which deregulate their target genes expression as shown when comparing healthy tissues with their cancer counterparts (Corces et al., 2018). In Clear Cell Renal Cell Carcinoma, a detailed deregulation mechanism was proposed: VHL deficiency was found to be involved in activation of a subset of enhancers through H3K27ac and some H3K4me1 gain, with some of these enhancers regulating *ZNF395* expression (Yao et al., 2017). These enhancers were found to acquire EPAS1/ARNT dimer binding further recruiting EP300 and increasing gene expression of target genes with little change in DNA interaction frequency between VHL deficient and VHL restored state in these VHL-dependent enhancers. This was done combining histone modifications with Capture-C and ChIP-seq of EP300 and taking enhancer-promoter interactions correlating enhancers H3K27ac signal with gene expression within TADs.

Combination of histone modifications and open chromatin have been used to determine oncogenic enhancer – promoter programs with clinical applications. For example, in a study involving ESR1 (Estrogen receptor alpha) positive breast cancers, putative enhancers were first obtained through histone modifications or open chromatin (Magnani et al., 2013). Then, it was concluded that ESR1-positive breast cancers responsive to endocrine therapy have active enhancers enriched in estrogen response elements (ESR1 binding motif) while ESR1-positive breast cancers resistant to treatments do not have the motif enrichment.

Disease	Alteration event	Effect on disease
AML	De novo *RARA* enhancer	Promotes sensitivity to potent RARA antagonists
B-cell lymphomas, multiple myeloma	t(8;14)	MYC driven by IgH enhancer
T-ALL	t(1;14)	TAL1 driven by TCR enhancers
T-ALL	Deletions	TAL1 driven by SIL enhancer
AML	t(3;3), inv(3)	EVI1 driven by GATA2 enhancer, hemizygous loss of expression of GATA2
T-ALL	Duplication at 8q24	Copy-number amplification of a NOTCH1-bound enhancer that drives MYC expression
AML	Copy-number amplifications 1.7 Mb downstream of *MYC*	Copy-number amplification of *MYC* enhancers
T-ALL	Focal indels 8 kb upstream of TAL1	Creation of de novo MYB binding site, generating a super enhancer that drives TAL1 expression
T-ALL	SNP 4 kb upstream of the *LMO1* transcription start site	Creation of de novo MYB binding site, generating an enhancer that drives LMO1 expression
CLL	Mutations at 9p13	Disruption of enhancer that regulates PAX5
CLL	Mutations at 15q15.1	Disruption of RELA enhancer that regulates *BMF*, leading to increased risk of CLL development
T-ALL	Aberrant NOTCH1 activity	NOTCH1 binds to an enhancer to drive LUNAR1 transcription. LUNAR1 is required for IGFR1 expression and T-ALL survival
AML	DNMT3A R882H mutations	Mutant DNTM3A leads to loss of methylation at broad enhancers, activation of self-renewal gene programs
AML	TET2 mutations	Mutant TET2 leads to hypermethylated DNA at enhancers, resulting in suppression of gene expression
AML	Cohesin complex mutations	Impaired differentiation, increased self-renewal in hematopoietic stem and progenitor cells
T-ALL	CTCF binding site deletions	Disruption of TAD insulation surrounding *TAL1*and *LMO2* genes, leading to aberrant enhancer activation of these genes

Table 1-1: Examples of alterations of enhancers and their effects in hematopoietic malignancies, adapted from (Bhagwat et al., 2018)**.**

B cell malignancies such as B-cell lymphoma have also favored from studying enhancer deregulation from combining different data sources: FAIRE-seq, H3K27ac, H3ac and RNA-seq (Koues et al., 2015). First, characterizing non-cycling centrocytes as the most similar healthy cell type of origin of the cancer in terms of regulatory regions (through chromatin accessibility) and gene expression. Then combining active enhancer marks with the chromatin accessible regions to derive altered regulatory regions in B-cell lymphoma compared with healthy and finding a significant amount of putative enhancers that were attenuated in cancer. Enhancer - promoter interactions were determined first by connecting altered distal enhancers (FAIRE-

seq, H3K27ac and H3ac) to their target genes within 500Kb. Then correlating chromatin accessibility changes (cancer vs. healthy) between distal enhancers and TSS and selecting only concordant changes in RNA-seq of the target genes (augmented, attenuated or unchanged in both chromatin accessibility and RNA-seq expression of putative target genes). Biologically relevant oncogenic pathways were found to be promoted by activation and decommissioning of enhancers, such as general cellular transformation gain and apoptosis loss respectively and enhancer associated TFs were derived. SNVs of patient-matched healthy peripheral blood mononuclear cells and B-cell lymphoma samples were used to infer enrichment of these events in altered cancer enhancers and examples of disruption of enhancer activity with relevant biological repercussions through decreased TF motif affinity were proposed (Corces et al., 2018) and tested (Koues et al., 2015). Linking putative enhancers to target genes within 500Kb through active enhancer signal correlation was also done in AML through H3K27Ac (Mckeown et al., 2017) with 62% super enhancer – gene interaction Hi-C validation in a close cell type. In this case, samples within certain AML subgroups were found to contain a super enhancer regulating the *RARA* gene with relevant prognosis and clinical applications. In another study, H3K27Ac was also used to identify enhancers and super enhancers in Adult T-cell Leukemia/Lymphoma, linking them to the nearest genes and determining super enhancer associated important critical genes for the malignancy (Ishida et al., 2017).

In MM, enhancer - promoter interactions have been analyzed comparing MM cells from bone marrow with a proxy for bone marrow Plasma cells: memory B-cells *in vitro* differentiated into Plasma cells in one study (Jin et al., 2018). MM exclusive, H3K27ac determined enhancers were assigned to target genes using GREAT (McLean et al., 2010) and genes within 200Kb up and downstream of each enhancer within CTCF boundaries. To my knowledge, super enhancers creating an oncogenic state were first identified in MM (Young et al., 2013), where a special group of large enhancers were found to be specially enriched in BRD4 and Mediator co-activators in comparison to regular enhancers by 16-fold and 18-fold respectively. Superenhancer target genes included MM biology relevant genes such as "*MYC*, *IRF4*, *PRDM1*, and *XBP1*", all of them associated with oncogenic processes when OE. MM super enhancers were determined from H3K27ac using Ranking Of Super Enhancer (ROSE: Whyte et al., 2013; Young et al., 2013) and TFs associated with MM super enhancers and enhancers were observed for TF motif enrichment relevant for MM such as ZFP36, PRDM1 or FLI1 (Jin et al., 2018). Additionally, integration of ATAC-seq using TF footprinting, ChIP-seq, RNA-seq yielded a TF network for MM including TF such as IRF4 or FLI1 having central roles.

Another attempt at creating a catalogue of MM enhancers and affected genes was performed on the basis of recurrent mutations using primary MM whole-exome and whole-genome sequencing data and the effect of mutations on gene expression with matched RNA-seq data (Hoang et al., 2018). 114 recurrently mutated (in MM samples) cis-regulatory regions (a proxy for enhancers) thought to interact with 271 genes in B-cells (through the use of B-cell promoter Capture Hi-C data). Despite this, the study found only seven of these interactions having a significant association between mutated candidate enhancers and altered gene expression with a frequency of 1%-6% of the total samples considered in the differential analysis. In terms of enrichment in MM subgroups according to cytogenetic event classification, only two interactions were associated with different subgroups with less than 5% of the subgroup samples affected (Hoang et al., 2018). Despite these efforts, to date, primary MM cells and PC from bone marrow have not been compared. Furthermore, subgroup specific changes have not been extensively determined. Leaving a field of opportunity to be explored.

Copy number variations in genes such as *MYC* (Nau et al., 1985; Schwab et al., 1983) and enhancer regions can also cause tumorigenesis. An example of the latter can be seen in a study with different cancer types, amplified copies of active (containing H3K27ac) cancer tissue-specific super-enhancers led to significant overexpression of target genes such as *MYC* in lung adenocarcinoma (Zhang et al., 2015). In the study, these interactions were verified to occur only in the relevant tissue through Chromosome Conformation Capture, EP300 binding, luciferase reporter assays showing duplication of the enhancer augmenting reporter gene transcription and CRISPR/CAS9 repressor assay targeted to the enhancer significantly reducing transcription of the target gene and the tumorigenesis properties. Amplifications in enhancer regions in AMLs can target *MYC* also in leukemia (Shi et al., 2013) and T-ALL (Herranz et al., 2014).

Carcinogenesis can occur by placing active enhancers close to aberrant target genes for example through disruption of TAD boundaries (as explained earlier) or by gene rearrangements such as translocations within TADs (Valton and Dekker, 2016). Chromosomal translocations can also put genes such as *MYC* under the influence of active enhancers such as the IgH enhancers in Burkitt lymphoma cells (Dalla-Favera et al., 1982) and MM (Affer et al., 2014). The *IgH* enhancer can also deregulate *CCND1, MAF* and *NSD2* genes in MM (see Multiple Myeloma and Plasma cells, section 1.7.2 and Chromosomal translocations, section 1.8.4). Normally, studies take into account the overexpression of the new gene put under the expression of the rearranged enhancer, but it is also important to take into account the under

expression of the gene under the normal influence of the enhancer before the translocation event. Such is the case with the key *GATA2* TF gene, which loses the translocated enhancer, and the reduced gene expression significantly accelerates leukemia initiation and progression in AML (Katayama et al., 2017).

In addition, aberrant repression of enhancers through gain of nucleosomes and methylation or activation through loss of nucleosomes and hypomethylation are other possible enhancer deregulation mechanisms. For example, in AML, a DNA methyltransferase-encoding gene *DNMT3A* is mutated and this can cause loss of 5-methylcytosine and subsequent gain of an active chromatin properties at enhancers thereby deregulating nearby genes (Lu et al., 2016). In AML too, mutations of the *TET2* gene generate a loss of demethylation function which results in hypermethylation at enhancers and loss of expression of the target genes (Rasmussen et al., 2015). Other examples involving methylation study methylation patterns in prostate cancer vs. healthy cells finding methylated and hypermethylated regions in cancer. Particularly intronic regions with a high degree of sequence conservation which points at potential enhancers (Yegnasubramanian et al., 2011). The effect of enhancer methylation on the target gene expression in cancer vs. healthy genomes was evaluated through machine learning algorithms (Aran et al., 2013), finding hypomethylated enhancers associated with upregulation of target genes and hypermethylated enhancers with downregulation in cancer vs. normal tissues. Hypermethylated enhancers with nucleosome repletion in breast and prostate cancer vs. healthy state were also found to have epigenetic repressive marks while nucleosome depleted and hypomethylated enhancers showed active epigenetic marks (Taberlay et al., 2014). Another striking finding was that in cancer, enhancer methylation state had a higher association than promoter methylation state with gene expression (Aran et al., 2013; Aran and Hellman, 2013).

Aberrant TF expression can also activate enhancers and initiate a cascade of deregulation events; this is the case in T-cell Acute Lymphoblastic Leukaemia, where *NOTCH1* activates an enhancer driving the expression of the lnc-RNA *LUNAR1. LUNAR1* in turn occupies with *NOTCH1* an *IGFR1* enhancer driving abnormal *IGFR1* expression (Trimarchi et al., 2014).

1.7.1. Cancer subtyping

Cancer is a very general term that refers to a group of diseases characterized by uncontrollable growth and division of cells with the possibility of invasion into other tissues. This classification only refers to some of the functions that a cell must possess in order to be malignant, but it does not address the specific causes and pathways taken to become malignant. Even a typical

classification of cancer by originating tissue cell type is very broad since every cancer is genetically and epigenetically different and even every cell within the same cancer varies to a large degree due to genome instability. It is therefore important to further sub-classify cancers with common features into groups where actions in terms of diagnosis, prognosis, treatment design and response prediction can be assigned to each subgroup based on its characteristics (Zhao et al., 2018).

Cancer classification has been traditionally performed with histological analysis (study of diseased tissue through a microscope). In this regard, presence or absence of chromosomal translocations (see Chromosomal translocations, section 1.8.4), or amplification/deletion of chromosome arms has also been used as biomarkers (Molist et al., 2004). As more in-depth data is available from new assays such as genomic sequencing, subtyping of different cancers based on this criteria has been possible (Zhao et al., 2018). This has been enabled by first elucidating a set of biomarkers that are relevant to a particular disease and then applying classifiers based on the patient data. There are many examples of biomarkers used in this process, a typical one used is gene expression whether done by Microarray or RNA-seq. Recently, RNA-seq is being more frequently used due its higher accuracy allowing it to discover novel and allele specific transcripts (Wang et al., 2009). A common process is to first determine a subset of genes that are most informative for the disease and then determine which patient samples group together using only these genes features through clustering and similar techniques. The typical gene expression profile for each of the sub classes is determined to create a classifier of the different sub classes, one of the first cancer type to benefit from this was human Acute Leukaemia (Golub et al., 1999).

Other general biomarkers used have been DNA methylation to identify patterns in pancreatic cancer among other cancer types (Kumar Mishra and Guda, 2017) or DNA mutations by themselves or in combination with biological pathway knowledge (Kuijjer et al., 2018). Micro RNAs (miRNA) can also serve as potential biomarkers, they are non-protein coding, around 22 nucleotides long and regulate gene expression through various mechanisms, they have been used in subtyping of breast cancer for example (Sherafatian, 2018). Other features used are structural variants in the DNA affecting stretches of 1Kb-3Mb, that, by either duplication or deletion, generate different number of copies are also another type of biomarker (Feuk et al., 2006). They are also used in cancer subtyping, for example, to stratify young squamous cell carcinoma of the oral tongue (SCCOT) patients into different prognostic groups which can determine the level of personalized treatment required to improve survival (Gu et al., 2018).

Combinations of the different information types above are also used (Kumar Mishra and Guda, 2017).

Another type of general profiling used in the literature is based on chromatin state and accessibility marking of regulatory elements. This can be done genome-wide, within the regulatory context of a certain oncogenic gene such as *MYC*, on particular regulatory elements types such as enhancers or promoters or subsets of elements, for example elements that are unique to a particular condition (Corces et al., 2018). Other enhancer features explained before such as correlations between enhancers and promoters or TF binding to active enhancers can also be used to this end.

There are also cancer-specific biomarkers for particular cancers, for example, for breast cancer there are different types based on the location of the malignancy: Ductal Carcinoma In Situ (DCIS), which means the disease, is within the ducts and Invasive Ductal Carcinoma (IDC) if the cancer has spread to the surrounding tissues. Breast cancer also has particular molecular subtypes based on Estrogen Receptor, Progesterone Receptor and *ERBB2* (HER2) gene amplification status (Gilcrease, 2015).

1.7.2. Multiple Myeloma and Plasma cells

MM is a malignancy of the terminally differentiated plasma cells (PC): B-cells, which are long-lived, produce immunoglobin (Bianchi and Munshi, 2015) and provide protection against antigens that activated them. The plasma cells originate in the bone marrow, during malignancy, they multiply uncontrollably, invading space and preventing the production of other blood cell types, additionally they produce abnormal immunoglobulin proteins that cannot fight disease and can be deleterious to the kidneys.

Studies involving multiple genomes, have shown MM to be characterized by a high heterogeneity within an individual population of cells, creating sub clonal populations of cells over time (Egan et al., 2012; Lohr et al., 2014). The heterogeneity is also spatial, since different clones of the disease are in different regions of the bone marrow with constrained access between them. All this makes MM a complex disease (Holstein et al., 2018). Despite advances in therapies, these characteristics make MM an incurable cancer to date (Ravi et al., 2018).

MM is nearly always preceded by Monoclonal Gammopathy of Uncertain Significance (MGUS) followed by Smoldering Myeloma precursor states in plasma cells (Landgren et al., 2014; Weiss et al., 2009). MGUS to Smoldering Myeloma is characterized by increasing percentage of "clonal PC in the bone marrow" (van de Donk et al., 2016). Investigations into the initiating

driver events of MM generally consider a set of partially overlapping characteristics occurring both in early MGUS and MM, these include chromosomal translocations in the IgH locus due to aberrant Class-switch Recombination (CSR, see below), hyperploidy and deletions of chromosome 13 components (del13). These events directly or indirectly all derail expression of the Cyclins: CCND genes (Kuehl and Bergsagel, 2012) (reviewed in Prideaux et al., 2014) and they account for a large proportion of the cases (Bergsagel and Kuehl, 2005). Secondary events found are secondary translocations (not derived from aberrant CSR), loss of heterozygosity (LOH), copy number variations (CNV) and epigenetic modifications coupled with a favorable microenvironment (reviewed in Prideaux et al., 2014).

Characterization of the primary driving event involving the deregulated oncogene gives rise to MM patient subgroups. Nearly half of all MM cases present a non-hyperdiploid profile with known oncogenes affected by chromosomal translocations of the IgH enhancers, these are: *CCND1*, *FGFR3* and *NSD2*, *MAF*, *MAFB*, and *CCND3*, with the remaining cases being hyperdiploid (Türkmen et al., 2014).

1.8. Plasma cell development

Since chromosomal translocations involving the IgH locus account for a significant proportion of MM driving events, it is important to understand the mechanisms that create this outcome during the development of plasma cells (see Table 1-2 for details regarding the expression of PC relevant genes) and characterize them. Human antibody molecules (and B cell receptors) are composed of Immunoglobulin Heavy chain (IgH) and light chains. The germ line IgH locus contains Constant (Cμ and Cδ), Variable (V), Diversity (D), Joining (J) gene segments (Birshtein, 2014) (see Figure 1-6 VDJ and Figure 1-7 regions in yellow). The gene expression of this locus is regulated by the μ-Enhancer and the 3' RR cluster of enhancers (composed of the HS3, HS4 and HS1,2 enhancers) and one per constant region at the 3' end in the human locus, which can be seen in Figure 1-6 in brown.

As can be seen in Table 1-2, induced DNTT (a gene encoding a template independent DNA-polymerase which contributes to generate antigen receptor diversity by adding nucleotides at the cleavage ends) and recombinase makes the Lymphoid early progenitor cells (marked as "early pro-B cell" in Table 1-2) undergo D-J joining on the Heavy chain (see Figure 1-7). In this process, any DNA between one D and one J gene segment of the heavy chain locus is deleted. Recombination-Activating Genes proteins and DNTT are also required for Somatic Hypermutation (SHM). At this point the expression of the Light Chain is the initial (not

rearranged) and the cells express Immunoglobulin variable alpha (α) and beta (β) chains (Ig αβ), throughout the whole cycle. Bruton's tyrosine kinase (BTK) is involved in the B cell maturation cycle (Satterthwaite, 2018) and activation and begins to be expressed on this stage (Table 1-2).

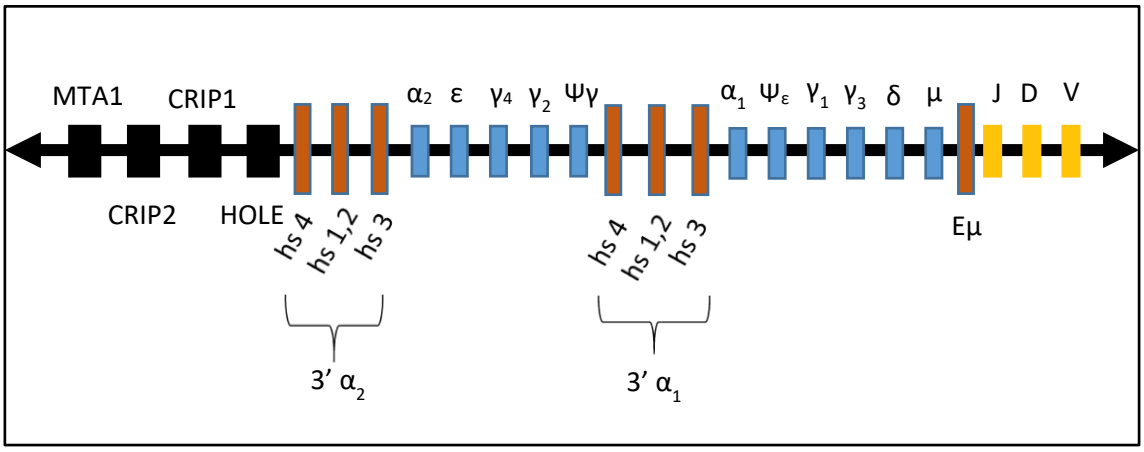


Figure 1-6: Schematic representation of the human Igh gene loci with emphasis on the 3' RR enhancers. Adapted from (Birshtein, 2014).

The IgH locus left to right corresponding to centromere to telomere (chromosome 14). The surrounding genes coloured in black, enhancers coloured in brown, in blue the constant regions involved in Class-Switch Recombination and in yellow the VDJ regions involved in VDJ Recombination. There are a group of enhancers per constant region: α_1 and α_2.

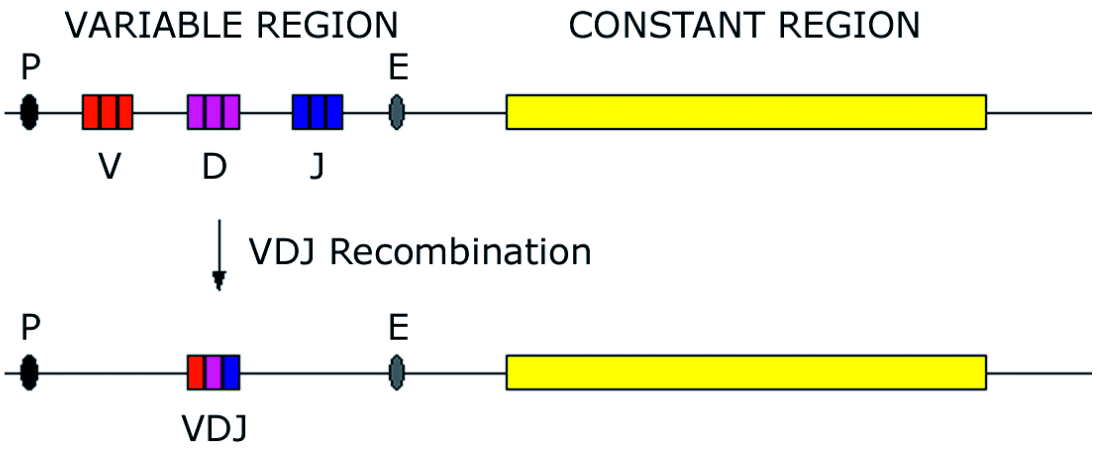


Figure 1-7: Simplistic mechanism of VDJ recombination. From (Market and Papavasiliou, 2003)

The figure shows VDJ recombination where excising of all DNA between the D and J segments selected and the V segment with the DJ combination occurs. The constant region (yellow rectangle) is unaltered during this process, so expression of the IgH gene includes the constant domain all throughout VDJ recombination. P: Promoter and E: Enhancer.

Stages in B Cell Development							
	stem cell	**early pro-B cell**	**late pro-B cell**	**large pre-B cell**	**small pre-B cell**	**immature B cell**	**mature B cell**
H chain genes	germline	D-J joining	V-DJ joining	VDJ rearranged	VDJ rearranged	VDJ rearranged	VDJ rearranged
L chain genes	germline	germline	germline	germline	V-J joining	VJ rearranged	VJ rearranged
Surface Ig	none	none	none	μ chain in pre-B receptor	μ chain in cytoplasm and on surface	membrane IgM	membrane IgM and IgD
RAG, DNTT expression	no	yes	yes	no	yes	yes	no
Surrogate L chain expression	no	yes	yes	yes	no	no	no
Ig αβ expression	no	yes	yes	yes	yes	yes	yes
btk*	no	little	yes	yes	yes	yes	yes
Membrane markers	CD34	CD34 CD45 (B220) Class II	CD45R Class II CD19 CD40	CD45R Class II pre-B-R CD19 CD40	CD45R Class II pre-B-R CD19 CD40	CD45R Class II IgM CD19 CD40	CD45R Class II IgM IgD CD19 CD21 CD40

Table 1-2: Stages in B Cell development. Adapted from (Decker, n.d.)

** Bruton's tyrosine kinase*

Late progenitor B cells (marked as "late pro-B cell" in Table 1-2) undergo V-DJ rejoining (see Figure 1-7): joining of one V gene segment, from a region upstream of the newly formed DJ complex, forming a rearranged VDJ gene segment.

Late progenitor B cells become large pre-B cells (Table 1-2) when RAG and DNTT expression stops and VDJ rearrangement ends and they express membrane μ chains with surrogate light chains in the pre-B receptor.

Small pre-B cells express μ chains in the cytoplasm and surface and undergo V-J joining in the light chain. Once Light chain has been successfully synthesized, it is expressed with μ chain on the cell membrane as IgM and the cell is called an immature B cell. The immature B cells migrate from the bone marrow to the blood undergoing further differentiation and becoming mature B cells.

When mature B cells bind to an antigen that is recognized both by the B-Cell Receptor (BCR) and T cell receptor (TCR) in T dependent responses, they cooperate together activating each other in a positive feedback loop where T cells activate B cells and B cells express molecules such as *CD80* and *CD86* required for the activation of T cells. In turn, T cells can differentiate into Th2 cells which can trigger differentiation of B cells into plasma cells, allowing CSR and SHM (Vale, 2010).

1.8.1. V(D)J Recombination

The IgH locus experiences multiple DNA rearrangements and transformations during the development and differentiation of B cells to give rise to the diverse human immunological response repertoire (Birshtein, 2014). V(D)J is one such recombination event (see Figure 1-7), which creates specific antigen receptor genes to recognize foreign antigens (Ebert et al., 2015).

In this process, first the Immunoglobin Heavy Chain locus is rearranged, combining one D-segment with one J-segment and eliminating the D and J segments in between. The mechanism (Figure 1-8) includes two phases: cleavage and joining. Recombination signal sequences (RSS) are brought together with the help of RAG proteins and RAG proteins produce nicks in both RSS. The cleaved DNA is then rearranged, processed and repaired and the ends are combined by non-homologous end joining (reviewed in Schatz & Swanson, 2011) where nucleotides are removed (Lieber, 1996) non-templated and palindromic nucleotides are added (Lieber, 1996). This is the highest determinant in diverse pathogen binding response repertoire. All this is facilitated by a process of long-range looping in pro-B cells (Ebert et al., 2015) thought to be mediated by CTCF and cohesin in mice, among other factors (reviewed in Aiden & Casellas, 2015).

In a similar process, the newly created D-J segment combines with a V-segment, eliminating all D and V segments in between (Aiden & Casellas, 2015). By this point, the cell is called a late

pro-B cell and it is expressing a transcript containing the VDJ region of the heavy chain and both of the constant chains: C_μ and C_δ (Vale, 2010). Pro-B cells become pre-B cells when they express membrane complete μ (Heavy chain) protein with surrogate (germ line) light chains to form the pre-B receptor, signalling the end of the heavy chain rearrangement (Vale, 2010).

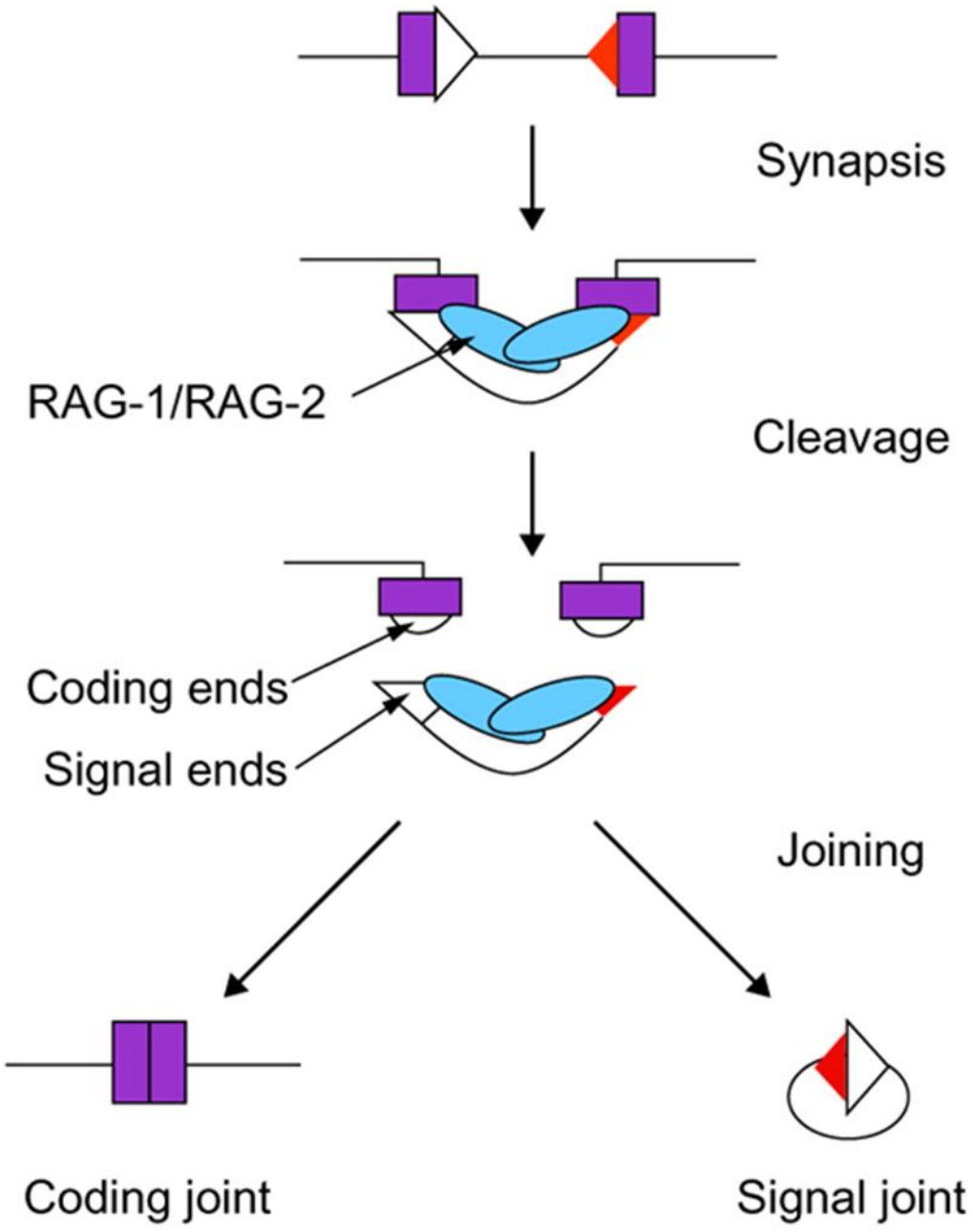


Figure 1-8: Rearrangement in VDJ recombination. From (Roth, 2000).

During the synapsis phase RAG proteins (blue) bind to the recombination signal sequences (RSS) in the DNA (marked by white and red triangles) in the proximity of V, D or J coding elements (purple rectangles). DNA cleavage occurs at both RSS sites, generating DNA coding ends and signal ends. These elements are then repaired and ends are combined by non-homologous end joining.

After four to six rounds of cell division, V-J segment joining occurs in the Immunoglobulin Light Chain and IgM is expressed on the cell membrane (Vale, 2010), at this point the cell is called an immature B cell. The cells migrate from the bone marrow to the blood undergoing further differentiation and becoming mature B cells with the ability to express IgM and IgD through alternative splicing (Vale, 2010).

1.8.2. Class-switch recombination

When the B cell binds to an antigen (such as bacteria or virus) via its B-cell receptor on secondary lymphoid organs, B cell activation begins. The activation can be T-cell dependent or independent and causes further differentiation (Vale, 2010). There are repetitive elements called switch (S) regions, which occur just before each region in the constant chain (Figure 1-6 blue regions). According to current knowledge, these regions when transcribed create R-loops. R-loops expose accessible regions of single stranded DNA that can be targeted by activation-induced cytidine deaminase (AID/AICDA) (Chaudhuri et al., 2007) and end up as double-strand breaks, fixed by means of the Base excision repair and DNA mismatch repair (MMR) pathways (reviewed in Matthews et al., 2014).

When activated B cells encounter specific signalling molecules they undergo antibody class switching, this is attained by removing of constant regions between the Eμ enhancer and a selected switch region, thereby putting the Eμ enhancer adjacent to the selected constant region. This process produces one of the different antibody isotypes, for example, if the constant region exon γ_3 is adjacent to the Eμ enhancer, the antibody isotype is *IgG3*, if instead, exon ε was selected, *IgE* will be produced (Figure 1-6 blue regions). Class-switch recombination allows a mature B cell to acquire new functions while at the same time maintaining the binding to a specific antigen.

1.8.3. Somatic hypermutation

Somatic hypermutation (SHM) occurs during B cell development in the Germinal Centers, single nucleotides are changed in a sequential manner on the rearranged V-region and the surrounding nucleotides. The mutations are not completely random since there are certain

preferences in the conversions of nucleotides (Vale, 2010). As with CSR, SHM allows B-cells to add functions to the immune response without affecting affinity binding.

1.8.4. Chromosomal translocations

It is possible that during V(D)J and class-switching recombination, a non-legitimate binding occurs when breaks in DNA are rejoined whether through homologous or non-homologous end joining to achieve an immune response (Aplan, 2006), (Figure 1-9 A: abnormal VDJ recombination and B: abnormal CSR). Thus, new regulatory sequences are introduced (Figure 1-9 A and B, Eδ enhancer, light blue rhombus) in the context of oncogenic genes and this can lead to deregulation of those oncogenes. It is important to note, that not all translocations occur as a result of plasma cell development processes but can take place in mature B cells (Affer et al., 2014).

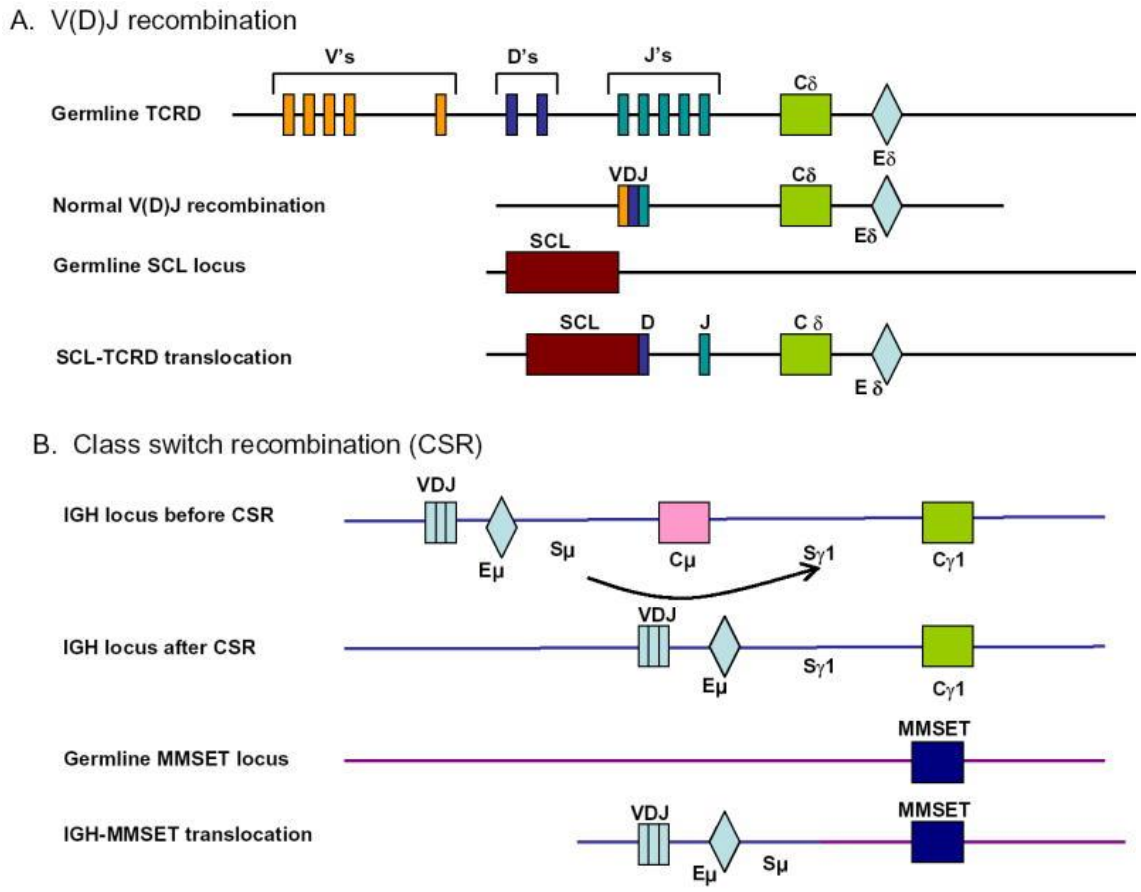


Figure 1-9: Chromosomal translocations during V(D)J recombination and CSR. From (Aplan, 2006)

(A): VDJ rearrangement of the germ line T-cell receptor (TCRD) containing the V (orange), D (dark purple) and J (dark green) regions, the constant region Cδ (light green) and the Eδ enhancer (light blue) showing the normal result of

the TCRD locus and the germline SCL locus (dark red) after rearrangement. "SCL-TCRD translocation" showing the result of a recombination signal sequence in the SCL locus being recognized by the V(D)J mechanism, generating a non-legit SCL-TCRD translocation where the expression of the SCL locus is affected by the TCRD Eδ enhancer.

(B): Before CSR recombination, IgM is produced beginning with an mRNA transcript including the VDJ segment (light blue rectangles) and stopping (spliced) at Cμ (pink rectangle). CSR occurs between the switches in the constant regions Sμ and Sγ1 (marked in the ends of the black arrow), denominating immunoglobin classes. When this process occurs normally, the regions between Sμ and Sγ1 are excised from the DNA and now the protein produced is IgG which corresponds to the transcript including VDJ and splicing at Cγ1 (green rectangle). When a translocation occurs, the IgH Eμ enhancer (light blue rhombus) is bounded to the MMSET (NSD2) gene (dark blue rectangle) chromosome region and the expression of NSD2/MMSET is affected by the IgH Eμ enhancer. This type of translocation can also occur with other genes, such as CCND1.

1.9. Aims

MM is a disease where around half of the cases are characterized by a critical driving translocation event involving the IgH enhancer resulting in abnormal gene expression (Türkmen et al., 2014). As such, this thesis aims at studying whether there are other clinically relevant examples of gene deregulation in MM (and particularly clinically determined MM subgroups) caused by associated altered enhancer states.

To address this, novel primary MM and PC ATAC-seq assay data will be used to first obtain a set of candidate enhancers based on the chromatin accessibility (see Properties of enhancers, section 1.2) and RNA-seq information will be combined to correlate regions with target genes (thought to be regulated in a condition specific manner: MM or MM subgroup). This will be done using principles and techniques outlined in Enhancer-promoter communication: how does a promoter determine its enhancer(s)?, section 1.6 and Enhancer – promoter deregulation in cancer, section 1.7. These regulatory candidate enhancers will then be studied in the context of publicly available histone modification signal data to determine whether they are novel to MM or active in other tissues, particularly in the B-cell lineage. Furthermore, regions with promising regulatory potential will be validated for being in close proximity with the target promoters in B-cells. To establish viability of predictions, candidate enhancer – gene pairs will be tested through one of the methods outlined in Enhancer-promoter communication: how does a promoter determine its enhancer(s)? section 1.6, thereby aiming to elucidate important interactions for MM biology.

Once relevant DNA regions causing Myelomagenesis are discovered, the present study will also try to elucidate what may be the key TF proteins that enable these candidate enhancers

involved in oncogenic gene expression. It will do so by studying the condition specific candidate enhancers for enrichment of binding sites, while putting observing how the expression of these TFs changes in the malignant states. Analogously, by analysing the deregulated genes, general pathways affected in the transition to the oncogenic state will also be determined. Moreover, gene expression alteration as a result of corresponding promoter accessibility will be studied to determine the importance of this factor in the cancer state.

Finally, it will be verified whether there is a genuine PC, MM and MM subgroup specific effect and whether new subgroups emerge when considering chromatin accessibility and gene expression fluctuations simultaneously in an unsupervised manner. Additionally, for relevant sample separations, associated elements previously stated in this section will be studied: enhancers, deregulated genes, interactions involving the former, TF proteins and gene pathways. Moreover, the relative contribution to the overall effects of the samples divisions will be calculated and the results compared with supervised analysis details as well as previous literature.

These interactions, key deregulated genes and TF activating enhancers should serve as a basis to design personalized treatment, such as the one based on enhancer biology as it has already been done before in Multiple Myeloma (Young et al., 2013).

2. Chapter 2: Materials and Methods

All data files referred to in this thesis are archived (Alvarez-Benayas, 2020a) unless specified so.

2.1. ATAC-seq and RNA-seq assay in primary PC and MM samples

Plasma cells and multiple myeloma samples from patients BM aspirates were subjected to red cell lysis. MM PCs were purified by two rounds of CD138 immunomagnetic selection (Miltenyi Biotech) following the manufacturer's instructions. Pre and post selection purity was assessed by FACs analysis (BD LSR-Fortessa) using CD138, CD45, CD19, CD56 and CD38 panel antibodies. In order for a sample to be processed (after CD138 + purification), only samples with a minimum of 85% of purity (by FACs for CD138, CD38, CD45, CD56 and CD19) and at least 500.000 pure cells were selected. Purified cells were immediately processed for ATAC-seq and RNA-seq.

For PC BM samples, mononuclear cells (MNCs) from BM aspirates were isolated by ficoll (histopaque, Sigma). The MNCs were pre-cleared of T cells and Monocytes by consecutive immunomagnetic negative selection (CD3/CD14- EasySep StemCell Technologies) following manufacturer's instructions. The sample was stained and sorted for positive CD138, CD319, CD27, CD45 and CD38, and negative for CD2, CD3, CD14, CD16,4TPA and 7AAD. (FACSAriaII, BD Biosciences) and CD19+ or CD19 neg. The sorted cells were immediately processed for ATAC-seq and RNA-seq.

ATAC-seq was performed as described in (Buenrostro et al., 2013). Briefly, 50.000 purified Plasma cells, myeloma plasma cells or cell lines, were washed with cold PBS (Sigma, UK) at 500g at 40C for 5 min. The cells were resuspended in 50 µL of cold Lysis Buffer (10 mM Tris-HCl, pH 7.4, 10 mM NaCl, 3 mM MgCl2, 0.1% IGEPAL CA-630) and washed at 500g at 40C for 10 min. The nuclei were subjected to transposase reaction for 30 min at 37.0°C; termination of the reaction and DNA purification was performed using a MiniElute Kit (Qiagen) and eluted twice with 10 µL. The purified DNA was amplified with NEBNext High-Fidelity 2x PCR Master Mix (New England Biolabs). The PCR amplified product was cleaned twice with (0.9X) AMPure beads (Beckman). The quality of the libraries was assessed with the Bioanalyzer High Sensitivity DNA kit (Agilent). The libraries were quantified using the NEBNext Library Quant Kit for Illumina (New Engand Biolabs) on a StepOne Plus Real-Time PCR (Applied Biosystems). The libraries were sequenced at the Genomics Facility at Imperial College London using the Illumina HiSeq 4000 platform to obtain paired-end 75bp reads.

Total RNA was isolated using the Nucleospin RNA kit (Macherey-Nagel) and quantified using the Qubit RNA Assay kit (Life Technologies) and RNA quality was assessed on the Bioanalyser using the RNA pico kit (Agilent). Total RNA libraries were prepared by removing the ribosomal RNA with NEBNext rRNA depletion kit (New Engand Biolabs) and NEBNext Ultra II Directional RNA Library Prep kit for Illumina (New Engand Biolabs), following manufacturer's instructions. Library quantity was determined using the Qubit High Sensitivity DNA kit (Life Technologies) and library size was determined using the Bioanalyser High Sensitivity DNA kit (Agilent). Libraries were diluted to 2nM and sequenced using the Illumina HiSeq 4000 platform the Genomics Facility at Imperial College London to obtain paired-end 75bp reads.

The processes were supervised by Valentina Caputo (Haematology, Division of Experimental Medicine Faculty of Medicine, Imperial College London). Further considerations for quality control of the samples are determined bioinformatically (explained later on). The details on the samples can be seen in Table 2-1.

Sample RNA id	Sample ATAC id	MM or PC	MM subgroup (or PC)	PC id	PC CD19 status	RNA batch	ATAC batch	Presence in analyses
RS_1.10	A24.10	MM	NA	NA	NA	pool1	pool24_pool27	MM-PC, MOFA
RS_1.11	A26.6B	MM	HD	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.12	A26.8	MM	CCND1	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.13	A26.9B	MM	HD	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.14	A26.10B	MM	MMSET	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.15	A26.11	MM	MAF	NA	NA	pool1	pool26_pool28_ pool28_RE	MM-PC, Subgroup, MOFA
RS_1.16	A26.12	MM	MAF	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.17	A27.12	MM	HD	NA	NA	pool1	pool27	MM-PC, Subgroup, MOFA

Sample RNA id	Sample ATAC id	MM or PC	MM subgroup (or PC)	PC id	PC CD19 status	RNA batch	ATAC batch	Presence in analyses
RS_1.18	A26.13	MM	HD	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.19	A26.14	MM	CCND1	NA	NA	pool1	pool26_pool28	MM-PC, Subgroup, MOFA
RS_1.2	A19.1	MM	HD	NA	NA	pool1	pool19	MM-PC, Subgroup, MOFA
RS_1.20	A26.15B	MM	HD	NA	NA	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.23	A26.19	PC	PC	160617	-	pool1	pool26	MM-PC, Subgroup, MOFA
RS_1.3	A19.8	MM	MMSET	NA	NA	pool1	pool19	MM-PC, Subgroup, MOFA
RS_1.4	A19.2	MM	CCND1	NA	NA	pool1	pool19	MM-PC, Subgroup, MOFA
RS_1.5	A24.7	MM	HD	NA	NA	pool1	pool24	MM-PC, Subgroup, MOFA
RS_1.7	A19.5	MM	HD	NA	NA	pool1	pool19	MM-PC, Subgroup, MOFA
RS_1.8	A19.6	MM	NA	NA	NA	pool1	pool19	MM-PC, MOFA
RS_1.9	A24.8	MM	HD	NA	NA	pool1	pool24_pool26	MM-PC, Subgroup, MOFA
RS_2.1	A17.5	MM	NA	NA	NA	pool2	pool17_pool26	MM-PC, MOFA
RS_2.3	A17.9	MM	HD	NA	NA	pool2	pool17	MM-PC, Subgroup, MOFA
RS_2.4	A24.11	MM	HD	NA	NA	pool2	pool24_pool26	MM-PC, Subgroup, MOFA
RS_2.5	A27.18	MM	NA	NA	NA	pool2	pool27	MM-PC, MOFA
RS_2.6	A27.19	MM	MMSET	NA	NA	pool2	pool27_pool28_ pool28_RE	MM-PC, Subgroup, MOFA
RS_2.7	A27.20	MM	NA	NA	NA	pool2	pool27	MM-PC, MOFA

Sample RNA id	Sample ATAC id	MM or PC	MM subgroup (or PC)	PC id	PC CD19 status	RNA batch	ATAC batch	Presence in analyses
RS_2.8	A27.21	PC	PC	271017	+	pool2	pool27	MM-PC, Subgroup, MOFA
RS_2.9	A27.22	PC	PC	271017	-	pool2	pool27	MM-PC, Subgroup, MOFA
RS_3B.1	A26.1	MM	HD	NA	NA	pool3	pool26	MM-PC, Subgroup, MOFA
RS_3B.2	A24.4	MM	HD	NA	NA	pool3	pool24	MM-PC, Subgroup, MOFA
RS_3B.3	A28.13	MM	MMSET	NA	NA	pool3	pool28	MM-PC, Subgroup, MOFA
RS_3B.4	A28.15	MM	CCND1	NA	NA	pool3	pool28_pool28_RE	MM-PC, Subgroup, MOFA
RS_4.25	A28c.14	PC	PC	230218	+	pool4	pool28_RE	MM-PC, Subgroup, MOFA
RS_1.22	A26.18.A26.20	PC	PC	160617	+	pool1	pool26	MM-PC, Subgroup, MOFA
RS_3B.11	A28.7	MM CL	OPM2 t(4;14) MMSET	NA	NA	pool3	pool28	MM-PC, Subgroup
RS_4.19	A28c.3	MM CL	MM1S t(14;16) MAF	NA	NA	pool4	pool28_RE	MM-PC, Subgroup
RS_4.9	A28c.5	MM CL	KMS12BM t(11;14) CCND1	NA	NA	pool4	pool28_RE	MM-PC, Subgroup
RS_4.16	A28c.6	MM CL	U266 - t(11;14) CCND1	NA	NA	pool4	pool28_RE	MM-PC, Subgroup
RSJJN3.1	AJJN3.1	MM CL	JJN3 t(14;16) MAF	NA	NA	pool3_pool4	pool28_pool28_RE	MM-PC, Subgroup

Table 2-1: All samples used in the analysis.

MM: Primary Multiple Myeloma, PC: Normal Donor, CL: Cell Line. MM subgroup (or PC): PC or MM subgroup based on primary oncogenic event [HD: Hyperdiploid, IgH translocation partners: CCND1: t(11;14) IGH/CCND1, HD: Hyperdiploid, MMSET: t(4;14) MMSET/IGH, MAF: t(14;16) IGH/MAF. NA: cytogenetic information not available]. PC id: Normal donor patient id. PC CD19 status: CD19 receptor status on the cell surface of NDs, positive (+) or negative

(-). Presence in analyses: whether the sample is used for the different analyses presented in this thesis: MM-PC (chapter 3), Subgroup (chapter 4) and MOFA (chapter 5).

2.2. General considerations for RNA-seq and ATAC-seq analysis

2.2.1. DESeq2 settings

Differential and more accessible/expressed regions and genes were obtained using DESeq2 version 1.18.1 and R. Unless otherwise noted default settings were used. In particular, count outliers were filtered using cook's distance for RNA-seq analyses (removal of outliers using a Cook's distance cutoff was disabled for ATAC-seq related analyses). Additionally, regions and genes were independently filtered for counts using mean counts across samples using an automatically chosen threshold, logFC changes are tested against a null hypothesis of no change, and p-values were corrected using the Benjamini-Hochberg procedure.

2.2.2. Tests performed and thresholds used to obtain significant regions and genes

Tests for differential regions and genes were performed using Deseq2 with the settings specified in the DESeq2 settings, section 2.2.1.

For the ATAC-seq and RNA-seq MM vs. PC analysis (primary samples), samples belonging to batches with only one sample were placed on the same meta batch. The following three tests were performed on the quantification matrix using the different covariates for each sample: condition, cd19, batch and donor id:

- MM vs average CD19 (PC)
- MM vs average donor (PC)
- MM vs PC Log Ratio Test (LRT) accounting for batch.

For the ATAC-seq and RNA-seq subgroup MM vs. PC analysis (primary samples), samples belonging to batches with only one sample were placed on the same meta batch. Only a MM subgroup vs. PC Log Ratio Test (LRT) accounting for batch was performed on the quantification matrix since the CD19 status and donor id are confounding.

For the DE genes between MM cell lines vs. PC primary samples analysis, batch effects couldn't be taken into account since there were very few overlaps in batches between the PC and MM CL samples and only a Wald test MM vs. PC was done. Thus MM CLs were used only in a qualitative way.

Similarly, for the DE genes between subgroup MM cell lines vs. PC primary samples analysis, a Wald test MM subgroup vs. PC was done, there were no HD cell lines.

The thresholds used to consider DE genes on all of the above tests were corrected p-value less than 0.05 in conjunction with absolute log_2FoldChange greater or equal to 1.5. In cases where multiple tests were performed, only results having met the specified criteria on all tests are kept. OE genes were obtained from DE genes with MM or MM subgroup log_2FoldChange over PC greater or equal to 1.5.

The thresholds used to consider DA regions were corrected p-value less than 0.05 in conjunction with absolute log_2FoldChange greater or equal to 1 on all tests. Over accessible regions are obtained from DA regions with MM or MM subgroup log_2FoldChange over PC greater or equal to 2.

In cases where samples were collapsed in the quantification matrix, the collapsed sample counts are added together, unless otherwise specified, the function collapseReplicates from the Deseq2 package is used.

2.2.3. Regularized log counts

Unless specified so, in cases where rLog or regularized log is mentioned, it was calculated using the rlog function from the Deseq2 package (blind parameter set to “True”).

2.2.4. Removing batch effects from samples

Unless specified so, in cases where batch effects are removed, the removeBatchEffect function from the R limma package (Ritchie et al., 2015) was used.

2.2.5. Reference genome and annotations

Unless specified so, the genome version used was hg38 with no alternative contigs, for the annotations, Ensembl version 85 was used.

2.3. RNA-seq analysis

2.3.1. Read cleaning and filtering

Paired-end RNA-seq reads which had already been adapter trimmed by the DNA sequencing facility at Centre for Haematology, Division of Experimental Medicine Faculty of Medicine, Imperial College London were processed with FastQC version 0.11.3 to confirm good quality reads. No further quality processing was necessary.

2.3.2. RNA-seq quantification

The paired-end reads for each sample were used with the pipeline_rnaseqdiffexpression pipeline from CGATPipelines (CGAT_Developers, 2018) and Salmon version 0.11.4 (Patro et al., 2017) to obtain an estimate of how many reads were mapping to each gene quantified (including all the associated transcripts). Salmon was used with fragment GC bias correction, 100 bootstrap samples and using an auxiliary k-mer hash over k-mers of length 31. A reference geneset from the reference genome was produced for this task. This process was taking into account the nature of the reads (uniquely mapping or multimapping to each transcript) and the relative abundance estimate for each transcript. For each sample, the number of reads for each gene were rounded to the nearest integer and a table was created with the integer number of reads for each gene in the reference geneset for each sample.

2.3.3. RNA-seq mapping

Pipeline_mapping pipeline from CGATPipelines was used to map the RNA-seq reads. Briefly, a reference geneset was created starting with the geneset from the human genome and filtering mitochondrial and non-standard chromosomes, removing long (>2Mb) and very short (<5bp) introns, ribosomal RNA. From this geneset, only the protein coding transcripts were retained and the splice junctions were curated. These known splice junctions and paired-end reads from each sample were supplied to Hisat version 0.1.6 (D. Kim et al., 2015) which maps the reads using index for the reference genome. These mapped files were used for RNA-seq signals shown throughout this thesis.

2.3.4. RNA-seq quality control statistics

The following statistics were generated as part of the analysis:

- *RNA Starting read pairs*: Reads pairs obtained.
- *RNA Mapping rate (Salmon)*: Percentage of the read pairs quantified in a gene transcript “RNA-seq quantification”.
- *RNA Mapped read pairs*: Calculated by multiplying starting read pairs by mapping rate.

2.3.5. Obtaining annotated and unannotated Transcription Start Sites

Unannotated TSS present in the PC and MM samples (primary and cell lines) were obtained using the pipeline for the detection of alternative polyadenylation (Sudbery, 2019a) developed by Dr. Ian Sudbery. Briefly, Stringtie version 1.2.3 (Pertea et al., 2015) was used with each sample’s RNA-seq mappings (RNA-seq mapping, section 2.3.3) in conjunction with the geneset from the human genome to generate cases of novel transcripts with alternate exon usage.

From all the novel transcripts, the ones that are one exon long were removed from this list (assuming they were eRNA transcripts). These novel transcripts were transformed using GNU Awk version 3.1.7 to obtain the 1bp long region representing their strand-aware five prime end. To obtain the promoter regions from these transformed novel transcripts, the strand-aware five prime end was extended 2kb upstream of the TSS, this way the TATA box, proximal and distal promoter were likely to be captured. In addition, they were extended 100bp downstream to cover the TSS site; these are referred to as unannotated promoter sites. This was done using Bedtools version 2.22.1 (Quinlan and Hall, 2010).

To obtain the annotated TSS, the annotations for the human genome were used. The TSS for the coding and non-coding genes were obtained and transformed in the same way as the unannotated TSS: the strand-aware five prime end was extended 2kb upstream and 100bp downstream of the TSS (strand-aware), these were referred to as annotated promoter sites.

The unannotated and annotated promoter sites were merged into one file using Bedtools merge.

2.3.6. DE and OE genes

2.3.6.1. Obtaining DEMM genes

The table with the quantified reads in each gene for each sample (Table 2-1 "MM-PC" category) was produced as explained in the RNA-seq quantification, section 2.3.2. The gene quantification table was inputted using the DESeq2 settings and test schemes outlined in General considerations for RNA-seq and ATAC-seq analysis, section 2.2 for RNA-seq MM vs. PC analysis (primary samples), DEMM genes were obtained.

2.3.6.2. Obtaining DESMM genes

The table with the quantified reads in each gene for each sample (Table 2-1 "Subgroup" category) was produced as explained in the RNA-seq quantification, section 2.3.2. The gene quantification table was inputted using the DESeq2 settings and test schemes outlined in General considerations for RNA-seq and ATAC-seq analysis, section 2.2 for RNA-seq subgroup MM vs. PC analysis (primary samples), DESMM genes were obtained.

2.3.6.3. Obtaining MM vs. PC DE genes between MM cell lines vs. PC primary samples

The table with the quantified reads in each gene for each sample PC and MM CL (Table 2-1) was produced as explained in the RNA-seq quantification, section 2.3.2. Sample RSJJN3.1 is formed of two technical replicates: RS_3B.10, RS_4.3, the replicates were collapsed. The gene quantification table was inputted using the DESeq2 settings and test schemes outlined in

General considerations for RNA-seq and ATAC-seq analysis, section 2.2 for DE genes between MM cell lines vs. PC primary samples analysis.

2.3.6.4. Obtaining subgroup MM vs. PC DE genes between MM cell lines vs. PC primary samples

The table with the quantified reads in each gene for each sample PC and MM CL (Table 2-1) was produced as explained in Obtaining MM vs. PC DE genes between MM cell lines vs. PC primary samples, section 2.3.6.3. The gene quantification table was inputted using the DESeq2 settings and test schemes outlined in General considerations for RNA-seq and ATAC-seq analysis, section 2.2, for DE genes between subgroup MM cell lines vs. PC primary samples analysis. There are no HD cell lines, so tables were created only for CCND1 vs. PC, MAF vs. PC and MMSET vs. PC.

2.3.6.5. Obtaining OEMM genes

OEMM genes were produced from DEMM genes as specified in General considerations for RNA-seq and ATAC-seq analysis, section 2.2.

2.3.6.6. OESMM genes

OESMM genes were produced from DESMM genes as specified in General considerations for RNA-seq and ATAC-seq analysis, section 2.2 (genes significantly OE in any MM subgroup vs. PC are filtered).

2.4. ATAC-seq analysis

2.4.1. Read cleaning and filtering

Raw paired-end ATAC-seq reads were processed with FastQC version 0.11.3 (Andrews, 2010) to confirm good quality reads (Figure 2-1). ATAC-seq adapters were removed, uncalled bases (N's) on ends of reads were trimmed using Cutadapt version 1.9.1 (Marcel, 2011). Only read pairs with both single end fragments remaining were kept. A second quality pass was performed using Sickle version 1.33 (Joshi and Fass, 2011), trimming was performed with a sliding window average Phred quality threshold of 30 (without five prime trimming). Only paired-end reads with each single end containing a minimum length of 20 bases of length were allowed. All the read filtering was done via the pipeline_readqc pipeline from CGATPipelines (CGAT_Developers, 2018).

After the quality control, metrics were obtained through the pipeline_readqc for the number of reads remaining after each quality step filtering and the total number dropped since the start of the process.

2.4.2. Mapping and calling chromatin accessible peaks

The remaining paired-end reads were mapped using Bowtie2 version 2.3.0 (Langmead and Salzberg, 2012) in paired-end mode to the human genome reporting up to 4 alignments per read, with maximum fragment length 2000bp (Figure 2-1). This was done within the context of the pipeline_mapping pipeline from CGATPipelines (CGAT_Developers, 2018).

With the help of some ideas from Dr. Ian Sudbery, I created pipeline_atacseq (Alvarez-Benayas, 2019) based on the ENCODE pipeline and recommendations for ATAC-seq processing (Dunham et al., 2012; Kundaje, 2019a) in order to process the mapped reads to produce peaks in signals and reads in peaks per sample (Figure 2-1). It uses the CGATPipelines framework.

Briefly the pipeline allows only correctly mapped read pairs. Reads removed using Samtools version 1.3.1 (Li et al., 2009) include unmapped (read pair or one of the reads), reads failing platform, orphan reads (one of the reads in the pair removed), read pairs mapping to different chromosomes (-F 524 -f 2 Samtools flags). Read pairs which were in the wrong orientation and contained no overlap were removed with Samtools and Bedtools. Mapped read pairs with more than single mapping sites were also removed. These last type of reads were removed with the script in (Kundaje, 2016).

For the remaining read pairs, duplicates were marked using Picard Markduplicates version 1.135 (Broad_Institute, 2018). The remaining read pairs were deduplicated and indexed using Samtools.

Each deduplicated read pair was converted into two single end tags (one for each end of the fragment) using bedtools and GNU Awk version 3.1.7 (Arnold, 2011). Any single end tag in any mitochondrial or not standard chromosome was eliminated with the use of GNU grep version 2.20 (Meyering, n.d.). The TN5 enzyme used in Atac-seq introduces a few base pairs on the 5' sites on each strand and these were removed from the remaining tags using GNU Awk, also any 5' trimming from the quality control was also extended on the 5' start sites in a strand-aware manner. The resulting elements are referred to as shifted tags (Figure 2-2).

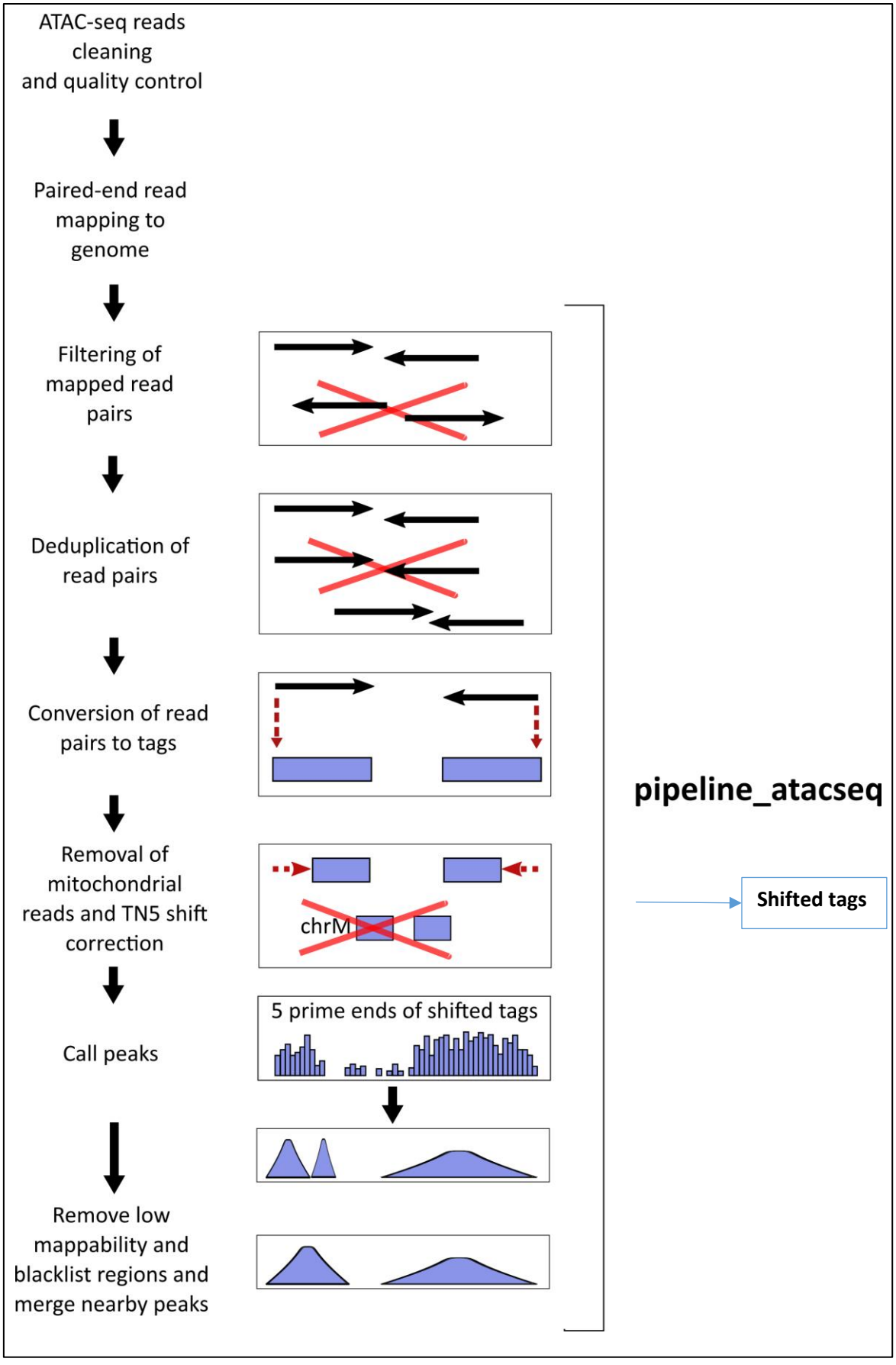


Figure 2-1: Bioinformatic process of ATAC-seq data processing.

The ATAC-seq paired reads are subject to quality control processing where low quality bases and N bases are removed, they are then mapped to the genome and filtered to remove incorrectly mapped pairs (for example, pairs mapping in the wrong orientation). Properly mapped pairs are then deduplicated and single end tags were created from each unique pair. Reads corresponding to the mitochondrial chromosome were removed and the remaining reads were applied TN5 shifting (the TN5 transposase introduces a few base pairs in the DNA cutting site) and any 5' removed bases during quality control were also accounted for. The remaining reads were used for peak-calling and blacklisted chromosomal regions based on known mappability issues were removed from the resulting peaks.

Broad and narrow peaks were called on each sample using MACS2 version 2.1.1.20160309 (Zhang et al., 2008) with the following options for narrow peaks:

-g hs -q 0.01 --nomodel --shift -100 --extsize 200 -B --SPMR --keep-dup all --call-summits

And for broad peaks:

-g hs -q 0.01 --nomodel --shift -100 --extsize 200 --broad --broad-cutoff 0.01 --keep-dup all

Narrow and broad peaks were filtered to remove areas of low mappability (Hoffman et al., 2013) downloaded from (Dunham et al., 2012; ENCODE_UCSC, 2011) after performing a coordinate "lift over" from hg19 to hg38 (Dunham et al., 2012; UCSC, 2019). ENCODE blacklist regions (Dunham et al., 2012) (ENCODE, n.d.) were also removed.

The narrow peaks and broad peaks produced per sample were merged, also any areas of 100 base pairs or less between peaks per sample are also considered peaks. This was also done using bedtools. These are referred to as sample peaks from here onwards.

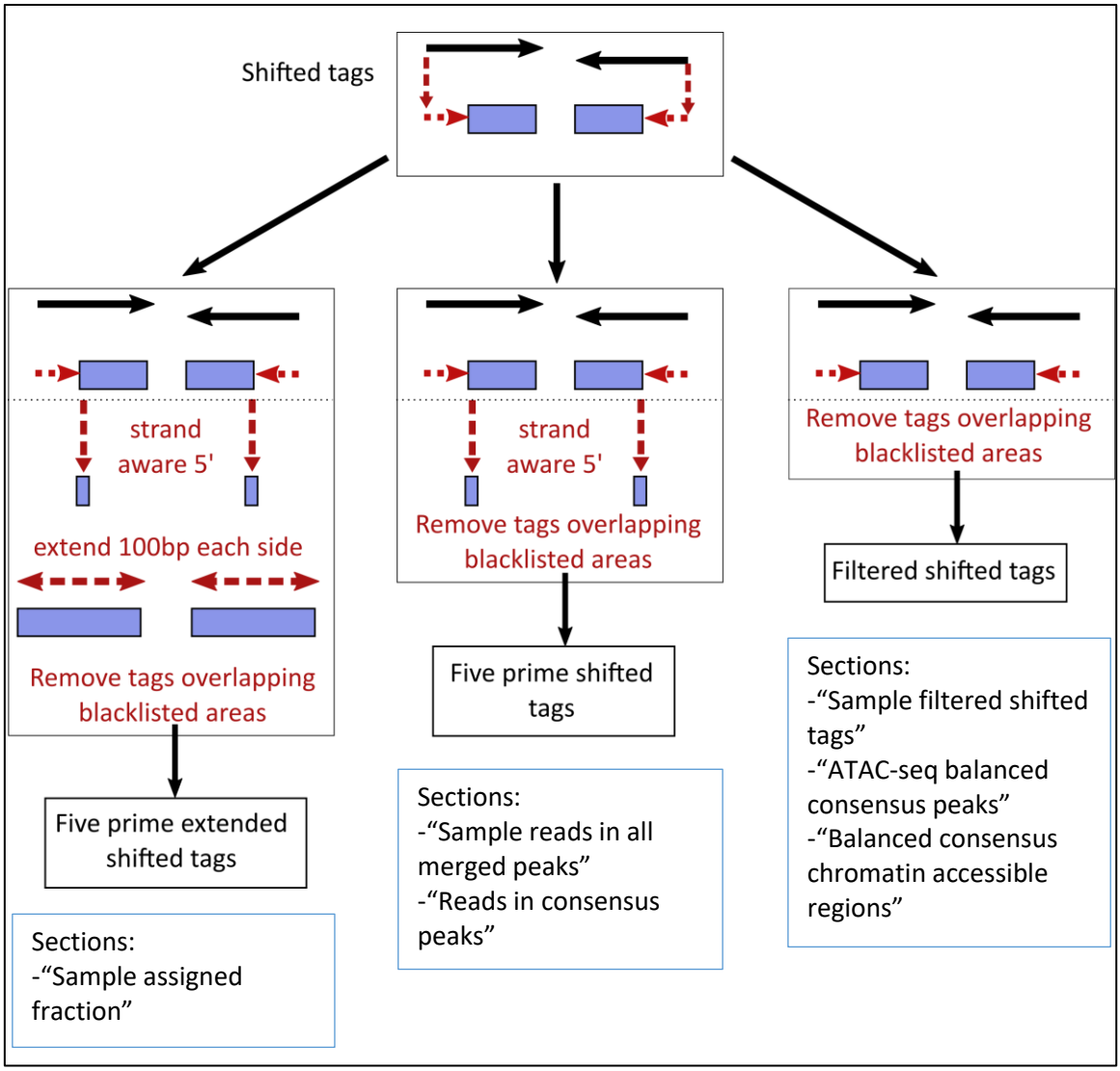


Figure 2-2: Types of ATAC-seq shifted reads.

Shifted tags are the result of deduplicated, correctly mapped read pairs being converted to individual single ends after applying TN5 and quality control shifting (Figure 2-1 before calling peaks).

Five prime extended shifted tags are shifted tags where the 5' end representing in each read in a pair (from the sequenced read fragment), is extended 100bp up and downstream.

Five prime shifted tags are shifted tags where only the 5' end representing in each read in a pair are taken into account.

In all of the above, the tags overlapping blacklisted areas (due to mappability issues) are removed, in the case of shifted tags, the resulting tags are called "filtered shifted tags".

2.4.3. Sample assigned fraction

The sample assigned fraction reflects the proportion of the total reads that are mapped to areas of high accessibility compared with background noise. To calculate it, sample shifted tags were transformed extending their five prime end 100bp upstream and then extending 200bp downstream (by the same values as with the peak calling). They were filtered to remove tags overlapping with areas of low and high mappability defined previously. The extended tags are referred to as five prime extended shifted tags (Figure 2-2). The sample assigned fraction is calculated in the following way:

$$\frac{\text{Five prime extended shifted tags overlapping merged sample peaks}}{\text{Total five prime extended shifted tags}}$$

2.4.4. Sample reads in all merged peaks

Sample shifted tags were also transformed to obtain only the 1bp long segment representing their strand-aware five prime end. They were then filtered to remove tags overlapping with areas of low and high mappability defined previously. They are referred to as five prime shifted tags (Figure 2-2). For each sample, the number of five prime shifted tags overlapping each common peak (merged peaks from all samples) is reported using Bedtools intersect. A table was created combining this data for all the samples and common peaks.

2.4.5. Sample filtered shifted tags

Sample shifted tags (Figure 2-2) were filtered to remove tags overlapping with areas of low and high mappability defined previously (referred to as filtered shifted tags).

2.4.6. ATAC-seq balanced consensus peaks

I created pipeline_atac_consensus_balanced_peaks (Alvarez-Benayas, 2020b) based on pipeline_atacseq in order to create a set of consensus peaks starting with filtered shifted tags from different samples (using the same number of filtered shifted tags for all samples).

First the number of filtered shifted tags per sample (Figure 2-2) is obtained, the minimum of all the samples is calculated. For each clinical sample, a random sample is extracted with the minimum number of filtered shifted tags, all the sample filtered shifted tags are pooled together and peaks are called, filtered for excluded regions and merged in the same manner as with pipeline_atacseq in Mapping and calling chromatin accessible peaks, section 2.4.2.

2.4.7. Balanced consensus chromatin accessible regions

Sample filtered shifted tags (Figure 2-2) were produced from each sample used, using the strategy explained in ATAC-seq balanced consensus peaks, section 2.4.6, first the minimum number of sample filtered shifted tags was obtained for all samples and samples were down-sampled to this level. Then filtered shifted tags were pooled separately for each of sample group specified (PC and MM or PC and each of the MM subgroups). For the PC pool, since samples A26.20 and A26.18 are biological replicates, the sample filtered shifted tags coming from each sample in the pool is half the minimum sample number of sample filtered shifted tags, considering both samples as one. Starting with the pooled, downsampled, filtered, shifted tags, the chromatin accessible peaks were called and processed as specified in Mapping and calling chromatin accessible peaks, section 2.4.2, for each pool (Figure 2-1), leaving a set of consensus peaks for each sample group.

The balanced consensus peaks found for each group were merged (any regions within two peaks of 200bp or less are also considered peak regions).

2.4.7.1. Balanced consensus chromatin accessible peaks for PC and MM

The consensus peaks for PC and MM were calculated as specified in Balanced consensus chromatin accessible regions, section 2.4.7, using the samples in Table 2-1 “MM-PC” category. These are referred to as consensus peaks for PC and MM.

2.4.7.2. Balanced consensus chromatin accessible peaks for primary PC and MM subgroups

The consensus peaks for PC and MM subgroups were calculated as specified in Balanced consensus chromatin accessible regions, section 2.4.7, using the samples in Table 2-1 “Subgroup” category. These are referred to as consensus peaks for PC and MM subgroups.

2.4.8. Annotations of the consensus peak regions

The balanced consensus chromatin accessible peaks for primary PC and MM subgroups were obtained as specified in Balanced consensus chromatin accessible peaks for primary PC and MM subgroups, section 2.4.7.2 and intersected using Bedtools independently with the different MM subgroup (CCND1, HD, MAF and MMSET) and PC balanced consensus peaks. An overlap of 10% of either the balanced consensus peaks or the subgroup peaks is required for a positive overlap result. Regions were annotated as specified in Genomic annotations of regions, section 2.7.

2.4.9. Reads in consensus peaks

Using all five prime shifted tags (Figure 2-2) from the samples selected, the number of tags for each sample in each consensus peak from the selected set was calculated as specified in Sample reads in all merged peaks, section 2.4.4. For the PC replicate (A26.20 and A26.18), the tag counts for each peak were added using the function collapseReplicates from the Deseq2.

2.4.9.1. Getting the sample five prime shifted tag counts in each consensus peak for PC and MM

The method specified in Reads in consensus peaks, section 2.4.9, was used to obtain the reads from the samples in Table 2-1 ("MM-PC" category) in each peak from the consensus peaks for PC and MM (see Balanced consensus chromatin accessible peaks for primary PC and MM subgroups, section 2.4.7.2).

2.4.9.2. Getting the sample five prime shifted tags in each consensus peak for PC and MM subgroups

The method specified in Reads in consensus peaks, section 2.4.9, was used to obtain the reads from the samples in Table 2-1 ("Subgroup" category) in each peak from the consensus peaks for PC and MM subgroups (see Balanced consensus chromatin accessible peaks for primary PC and MM subgroups, section 2.4.7.2).

2.4.10. ATAC-seq quality control statistics

The following statistics are generated as part of the analysis:

- ATAC Starting read pairs: Raw reads pairs sequenced.
- ATAC After QC read pairs: Read pairs after read cleaning and filtering (see Figure 2-1).
- ATAC % read pairs dropped start to After QC.
- ATAC Aligned read pairs (Bowtie2): Aligned pairs after mapping (see "Mapping and calling chromatin accessible peaks")
- ATAC % reads pairs aligned from After QC: $\frac{\text{ATAC Aligned read pairs (Bowtie2)}}{\text{ATAC After QC read pairs}}$
- ATAC Reads pairs examined (MARK DUPLICATES): Correctly mapped read pairs inputted to Mark Duplicates version: 1.135 (Broad_Institute, 2018) after filtering as specified in Mapping and calling chromatin accessible peaks, section 2.4.2.
- ATAC Unpaired reads examined (MARK DUPLICATES): Quality control, should always be 0 as only correctly mapped read pairs are inputted to Mark Duplicates.
- ATAC Unpaired read duplicates (MARK DUPLICATES): Quality control, should always be 0 as only correctly mapped read pairs are inputted to Mark Duplicates.

- ATAC Unmapped reads (MARK DUPLICATES): Quality control, should always be 0 as only correctly mapped read pairs are inputted to Mark Duplicates.
- ATAC Read pairs marked as duplicates (MARK DUPLICATES): Number of paired end reads marked as duplicates (counts all duplicates regardless of source).
- ATAC Read pairs duplicates that were caused by optical (machine) duplication (MARK DUPLICATES): A particular type of duplicate from all duplicates caused by the sequencing machine.
- ATAC Fraction duplication: Obtained by Mark Duplicates $\frac{\text{Read pairs marked as duplicates}}{\text{Reads pairs examined}}$
- ATAC Estimated library size: Obtained by Mark Duplicates. Estimated number of total paired end fragments (at saturation).

ENCODE estimated library complexity metrics (Kundaje, 2019b) starting from correctly mapped read pairs after filtering as specified in Mapping and calling chromatin accessible peaks, section 2.4.2:

- ATAC Total read pairs: Total number of read pairs in the starting file (included duplicated).
- ATAC Distinct read pairs: Total unique read pairs, repeated fragments count only once.
- ATAC One read pair: Total number of read pairs that only appear once (do not have duplications).
- ATAC Two read pairs: Total number of read pairs that only appear twice (have one duplication).
- ATAC NRF=Distinct/Total: Non Redundant Fraction: $\frac{\text{Distinct read pairs}}{\text{Total read pairs}}$
- ATAC PBC1=OnePair/Distinct: Polymerase chain reaction (PCR) Bottleneck coefficient 1 $\frac{\text{One read pair}}{\text{Distinct read pairs}}$
- ATAC PBC2=OnePair/TwoPair: Polymerase chain reaction (PCR) Bottleneck coefficient 2 $\frac{\text{One read pair}}{\text{Two read pairs}}$
- ATAC Total read pairs in proper pairs after deduplication (FLAGSTATS): Total deduplicated read pairs after deduplication obtained by Samtools.
- ATAC Final number of single ends inputted to the peak caller: Number of single ends after filtering deduplicated read pairs removing any fragment in mitochondrial or not standard chromosomes.

MACS2 peak calls:

- ATAC Narrow pre-processing (MACS2)
- ATAC Broad pre-processing (MACS2)
- ATAC Narrow post-processing (MACS2)
- ATAC Broad post-processing (MACS2)

Other:

- ATAC Assigned fraction %: Calculated as explained in Sample assigned fraction, section 2.4.3.
- ATAC Single ends in peaks: Calculated by multiplying the assigned fraction by the number of single ends inputted to the peak caller.
- ATAC Drop % in mapped - filtered (ATAC-seq pipeline): Filtered out paired end reads between mapping and producing correctly mapped read pairs as explained in Mapping and calling chromatin accessible peaks, section 2.4.2, as a fraction of the total mapped read pairs.
- ATAC Drop % in mapped - input read PAIRS peak caller (ATAC-seq pipeline): Filtered out paired end reads between mapping and entering the peak caller producing correctly mapped read pairs as explained in Mapping and calling chromatin accessible peaks, section 2.4.2, as a fraction of the total mapped read pairs.

2.4.11. Obtaining PC accessible peaks

Sample peaks from all PC samples (Table 2-1 “PC” samples) were obtained as specified in Mapping and calling chromatin accessible peaks, section 2.4.2, considering the sample A26.18.A26.20 as two individual samples. The resulting peaks were merged (any regions within two peaks of 200bp or less were also considered peak regions). The resulting regions are referred to as “PC accessible peaks” (Figure 2-3).

2.4.12. Obtaining cell line chromatin accessible peaks

Sample peaks for all individual MM cell lines samples (Table 2-1 “MM CL” samples) were obtained as specified in Mapping and calling chromatin accessible peaks, section 2.4.2.

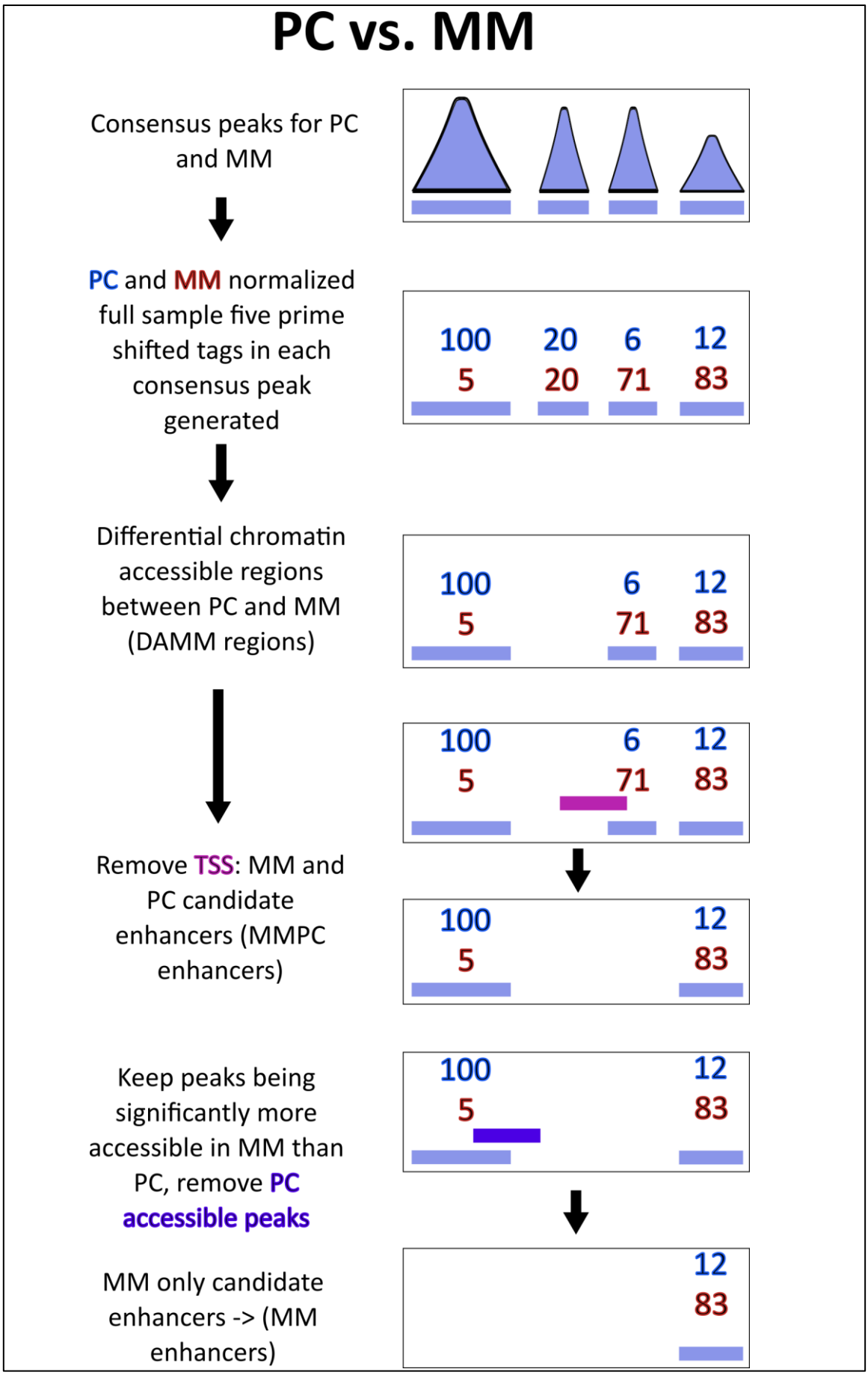


Figure 2-3: Chromatin accessible region types for MM vs. PC analysis.

Consensus peaks for PC and MM are obtained and using the full sample reads in peaks, DAMM regions are found through differential analysis. From these regions, annotated and unannotated TSS (except 1 exon TSS generating transcripts) are removed to obtain MMPC enhancers. From

these, MM enhancers are obtained by keeping regions with significantly more accessible chromatin in MM and removing regions overlapping PC accessible peaks.

2.4.13. DA regions

2.4.13.1. Obtaining DAMM regions

The table with the reads in PC and MM consensus peaks was obtained as specified in Getting the sample five prime shifted tag counts in each consensus peak for PC and MM, section 2.4.9.1. The reads in peaks quantification table was inputted using the DESeq2 settings and test schemes outlined in General considerations for RNA-seq and ATAC-seq analysis, section 2.2, for ATAC-seq MM vs. PC analysis (primary samples), DAMM regions were obtained (Figure 2-3).

2.4.13.2. Obtaining DASMM regions

The table with the reads in PC and MM subgroups consensus peaks was obtained as specified in Getting the sample five prime shifted tags in each consensus peak for PC and MM subgroups, section 2.4.9.2. The reads in peaks quantification table was inputted using the DESeq2 settings and test schemes outlined in General considerations for RNA-seq and ATAC-seq analysis, section 2.2 for ATAC-seq subgroup MM vs. PC analysis (primary samples), DASMM regions were obtained (Figure 2-4).

2.4.14. Candidate enhancer sets

2.4.14.1. Obtaining MMPC enhancers

The DAMM regions that overlap the unannotated and annotated promoter sites (obtained as explained in Obtaining annotated and unannotated Transcription Start Sites, section 2.3.5) were filtered out using Bedtools. These regions are referred to as MMPC enhancers (Figure 2-3).

2.4.14.2. Obtaining MM enhancers

MM enhancers were produced from MMPC enhancers as specified in General considerations for RNA-seq and ATAC-seq analysis, section 2.2, applying thresholds for over accessible regions. Additionally, using Bedtools, only regions, which were not overlapping any PC accessible peaks (Obtaining PC accessible peaks, section 2.4.11), were maintained. The remaining regions are referred as MM enhancers (Figure 2-3).

2.4.14.3. Obtaining DASMM enhancers

Starting with the DASMM regions (Obtaining DASMM regions, section 2.4.13.2), annotated and unannotated Transcription Start Sites were removed using Bedtools. These regions are referred to as DASMM enhancers (Figure 2-4).

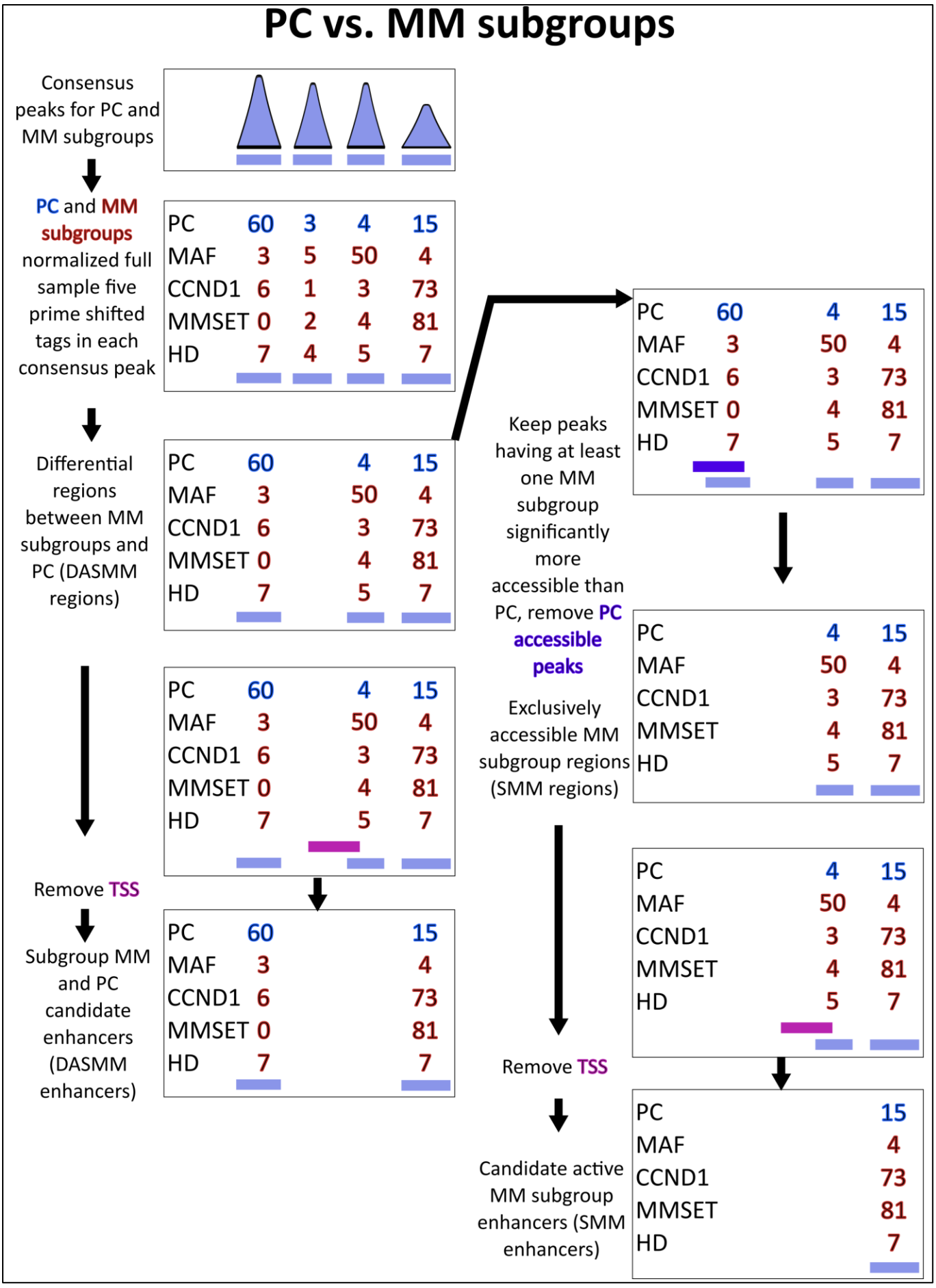


Figure 2-4: Chromatin accessible region types for MM subgroups vs. PC analysis.

Consensus peaks for PC and MM subgroups are obtained and through differential analysis of the reads in peaks, DASMM regions are attained. Annotated and unannotated TSS of more than one exon are removed from these regions to obtain DASMM enhancers. Additionally, DASMM regions overlapping PC accessible peaks are removed to obtain SMM regions. By also removing TSS from SMM regions, SMM enhancers are obtained.

2.4.14.4. Obtaining SMM regions

SMM regions were produced from DASMM regions as specified in General considerations for RNA-seq and ATAC-seq analysis, section 2.2, applying thresholds for over accessible regions and maintaining regions with at least one MM subgroup vs PC over accessible. The ATAC DE subgroup vs. PC columns were updated to reflect whether the subgroup is DA with respect to PC in terms of the log_2FoldChange with these new thresholds. Using Bedtools, only regions, which were not overlapping any PC accessible peaks (see Obtaining PC accessible peaks, section 2.4.11), were maintained. The regions are referred as "SMM regions" (Figure 2-4). These regions were annotated following the same method specified in Annotations of the consensus peak regions, section 2.4.8.

2.4.14.5. Obtaining SMM enhancers

Starting with the SMM regions, annotated and unannotated Transcription Start Sites were removed using Bedtools (Figure 2-4). The remaining regions are referred to as "SMM enhancers".

2.5. Integrated ATAC and RNA analysis

2.5.1. Relating candidate enhancer regions with altered expression genes

2.5.1.1. Linking candidate enhancers to target protein coding genes

The annotated protein coding genes' TSS were obtained from the Geneset annotations for the human genome. The TSS were strand-aware extended to cover the promoters: 2kb upstream (to overlap the TATA box, proximal and distal promoter) and 100bp downstream of the TSS (to cover the TSS) using Bedtools. Each candidate enhancer from the set used was then extended 1Mb upstream and downstream (using Bedtools) and the promoters of all protein coding genes within that area were related (Figure 2-5 black and dark yellow boxes).

Ensembl gene ids were converted to symbols and added gene function descriptions using the AnnotationDbi R package (Pages et al., 2010).

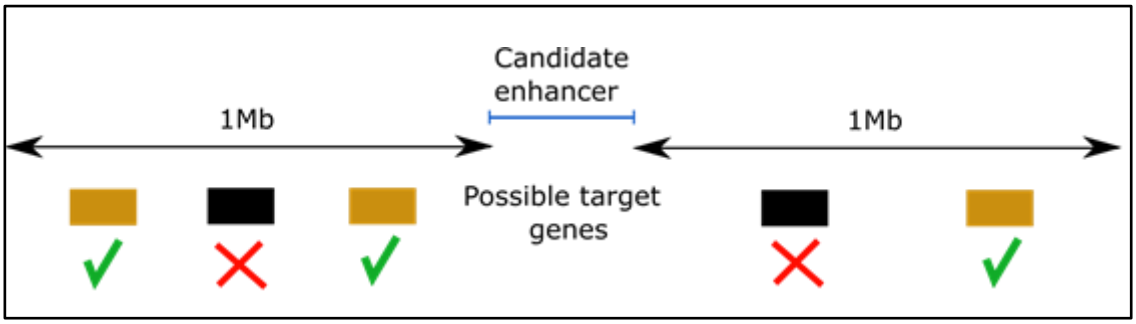


Figure 2-5: Criteria for enhancer - promoter interactions.

Interactions are determined on the basis of candidate enhancer regions within 1Mb of genes with altered expression in the malignant compared with the healthy state. Black boxes: genes with no altered expression. Dark yellow boxes: genes with altered expression.

2.5.1.2. Obtaining MMPC enhancers near DEMM protein coding genes

The interactions between MMPC enhancers and all protein coding genes were determined as specified in Linking candidate enhancers to target protein coding genes, section 2.5.1.1. From these interactions, only interactions containing DEMM genes were kept. Information on whether each region – gene pair contains a MM CL vs. PC DE gene (see Obtaining MM vs. PC DE genes between MM cell lines vs. PC primary samples, section 2.3.6.3) or overlaps a MM cell line chromatin accessible peak, was added to the interactions (overlapping with Bedtools the regions obtained in Obtaining cell line chromatin accessible peaks, section 2.4.12). These interactions are referred to as MMPC enhancers near DEMM protein coding genes (Figure 2-5 dark yellow boxes).

2.5.1.3. Obtaining MM enhancers regulating OEMM protein coding genes

From the table produced in Obtaining MMPC enhancers near DEMM protein coding genes, section 2.5.1.2, interactions were only maintained if both the region and gene complied with the over accessible and OE criteria respectively explained in Tests performed and thresholds used to obtain significant regions and genes, section 2.2.2. Using the guidelines explained in this section, the remaining interactions were annotated with information as to whether the gene involved are over-expressed in cell lines.

Finally, using Bedtools, information regarding whether an interaction contained a region which was overlapping any PC accessible peaks (Obtaining PC accessible peaks, section 2.4.11), was annotated. Interactions containing these overlaps were removed and the remaining

interactions are referred to as “MM enhancers regulating OEMM protein coding genes” (Figure 2-5 dark yellow boxes).

2.5.1.4. Obtaining DASMM enhancers - protein coding DESMM genes

The interactions between DASMM enhancers and all protein coding genes were determined as specified in Linking candidate enhancers to target protein coding genes, section 2.5.1.1.

From these interactions, only interactions containing DESMM genes and DASMM enhancers (for at least one subgroup vs. PC), were kept. Information on whether each region – gene pair interaction contains a MM subgroup CL vs. PC DE gene (see Obtaining subgroup MM vs. PC DE genes between MM cell lines vs. PC primary samples, section 2.3.6.4) was added. Furthermore, using Bedtools, information regarding whether an interaction contained a region which was overlapping any PC accessible peaks (see Obtaining PC accessible peaks, section 2.4.11) and whether an overlap with a MM cell line chromatin accessible peak (overlapping with Bedtools the regions obtained in Obtaining cell line chromatin accessible peaks, section 2.4.12), was annotated. These region – gene pairs are referred to as DASMM enhancers - protein coding DESMM genes (Figure 2-5 dark yellow boxes).

2.5.1.5. SMM enhancers regulating protein coding OESMM genes

From the table produced in Obtaining DASMM enhancers - protein coding DESMM genes, section 2.5.1.4, interactions were only maintained if both the region and gene complied with the over accessible and OE criteria respectively explained in Tests performed and thresholds used to obtain significant regions and genes, section 2.2.2, in at least one MM subgroup and not overlapping any PC peaks. These interactions are referred to as SMM enhancers regulating protein coding OESMM genes (Figure 2-5 dark yellow boxes).

2.5.2. Obtaining candidate enhancer regions sets within 1Mb of candidate regulated genes sets

Promoters for protein coding and non-coding genes were obtained by extending the TSS as specified in Linking candidate enhancers to target protein coding genes, section 2.5.1.1. Three promoter sets were created: all promoters, DEMM genes and OEMM genes by filtering out the corresponding gene set in each case.

MMPC enhancers and MM enhancers previously obtained were extended upstream and downstream by 1Mb (Figure 2-5). For the following combinations, the number of extended candidate regions intersecting each promoter were obtained (variations for protein coding and non-protein coding of the below):

- All promoters with extended MMPC enhancers.

- Promoters of DEMM genes with extended MMPC enhancers.
- All promoters with extended MM enhancers.
- Promoters of OEMM genes with extended MM enhancers.

2.5.3. Obtaining states for the MMPC enhancers near DEMM protein coding genes on other cell types

From MMPC enhancers near DEMM protein coding genes (Obtaining MMPC enhancers near DEMM protein coding genes, section 2.5.1.2) the unique regions were acquired. The human genome was divided into 200bp windows and each window overlapping the enhancer regions used was kept using Bedtools intersect.

Next, the Chromatin State Segmentations (12 states) by ChromHMM for the 173 cell types for the GRCh38 genome available to date in The DeepBlue Epigenomic Data Server (Albrecht et al., 2016) were retrieved. The chromatin state segmentations are already divided into the same 200bp windows where each row contains a region and the state for each of the cell type. Each 200bp enhancer region was intersected with the cell states table to obtain the chromatin state for each cell state in the region.

Using manually curated annotations for each cell type provided by our collaborator Nikolaos Trasanidis (Centre for Haematology - Imperial College London), the cell types categories were refined and classified into disease/normal status.

Heatmaps were produced for the regions and states in each cell type. Gower distance between the different states in cell types and enhancers and average clustering linkage was employed for the dendrograms.

2.5.4. Relationship between protein coding promoter accessibility and gene expression in MM and PC

From the quantified genes in the RNA-seq experiments (RNA-seq quantification, section 2.3.2) only protein coding genes that had annotated TSS in the human genome geneset were selected. In total 19,957 genes.

To obtain chromatin accessibility of the promoters in each condition (MM and PC), the promoters for protein coding genes were obtained by extending the TSS as specified in Linking candidate enhancers to target protein coding genes, section 2.5.1.1. For each condition (MM and PC), each promoter was overlapped with the corresponding consensus peaks for PC and MM (see “Balanced consensus chromatin accessible peaks for PC and MM”) using Bedtools.

To obtain the gene expression for each sample in terms of Transcripts Per Million (TPM), the gene quantification pipeline in RNA-seq quantification, section 2.3.2 was used. The average TPM for each gene for MM samples and PC samples was calculated and a threshold of 5 TPM or more than was used to establish an expressed gene.

For each condition, lists of genes were created with all combinations of:

- Promoter chromatin accessibility.
- Gene expression (expressed or not based on threshold).

OEMM genes were obtained as specified in Obtaining OEMM genes, section 2.3.6.5, and only protein coding genes were selected. These gene identifiers were intersected with the gene categories established earlier to get all combinations.

2.5.5. Reads in peaks tables

2.5.5.1. rLog ATAC-seq counts removing batch effects accounting for condition (MM or PC) effect for the MM and PC consensus peaks

The table with the reads in PC and MM consensus peaks were obtained as referred to in Getting the sample five prime shifted tag counts in each consensus peak for PC and MM, section 2.4.9.1. The rLog counts with batch effects removed (taking into account MM and PC effects) were obtained with samples belonging to batches with only one sample were placed on the same meta batch. rLog and the batch effects were removed using the tools specified in General considerations for RNA-seq and ATAC-seq analysis, section 2.2.

2.5.5.2. rLog ATAC-seq counts removing batch effects accounting for subgroup effect for the subgroup MM and PC consensus peaks

The table with the reads in each subgroup consensus peak for PC and MM subgroups was obtained as described in Getting the sample five prime shifted tags in each consensus peak for PC and MM subgroups, section 2.4.9.2. The rLog counts, with batch effects removed (taking into account MM subgroup and PC effects) of the counts per peak were obtained with samples belonging to batches with only one sample were placed on the same meta batch. Non-cytogenetically annotated samples were placed in the same subgroup (MM_OTHER). rLog and the batch effects were removed using the tools specified in General considerations for RNA-seq and ATAC-seq analysis, section 2.2.

2.5.6. Gene quantification tables

2.5.6.1. rLog RNA-seq counts removing batch effects accounting for condition (MM or PC) effect

The table of the quantified reads in each gene for each sample (Table 2-1 "MM-PC" category) was produced as explained in RNA-seq quantification, section 2.3.2. rLog and batch correction, taking into account MM and PC effects, was performed as previously specified, with samples belonging to batches with only one sample placed on the same meta batch.

2.5.6.2. rLog RNA-seq counts removing batch effects accounting for subgroup effect

The table of the quantified reads in each gene for each sample (Table 2-1 "Subgroup" category) was produced as explained in RNA-seq quantification, section 2.3.2. rLog and batch correction, taking into account MM subgroup and PC effects was performed as previously specified with samples belonging to batches with only one sample were placed on the same batch.

2.5.7. MM and PC consensus peaks chromatin accessibility and RNA-seq profiles for all genes

For the chromatin accessibility profiles, the normalized reads in MM and PC consensus peaks were obtained as described in rLog ATAC-seq counts removing batch effects accounting for condition (MM or PC) effect for the MM and PC consensus peaks, section 2.5.5.1. The average rlog count for the PC samples was calculated. The $\log_2$ fold changes between MM and PC were obtained for each consensus peak for PC and MM by the same procedure as the DAMM regions (see Obtaining DAMM regions, section 2.4.13.1) without filtering significant DA regions.

For the RNA-seq profiles, the normalized gene counts were obtained as mentioned in rLog RNA-seq counts removing batch effects accounting for condition (MM or PC) effect, section 2.5.6.1. The average rlog for the PC samples was calculated. The $\log_2$ fold changes between MM and PC were obtained for each gene for PC and MM by the same procedure as the DEMM genes (see Obtaining DEMM genes, section 2.3.6.1) without filtering out significant DE genes.

For both ATAC and RNA profiles, since raw counts were normalized and batch effects removed accounting for condition, only the MM vs PC Log Ratio Test (LRT) accounting for batch and its adjusted p-value was used. The average rlog count for the PC and $\log_2$ fold changes between MM and PC were plotted taking into account that adjusted p-values of less than 0.05 were considered significant. The regions and genes either with 0 basemeans, containing a MM or PC sample with an extreme count outlier or filtered by Deseq2 automatic independent filtering for having a low mean normalized count were not shown.

2.5.8. Subgroup MM and PC consensus peaks chromatin accessibility and RNA-seq profiles for all genes

An analogous version to MM and PC consensus peaks chromatin accessibility and RNA-seq profiles for all genes, section 2.5.7, for MM subgroups vs. PC, was computed.

In this case, for the chromatin accessibility profiles, the normalized read counts in subgroup MM and PC consensus peaks (see rLog ATAC-seq counts removing batch effects accounting for subgroup effect for the subgroup MM and PC consensus peaks, section 2.5.5.2) were obtained and mean rLog for the PC samples in each consensus peak calculated. The $\log_2$ fold changes between each subgroup and PC, and cognate adjusted p-values obtained from the same analysis described in Obtaining DASMM regions, section 2.4.13.2, without filtering significant DA regions.

For the RNA-seq profiles, the normalized gene counts were obtained as described in rLog RNA-seq counts removing batch effects accounting for subgroup effect, section 2.5.6.2. The mean rLog counts for the PC samples was calculated. The $\log_2$ fold changes between each subgroup and PC and adjusted p-values were obtained for all genes by the same procedure as with the DESMM genes (see Obtaining DESMM genes, section 2.3.6.2) without filtering the significant DE genes.

Threshold for significance is the same as in MM and PC consensus peaks chromatin accessibility and RNA-seq profiles for all genes, section 2.5.7.

2.5.9. Chromatin accessibility and gene expression subtyping classification profiles for MMPC enhancers near DEMM coding genes

For the chromatin accessibility profiling, the normalized read counts in the MM and PC consensus peaks were obtained as described in rLog ATAC-seq counts removing batch effects accounting for condition (MM or PC) effect for the MM and PC consensus peaks, section 2.5.5.1. From this table, only the regions corresponding to the MMPC enhancers near differential protein coding genes (see Obtaining MMPC enhancers near DEMM protein coding genes, section 2.5.1.2) were considered.

For the RNA-seq profiles, the normalized gene counts were obtained as described in rLog RNA-seq counts removing batch effects accounting for condition (MM or PC) effect, section 2.5.6.1. From this table, only genes corresponding to the MMPC enhancers near DEMM protein coding genes were considered.

Different heatmaps were produced for the selected regions and genes using the normalized counts and the heatmap.2 function from the gplots R package. For the hierarchical clustering of samples and features 1 – Pearson correlation was used as distance metric and average linkage as the clustering method.

2.5.10. Chromatin accessibility and gene expression subtyping classification profiles for DASMM enhancers regulating protein coding DESMM genes

For the chromatin accessibility profiling, the normalized read counts in the subgroup MM and PC consensus peaks were obtained as described in rLog ATAC-seq counts removing batch effects accounting for subgroup effect for the subgroup MM and PC consensus peaks, section 2.5.5.2. From this table, only the regions corresponding to the DASMM enhancers - protein coding DESMM genes (see Obtaining DASMM enhancers - protein coding DESMM genes, section 2.5.1.4) were considered.

For the RNA-seq profiles, the normalized gene counts were obtained as mentioned in rLog RNA-seq counts removing batch effects accounting for subgroup effect, section 2.5.6.2. From this table, only DASMM enhancers - protein coding DESMM genes.

Different heatmaps (R heatmap.2 package) were produced for the selected regions and genes using the normalized counts with 1 – Pearson correlation was used as distance between features and samples, average linkage used for regions and complete linkage used for the clustering of genes.

2.5.11. Accessibility and CCND2 expression correlation plots for candidate enhancer regions regions

The normalized read counts in the subgroup MM and PC consensus peaks were obtained as specified in rLog ATAC-seq counts removing batch effects accounting for subgroup effect for the subgroup MM and PC consensus peaks, section 2.5.5.2, from this the values for the candidate enhancer regions for the CCND2 gene studied were selected. The normalized gene counts (see rLog RNA-seq counts removing batch effects accounting for subgroup effect, section 2.5.6.2) were also obtained for the CCND2 gene.

The correlation (R^2 and Pearson) between both was calculated.

2.6. Motif enrichment

Motif enrichment was performed on selected regions by inputting them to the pipeline_denovo_motifs pipeline using CGATPipelines framework (Sudbery, 2019b) which was

developed by Ian Sudbery and contains some fixes done by myself. Briefly, the analysis pipeline runs MEME (Bailey and Elkan, 1994) and DREME (Bailey, 2011) version 4.12.0 from the MEME suite. MEME was used to detect 40 maximum motifs, using DNA alphabet, allowing sites on both strands, finding distribution of motifs with any number of repetitions. DREME was used with minimum width of core motif 5 and maximum 30. This was performed to find novel motifs enriched in the provided regions. The motifs found were inspected for similarities between them using Tomtom (Gupta et al., 2007), which is also included in the MEME suite. First clusters of motifs were created by linking *de novo* motifs that were significantly similar (q-value less than 0.05). The *de novo* motif having the most significant E-value and number of found binding sites was selected as the representative motif of the cluster. Then clusters were merged if their representative *de novo* motifs were similar (q-value less than 0.1).

The reference motifs for each cluster were then checked for similarity to consensus binding sites for TFs from multiple databases using Tomtom. The databases used are:

- Jaspar databases: JASPAR_CORE_2016, JASPAR_CORE_REDUNDANT_2016, JASPAR_CORE_2016_vertebrates, JASPAR_CORE_REDUNDANT_2016_vertebrates (Mathelier et al., 2016).
- HOCOMOCO databases: HOCOMOCOv10_HUMAN, HOCOMOCOv10_MOUSE (Kulakovskiy et al., 2016).
- CIS-BP databases: Homo_sapiens, Mus_musculus (Weirauch et al., 2014).
- EUKARYOTE wei2010 human and mouse databases included with MEME version 4.12.0.

2.6.1. TF binding enrichment in the unique MM enhancers near OEMM protein coding genes

Motif enrichment was performed on MM enhancers near OEMM protein coding genes, by getting the unique regions from all the interactions and merging contiguous regions using Bedtools. These regions were then inputted and processed as specified in Motif enrichment, section 2.6.

2.6.2. TF binding enrichment in the SMM enhancers regulating protein coding OESMM genes

Motif enrichment was performed on the SMM enhancers regulating protein coding OESMM genes, by getting only the unique regions from all the interactions and merging contiguous regions using Bedtools. These regions were then input and processed as specified in the start of the Motif enrichment, section 2.6.

A threshold of E-value < 0.05 was used for *de novo* motifs which were enriched in the regions for each enhancer set and subsequent similarity of TFs motifs with these *de novo* motifs. Only unique motifs - corresponding TFs combinations were reported. In cases where a motif for a TF was found for multiple species, only the human version was shown, if the human version was not present, only the mouse version was shown if it was available. The results using DREME and MEME were combined for each enhancer set. In turn, the results for each enhancer set were combined into a final table, showing whether each enriched motif in at least one MM subgroup enhancer was enriched in the other MM subgroup enhancers.

2.6.3. Gene expression comparison for TF genes binding to SMM enhancers regulating protein coding OESMM genes for the different conditions

Using the TF binding enrichments in the regulatory SMM enhancers (see TF binding enrichment in the SMM enhancers regulating protein coding OESMM genes, section 2.6.2). A list of 425 TF genes was obtained by including genes having motif enrichment in at least one subgroup for regulatory SMM enhancers (38 TFs) and then overlapping and getting unique gene ids (Ensembl id) from:

- TF genes found in the HOCOMOCOv10 HUMAN database (641 TFs).
- Genes with measured RNA-seq data in this study (57992 genes).

The 425 TF genes are referred to as annotated TF genes. To obtain Ensembl ids from gene names, the human gene symbol to Ensembl id conversions with the R package *org.Hs.eg.db* version 3.7.0 (Carlson, 2018) and a curated list for human genes (National Center for Biotechnology Information, 2019) were used.

For this list, the rLog with batch effects removed of the gene expression were obtained as previously specified on the sections 2.2.3 and 2.2.4 respectively for quantified genes (section 2.3.2) for all quality samples (Table 2-1, including the PC, MM CL and all MM samples). Since there are comparisons between PC, MM and MM CLs and given that MM CLs subgroups may not be comparable to MM subgroups, it was decided to not account for subgroup effect in this calculation.

Gene expression was obtained for each TF list and all combinations of:

- Each group of samples PC, MM, MM CLs.
- Two groups from the annotated genes: one with the TF genes with significant binding in MM any subgroup and another one with no enrichment in MM subgroup.

For each combination, the mean regularized gene expression was calculated for the corresponding samples and genes. Then a two-sample one tail Kolmogorov-Smirnov test within each group of samples (MM, PC and MM CLs) was performed under the null hypothesis that the distribution of gene means was not greater for TFs with significant motif enrichment in any MM subgroup compared with TFs without significant motif enrichment in any MM subgroup.

100,000 permutation tests were performed through random sampling of a number of annotated genes equal to the number of TFs with a motif significantly enriched in any MM subgroup from all annotated genes. For MM, PC and MM CLs, for each sample the average expression of the selected TFs was calculated. The mean of all gene expression means (average expression of the selected TFs for the random sample of TFs) was compared to that of MM subgroup enriched annotated genes in any MM subgroup. The number of tests having the random sample average TF gene average expression higher or equal to that of TF motif enriched annotated genes in any MM subgroup was obtained and p-values were calculated.

Similarly, as done with each combination of PC, MM and MM CLs samples, another set of combinations was performed for:

- Each MM subgroup.
- Two groups from the annotated genes: one with the TF genes with significant binding in the particular MM and another one with no enrichment in the particular MM subgroup.

For each combination, the mean regularized gene expression was calculated for the corresponding samples and genes. Then a two-sample one tail Kolmogorov-Smirnov test within each group of samples (MM subgroups) was performed under the null hypothesis that the distribution of gene means was not greater for TFs with significant motif enrichment in the particular MM subgroup compared with TFs without significant motif enrichment in the particular MM subgroup.

For each subgroup, 100,000 permutation tests were performed through random sampling of all annotated genes, each sample's number of elements was equal to the number of TF motif enriched annotated genes in that particular MM subgroup. The average of all gene expression averages (average expression of the selected TFs for the random sample) was compared to that of MM subgroup enriched annotated genes in that MM subgroup. The number of tests having the random sample average TF gene average expression higher or equal to that of TF

motif enriched annotated TF genes in that MM subgroup was obtained and p-values were calculated.

2.7. Genomic annotations of regions

To annotate regions, the R library Annotatr version 1.8.0 (Cavalcante and Sartor, 2017) was used, with annotations from the library TxDb.Hsapiens.UCSC.hg38.knownGene (Team_BC and Maintainer_BP, 2019). Any region can overlap multiple different types of genomic annotations on both strands but for each region, a particular type is only reported once. The annotations cover the following types:

- Gene promoters (<1Kb upstream of the TSS).
- Genes 5' UTRs
- Genes 3' UTRs
- Genes coding sequences (cds): All exons after removing the 5' UTRs and 3' UTRs
- Genes introns
- Genes intergenic regions

2.8. Multi Omics Factor Analysis (MOFA)

2.8.1. rLog ATAC-seq counts removing batch effects accounting for subgroup effect for the MM and PC consensus peaks

Reads in each consensus PC and MM peak were obtained as referred to in Getting the sample five prime shifted tag counts in each consensus peak for PC and MM, section 2.4.9.1. The rLog with batch effects removed (taking into account MM subgroup and PC effects) of the counts per peak were obtained, samples belonging to batches with only one sample were placed on the same batch, non-cytogenetically annotated samples were placed on the same subgroup (MM_OTHER). rLog and batch effects were removed from the rLog expression as previously specified.

2.8.2. MOFA input features and execution

For ATAC-seq, normalized reads in each consensus PC and MM peak were obtained as specified in rLog ATAC-seq counts removing batch effects accounting for subgroup effect for the MM and PC consensus peaks, section 2.8.1. Regions corresponding to gender chromosomes (chrX and chrY) and annotated and unannotated TSS (obtained as explained in

in Obtaining annotated and unannotated Transcription Start Sites, section 2.3.5) were removed from these regions using GNU Awk and Bedtools respectively, yielding 273,216 remaining regions. The variance per peak was calculated and only the 5000 (or 10,000) peaks with the highest variance were selected.

For RNA-seq, normalized gene counts were obtained as specified in rLog RNA-seq counts removing batch effects accounting for subgroup effect, section 2.5.6.2. Genes with the TSS in gender chromosomes in the human genome were removed from the table. The sample names from the RNA-seq were converted to ATAC-seq sample names in this table, the variance per gene was calculated and only the 5000 (or 10,000) genes with highest variance.

Using the R library MOFAtools version 0.99.0 (Argelaguet et al., 2018), the ATAC-seq and RNA-seq tables were inputted to MOFA. The default data and model options were used and Gaussian data was selected both for ATAC-seq and RNA-seq. The training options used that were different to the defaults were: dropping factor threshold of 0.01, 10,000 maximum iterations, minimum tolerance convergence threshold of 0.01. Information on each samples' subgroup (cytogenetic MM subgroup, PC or "MM_OTHER" for MM samples with no cytogenetic information), condition (PC or MM) was not used by MOFA but was included for the analysis of the results.

The R script used to train the MOFA model can be found in:

MOFA/MOFA_top5k_var_peaks_and_genes_no_gender_no_TSS_train_model.R

The bash script to run the MOFA R script can be found in:

MOFA/MOFA_train_model_script.sh

2.8.3. Silhouette score for samples

The sample and LF weights table (for all LFs) resulting from running MOFA as specified in MOFA input features and execution, section 2.8.2, were obtained. Additionally, the same table from running a MOFA model using only the same RNA-seq features and with the same parameters was obtained. For each model, each sample was assigned a label belonging to the cluster determined by its cytogenetic subgroup, the per sample silhouette score was obtained by calculating Euclidean distances between samples using all LFs as dimensions and also LF1 to LF5. This was done using the "silhouette" function from the "cluster" version 2.0.6 R package (Maechler et al., 2016). The mean silhouette score for each subgroup and model was calculated.

2.8.4. MOFA gene features and MM disease – gene association scores

The genes - LF weights table (for all LFs) resulting from running MOFA as specified in MOFA input features and execution, section 2.8.2, was obtained. The 1,311 available genes and their association score with Multiple Myeloma (Piñero et al., 2020) by selecting "Multiple Myeloma, C0026764" were obtained and added to the genes - LF weights table.

2.8.5. MOFA features interactions between candidate enhancers and genes with supervised analysis details

The LF factor weights for each MOFA accessibility and gene feature and each LF was obtained, the corresponding annotated promoter sites for these gene features were obtained as described in Obtaining annotated and unannotated Transcription Start Sites, section 2.3.5. As done previously to obtain interactions, the promoters were extended 1Mb upstream and downstream and overlapped with the MOFA accessibility features (Figure 2-5 and Linking candidate enhancers to target protein coding genes, section 2.5.1.1), this was done for each LF.

For the resulting candidate enhancer – gene interactions it was determined if they were overlapping MM enhancers near OEMM protein coding genes and MMPC enhancers near DEMM protein coding genes. Additionally, details for the accessibility and expression mean counts and for the all log_2foldchanges and its averages (see Tests performed and thresholds used to obtain significant regions and genes, section 2.2.2 for the MM vs. PC analysis) were added to these interactions.

Furthermore, the interactions were also annotated with MM subgroup vs. PC accessibility and expression details from the supervised analysis. For accessibility features, each MOFA feature was overlapped with the consensus peaks for primary PC and MM subgroups, with Bedtools version 2.22.1. In cases where there were multiple overlaps, the details for the PC and MM subgroup peak with the largest overlap was selected. The subgroup details added include accessibility and expression details (see Tests performed and thresholds used to obtain significant regions and genes, section 2.2.2, for the subgroup MM vs. PC analysis) for log_2FoldChanges, whether each feature is DE/DA, OE or over accessible in each subgroup.

The fields in the MOFA/MOFA_all_LFs_ATAC_1Mb_RNA_promoters_MM_vs_PC_and_subgroup_MM_vs_PC_details.tsv.gz table are:

- Chr: Chromosome of the MOFA region.
- Start: Coordinate start of the MOFA region.

- End: Coordinate end of the MOFA region.
- Factor: MOFA model factor.
- Weight_ATAC: Weight assigned to the ATAC-seq feature.
- Weight_RNA: Weight assigned to the RNA-seq feature.
- SYMBOL: Gene symbol of the RNA-seq feature.
- GENENAME: Gene description of the RNA-seq feature.
- Present_MM_vs_PC_strict: Whether the interaction is a MM enhancer near OEMM protein coding gene.
- ATAC_baseMean: Accessibility counts mean for the MOFA region.
- ATAC_log2FoldChange_MM_vs_average_ND_CD19: Accessibility MM vs. average CD19 (PC) log_2FoldChanges.
- ATAC_padj_MM_vs_average_ND_CD19: Accessibility MM vs. average CD19 (PC) adjusted p-value.
- ATAC_log2FoldChange_MM_vs_average_ND_donor: Accessibility MM vs. average donor (PC) log_2FoldChanges.
- ATAC_padj_MM_vs_average_ND_donor: Accessibility MM vs. average donor (PC) adjusted p-value.
- ATAC_log2FoldChange_LRT_batch_condition_MM_vs_PC: Accessibility MM vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChanges.
- ATAC_padj_LRT_batch_condition_MM_vs_PC: Accessibility MM vs. PC Log Ratio Test (LRT) accounting for batch adjusted p-value.
- ATAC_MM_vs_PC_log2FoldChange_avg: Average accessibility MM vs. PC log_2FoldChanges (ATAC_log2FoldChange_MM_vs_average_ND_CD19, ATAC_log2FoldChange_MM_vs_average_ND_donor, ATAC_log2FoldChange_LRT_batch_condition_MM_vs_PC).
- RNA_baseMean: Expression counts mean for the MOFA gene.
- RNA_log2FoldChange_MM_vs_average_ND_CD19: Expression MM vs. average CD19 (PC) log_2FoldChanges.
- RNA_padj_MM_vs_average_ND_CD19: Expression MM vs. average CD19 (PC) adjusted p-value.
- RNA_log2FoldChange_MM_vs_average_ND_donor: Expression MM vs. average donor (PC) log_2FoldChanges.
- RNA_padj_MM_vs_average_ND_donor: Expression MM vs. average donor (PC) adjusted p-value.

- RNA_log2FoldChange_LRT_batch_condition_MM_vs_PC: Expression MM vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChanges.
- RNA_padj_LRT_batch_condition_MM_vs_PC: Expression MM vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChanges.
- RNA_MM_vs_PC_log2FoldChange_avg: Average expression MM vs. PC log_2FoldChanges (RNA_log2FoldChange_MM_vs_average_ND_CD19, RNA_log2FoldChange_MM_vs_average_ND_donor, RNA_log2FoldChange_LRT_batch_condition_MM_vs_PC)
- Present_MM_vs_PC_generic: Whether the interaction is a MM, PC candidate enhancers near a differential protein coding gene.
- Gene: The Ensembl id of the MOFA gene.
- ATAC_padj_LRT_MM_subgroup_vs_PC: Accessibility MM subgroup vs. PC Log Ratio Test (LRT) accounting for batch adjusted p-value.
- ATAC_log2FoldChange_HD_vs_ND: Accessibility HD vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- ATAC_lfcSE_HD_vs_ND: Accessibility HD vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- ATAC_log2FoldChange_CCND1_vs_ND: Accessibility CCND1 vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- ATAC_lfcSE_CCND1_vs_ND: Accessibility CCND1 vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- ATAC_log2FoldChange_MAF_vs_ND: Accessibility MAF vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- ATAC_lfcSE_MAF_vs_ND: Accessibility MAF vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- ATAC_log2FoldChange_MMSET_vs_ND: Accessibility MMSET vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- ATAC_lfcSE_MMSET_vs_ND: Accessibility MMSET vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- ATAC_DE_HD_vs_ND: Accessibility consensus subgroup MM and PC region is DA in HD vs. PC, this is, accessibility MM subgroup vs. PC LRT accounting for batch adjusted p-value < 0.05 and:
    - -1 if log_2FoldChange <= -1
    - +1 if log_2FoldChange >= +1

    - 0 if abs(log_2FoldChange) < 1
- ATAC_DE_CCND1_vs_ND: Accessibility consensus subgroup MM and PC region is DA in CCND vs. PC.
- ATAC_DE_MAF_vs_ND: Accessibility consensus subgroup MM and PC region is DA in MAF vs. PC.
- ATAC_DE_MMSET_vs_ND: Accessibility consensus subgroup MM and PC region is DA in MMSET vs. PC.
- RNA_padj_LRT_MM_subgroup_vs_PC: Expression MM subgroup vs. PC Log Ratio Test (LRT) accounting for batch adjusted p-value.
- RNA_log2FoldChange_HD_vs_ND: Expression HD vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- RNA_lfcSE_HD_vs_ND: Expression HD vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- RNA_log2FoldChange_CCND1_vs_ND: Expression CCND1 vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- RNA_lfcSE_CCND1_vs_ND: Expression CCND1 vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- RNA_log2FoldChange_MAF_vs_ND: Expression MAF vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- RNA_lfcSE_MAF_vs_ND: Expression MAF vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- RNA_log2FoldChange_MMSET_vs_ND: Expression MMSET vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange.
- RNA_lfcSE_MMSET_vs_ND: Expression MMSET vs. PC Log Ratio Test (LRT) accounting for batch log_2FoldChange standard error.
- RNA_DE_HD_vs_ND: HD vs. PC differential expression, this is, expression of MM subgroup vs. PC LRT accounting for batch adjusted p-value < 0.05 and:
    - -1 if log_2FoldChange <= -1.5
    - +1 if log_2FoldChange >= +1.5
    - 0 if abs(log_2FoldChange) < 1.5
- RNA_DE_CCND1_vs_ND: CCND1 vs. PC differential expression.
- RNA_DE_MAF_vs_ND: MAF vs. PC differential expression.
- RNA_DE_MMSET_vs_ND: MMSET vs. PC differential expression.

- Present_CCND1_vs_PC_strict: Whether the MOFA interaction is more accessible and OE in CCND1 compared with PC (supervised analysis).
- Present_HD_vs_PC_strict: Whether the MOFA interaction is more accessible and OE in HD compared with PC (supervised analysis).
- Present_MAF_vs_PC_strict: Whether the MOFA interaction is more accessible and OE in MAF compared with PC (supervised analysis).
- Present_MMSET_vs_PC_strict: Whether the MOFA interaction is more accessible and OE in MMSET compared with PC (supervised analysis).
- Present_CCND1_vs_PC_generic: Whether the MOFA interaction is DA and DE in CCND1 compared with PC (supervised analysis).
- Present_HD_vs_PC_generic: Whether the MOFA interaction is DA and DE in HD compared with PC (supervised analysis).
- Present_MAF_vs_PC_generic: Whether the MOFA interaction is DA and DE in MAF compared with PC (supervised analysis).
- Present_MMSET_vs_PC_generic: Whether the MOFA interaction is DA and DE in MMSET compared with PC (supervised analysis).

2.9. Gene Ontology analysis

Gene ontology analysis was performed using R version 3.5.1 and the packages goseq version 1.34.0, TxDb.Hsapiens.UCSC.hg38.knownGene version 3.4.0, geneLenDataBase version 1.18.0, org.Hs.eg.db version 3.7.0, KEGGREST 1.22.0. The analysis Rscript can be found in:

Analysis_scripts/Goseq-analysis_human_parameters.R

Briefly, the script is provided with background genes which include the particular selected genes for enrichment analysis. It calculates a Probability Weighting Function for all background genes supplied taking into account gene length using the Goseq (Young et al., 2010) 'nullp' function and gets the Gene Ontology categories for all background genes. Results are obtained by testing different enriched statistics. FDRs are calculated using the Benjamini-Hochberg procedure and thresholded using the FDR specified. An enrichment metric for each category is calculated as: the ratio of proportion of selected genes in the category from all selected genes with category and proportion of background genes in the category from all background genes with category. The different statistics used in the testing of all categories (Cellular Component, Biological Process and Molecular Function): Wallenius, Random sampling to generate the null

distribution, Hypergeometric with no gene length correction. Unless stated otherwise, only Wallenius is reported. Additionally, Wallenius approximation is done only on Biological Process and Molecular Function independently to reduce the multiple testing burden. Finally, a KEGG (Kyoto Encyclopedia of Genes and Genomes) analysis pathway is performed. To reduce the number of tests to perform for each approximation, only categories with at least 10 background genes are taken into account.

2.9.1. Gene Ontology analysis on DEMM genes

Gene Ontology analysis was performed as specified in Gene Ontology analysis, section 2.9, using as background genes all genes quantified and selected genes, the DEMM genes, using FDR of 0.1.

2.9.2. Gene Ontology analysis on OEMM genes

Gene Ontology analysis was performed as specified in Gene Ontology analysis, section 2.9, using as background genes all genes quantified and selected genes, the OEMM genes, using FDR of 0.1.

2.9.3. Gene Ontology analysis on DESMM genes

First subsets of the DESMM genes were calculated for each subgroup (see thresholds for DE genes in Tests performed and thresholds used to obtain significant regions and genes, section 2.2.2). Each of these sets is referred from here onwards as differential expressed (DE) genes between MM particular subgroup and PC.

Different Gene Ontology analysis were performed as specified in Gene Ontology analysis, section 2.9, using as background all genes quantified and as selected genes, the DESMM genes and each of the sets of DE genes between MM particular subgroup and PC using FDR of 0.1. Wallenius significant test results for each of the ontology analysis were combined filtering using a category enrichment of genes of 2 or more over background was performed for each of the test sets.

2.9.4. Gene Ontology analysis on OESMM genes

Subsets of the OESMM were calculated for each subgroup (see thresholds for OE genes in Tests performed and thresholds used to obtain significant regions and genes, section 2.2.2). Each of these sets is referred from here onwards as particular subgroup OESMM genes (for example MAF OESMM genes).

Different Gene Ontology analysis were performed as specified in Gene Ontology analysis, section 2.9, using as background all genes quantified and as selected genes, the OESMM genes

and each of the sets of particular subgroup OESMM genes using FDR of 0.1. Wallenius significant test results for each of the ontology analysis were combined and filtered using a category enrichment cutoff of 2 or more over background.

2.9.5. Gene Ontology analysis on top MOFA LF1 and LF2 MM activated genes

Separate gene Ontology analyses were performed for LF1 and LF2, as specified in Gene Ontology analysis, section 2.9, using as background all genes used as MOFA features and as selected genes, the top 10% (500) genes in each LF with the most extreme negative weights (reflecting MM activation) using FDR of 0.1. Wallenius significant test results were performed using a category enrichment cutoff of 2 or more over background.

2.9.6. Gene Ontology analysis on top MOFA LF3 genes by absolute loading

Gene Ontology analysis was performed for LF3, as specified in Gene Ontology analysis, section 2.9, using as background all genes used as MOFA features and as selected genes, the top 10% (500) genes by LF3 absolute weights only for only Molecular Function categories at FDR of 0.1. Wallenius significant test results were outputted, filtering using a category enrichment cutoff of 2 or more over background.

2.9.7. Gene Ontology analysis on top MOFA LF5 genes by absolute loading

Gene Ontology analysis was performed for LF5, as specified in Gene Ontology analysis, section 2.9, using as background all genes used as MOFA features and as selected genes, the top 10% (500) genes by LF5 absolute weights only for only Molecular Function categories at FDR of 0.1. Wallenius significant test results were outputted.

3. Chapter 3: MM vs. PC chromatin and gene expression analysis

3.1. Acronyms and Abbreviations used in the chapter

Acronym	Definition
3C	Chromosome Conformation Capture
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BM	Bone Marrow
BMP	Bone Morphogenetic Protein
bp	Base pair
BP	Biological Process
CL	Cell Line
DA	Differentially Accessible
DAMM	Differentially Accessible MM
DE	Differentially expressed
DEMM	Differentially Expressed MM
eRNA	Enhancer RNA
FACS	Fluorescence Activated Cell Sorting
FDR	False Discovery Rate
H3K27ac	Acetylation of histone H3 lysine 27
H3K4me1	Histone H3 lysine 4 monomethylation
H3K4me3	Histone H3 lysine 4 trimethylation
HD	Hyperdiploid
IgH	Immunoglobulin Heavy Chain
Kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
lnc-RNA	Long non-coding RNA
log_2foldchange	Log (base 2) fold change
Mb	Megabase
mESC	Mouse Embryonic Stem Cells
MF	Molecular Function
miRNA	Micro RNA
MM	Multiple Myeloma
MMPC	Multiple Myeloma and Plasma Cell
mRNA	Messenger RNA
ncRNA	Non-coding RNA
ND	Normal Donor
OE	Overexpressed
OEMM	Over Expressed MM
PC	Plasma Cell (used interchangeably with ND)
PCR	Polymerase Chain Reaction
RNA-seq	RNA sequencing
SNP	Single Nucleotide Polymorphism
TF	Transcription Factor
TPM	Transcripts per Million

TSS	Transcription Start Site
UTR	Untranslated Region

3.2. Introduction

MM is a cancer where the primary driver events in half of the patients involve rearrangement of the IgH enhancers deregulating the expression of oncogenes. MM enhancer biology is therefore a critical subject to study to understand how the PCs transition to the malignant state. Efforts in this regard have been previously performed extensively in hematopoietic cancers, for example, in different leukemia types (Gröschel et al., 2014; Yamazaki et al., 2014) where through chromosomal inversions and translocations, the expression of EVI1 is altered, or in Blastic plasmacytoid dendritic cell neoplasms where the RUNX2 super-enhancer activates MYC (Kubota et al., 2019). In particular, B-cell lymphomas (Taub et al., 1982) and MM (Affer et al., 2014) have also benefited from these studies.

3.2.1. ATAC-seq quality control metrics

More recently, MM cells from patient bone marrow (BM) were compared with memory B-cells which were in vitro differentiated into PC (Jin et al., 2018). MM putative enhancer regions were determined on the basis of H3K27ac signal and approximately 20,000 candidate enhancers were found having a different signal between the differentiated PCs and MM samples. Candidate enhancer regions were assigned to genes within 200Kb accounting for CTCF sites. Since this study uses in vitro differentiated cells, only 11 MM samples (none reported to include MAF translocations) and H3K27ac signals which are known to generate different enhancer sets than chromatin accessibility (Kleftogiannis et al., 2015), it is possible that the study shown here can extend the MM enhancer biology knowledge. Samples quality is a key added value in the study in this thesis, recommendations in terms of ATAC-seq quality control include assigned fraction (reads in called chromatin accessible peaks from the total reads), fragment length, input reads and called peaks (Alasoo et al., 2017) have been proposed and taken into account.

Read number: For ATAC-seq experiments, 50 million mapped reads have been proposed as a minimum to infer chromatin accessibility changes (Buenrostro et al., 2015; Neph et al., 2012). Other studies have produced at least 30 million filtered reads entering the peak caller (Ackermann et al., 2016).

Assigned fraction: The assigned fraction is a measure of the ratio of reads used to form chromatin accessible regions with respect to the background reads produced by the assay

(similar to a signal to noise ratio). An assigned fraction of 10% or more has been used as a cutoff (Alasoo et al., 2017).

Number of peaks called: Since the number of called peaks to be expected is correlated with the assigned fraction and the number of input reads, once these two metrics are accomplished, around 30,000 peaks per sample seems to be a reasonable lower limit (Alasoo et al., 2017).

The ENCODE consortium has also produced ATAC-seq analysis guidelines and an analysis pipeline based on the work of the Kundaje lab (Kundaje, 2019a, 2019b). Together with Dr. Ian Sudbery, we also implemented many of the features of this pipeline when analyzing this data. Of particular interest are metrics which inform about the sample library complexity (where the sample lies in the unique fragment saturation curve).

Different peak calling tools are available for ATAC-seq, they are mainly imported from the ChIP-seq data analysis. The most widely used are MACS2 (Zhang et al., 2008), Epic2 (Stovner and Sætrom, 2019) or ZINBA (Rashid et al., 2011). Following ENCODE guidelines, MACS2 was used to analyze the data.

3.2.2. Chapter Aims

The ultimate aim of this chapter is to obtain a set of active MM enhancers (Figure 2-3) and relate them to OEMM genes. For this, quality samples according to the guidelines stated are used to establish the chromatin accessible landscape of MM and the healthy state and obtaining significant differences between them pointing at condition specific enhancers. In a similar way, the gene expression program for the healthy and malignant state and associated expression changes during Myelomagenesis are also elucidated. Altered genes are studied in terms of the pathways involved through gene ontology. In addition, the influence of promoter accessibility in cancer altered genes is also examined to determine how critical of a factor it is.

Specific and shared interactions between PC and MM are inferred from candidate enhancer regions being associated to genes though condition specific changes. This catalogue of regulatory mechanisms is also cross checked with MM cell lines to produce testable candidates. Furthermore, the regulatory regions inferred from these interactions are also studied on other cell types to determine if these candidate enhancer regions are active or are novel on other lineages. Moreover, from these interactions, the exclusively active ones in MM are studied for TF networks involved in them.

3.3. Results

3.3.1. Quality control statistics

For this study, 60 MM patients were recruited into the study following at diagnosis and relapse with all clinical information at recruitment. Bone marrow (BM) samples were obtained either at the Hammersmith Hospital, Imperial College Healthcare NHS Trust in London, or at the AHEPA University Hospital of Thessaloniki in Greece, patients were consented by Alexia Katsarou, (Department of Haematology) and Evdoxia Hatjiharissi at each facility respectively. Written informed consent and research ethics committee approval was obtained (Research Ethics Committee reference: 11/H0308/9).

ATAC-seq was performed on 60 samples and RNA-seq on 54 as specified in section 2.1. Sample input material for ATAC-seq and RNA-seq was decided after purification aiming for 50,000 and 100,000 paired-end reads for ATAC and RNA respectively (when available). The samples were then bioinformatically analysed post-hoc to maintain only samples complying with the guidelines mentioned in the introduction section of this chapter.

Due to too low assigned fraction, a low number of single ends entering the peak caller, low quality when manually inspecting a sample's peaks produced, ATAC – RNA pairing not available, non-comparable RNA type, RNA-seq contamination, only 38 paired ATAC – RNA were used (see Table 2-1), of which 33 are primary samples and 5 MM CL samples. Cytogenetic information was obtained for some samples prior to sequencing and was later confirmed by in-house methods by Philippa May. Cytogenetics is available for primary MM samples creating subgroups based on translocations involving the IgH locus and the MMSET/NSD2, MAF and CCND1 genes and the Hyperdiploidy (HD) status (chapter 4).

Since the ATAC-seq sequencing material was of significantly lower depth for MM CLs compared to PC and MM primary samples, the former samples were used only in a qualitative (and not quantitative) manner and correspondingly different quality control criteria were employed. Primary samples were used in determining the candidate enhancer regions regulating genes, only samples with more than around 30M single ends entering the peak caller (except sample A26.11 with 28,278,284) and in general greater than 10% assigned fraction (except sample A26.6B with 7%) were considered. Cell lines were only required to have an assigned fraction of around 10% or more and were used to obtain primary MM interactions also activated in CLs.

The 38 samples consisted of: 5 primary PC samples comprising of 3 donors (with $CD19^+$ and $CD19^-$ variant samples for two of them and a technical replicate for one of these) and 28 MM primary samples. In total, 3,124,720,754 RNA-seq read pairs were generated with an average mapping rate of 83% when quantifying reads in transcripts, generating an average of 67,937,531 mapped reads per sample. 3,287,274,493 ATAC-seq read pairs were sequenced and 2,258,363,658 total unique read single ends (average of 59,430,623 per sample) were input into the peak caller to generate a total of 2,350,508 and 2,645,283 sample narrow and broad peaks respectively. The average sample assigned fraction is 21%. The table with all the samples and details can be seen at: MM_vs_PC_supervised_analysis/ATAC_and_RNA-seq_stats.xlsx

To determine that the ATAC-seq reads were piling up in patterns reflecting open chromatin consistent with previous studies (Alasoo et al., 2017), the number of single ends entering peak calling, the peaks produced and the assigned fraction were studied and classified in groups of samples (PC, MM and MM CL). The number of peaks per sample after filtering for areas of high and low mappability is greater for PC than for MM or MM CL with medians of 80,000, 60,000 and 40,000 respectively (Figure 3-1 A). The number of single ends (one read pair has two single ends) input into the peak caller per sample with each category is only slightly higher for PC than for MM (median of 55M vs. 50M respectively), but significantly fewer for MM CL (around 15M) (Figure 3-1 B). This is consistent with the strong correlation between the number of single ends used for peak calling and the number of peaks produced (Figure 3-1 D).

The assigned fraction is in the great majority of cases above the required 10% threshold, which complies with the guidelines (Alasoo et al., 2017) (Figure 3-1 C). In the case of MM samples, the variability in the assigned fraction is very high, this can be due to the nature of the different subgroups, for example in terms of overall chromatin accessibility and also due to different Hyperdiploid states which may alter the piling up of chromatin accessibility signal in certain areas (studied in greater detail in chapter 4). PC samples have a higher number of peaks than MM, albeit having similar reads entering peak calling, this could mean that the chromatin accessibility signal is more concentrated in regions that are more spread out for PC, hence explaining its lower assigned fraction in the healthy condition.

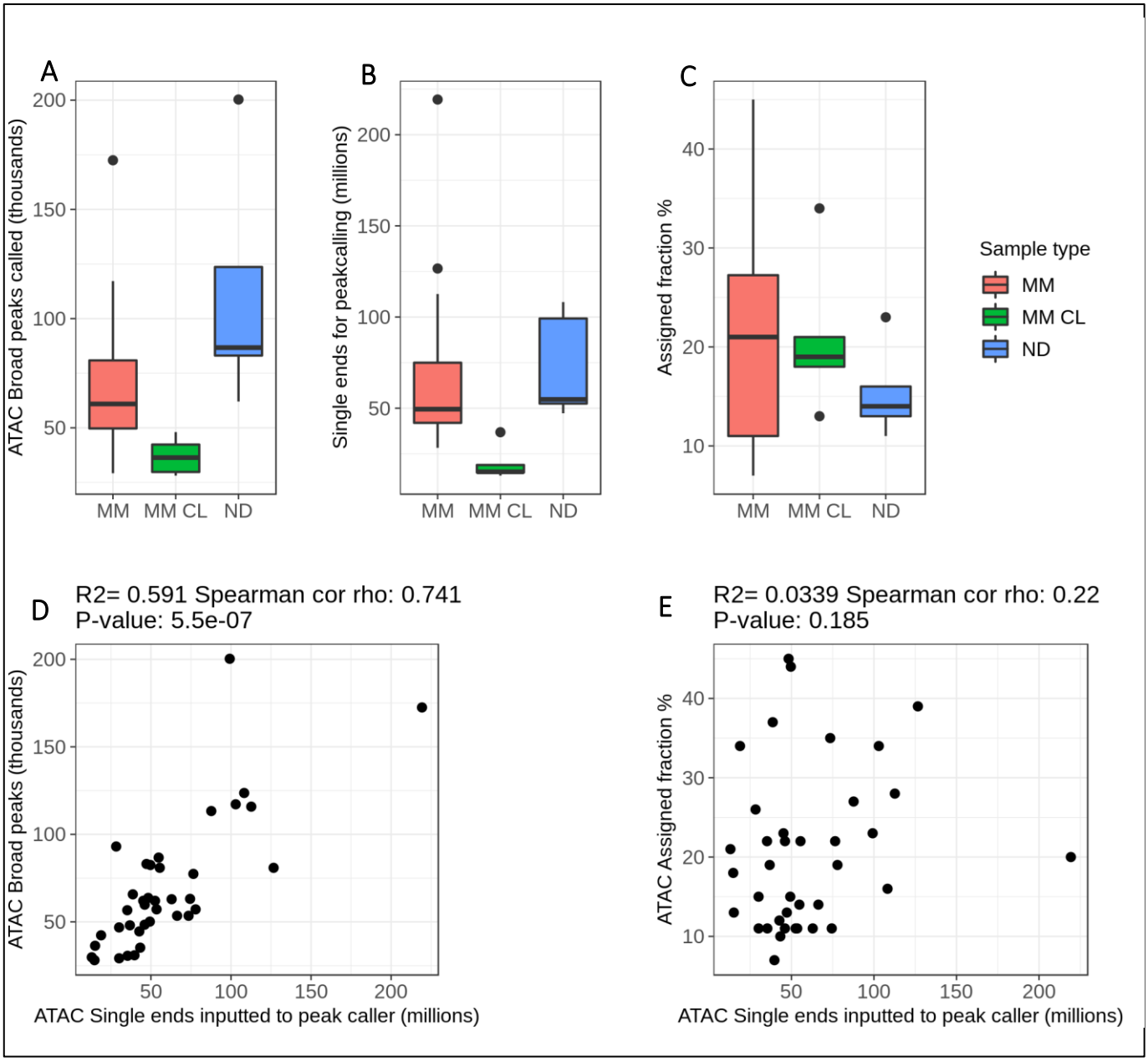


Figure 3-1: Quality control of ATAC-seq data.

A-C: Distribution of ATAC-seq statistics for different groups of all samples used in the study (Table 2-1): MM and ND (PC) primary samples, MM CL (MM cell line). A) Broad peaks called by MACS2. B) Number of filtered reads used for peak calling. C) Assigned fraction. D) Relationship between the number of filtered broad peaks called and the number of filtered reads used for peak calling. E) Relationship between the number of filtered reads used for peak calling and the sample assigned fraction.

In PC, maybe for a fraction of the signal, there is not sufficient sequencing material to surpass the threshold to be considered an accessible region, or perhaps there is more background noise. The higher assigned fraction in MM can be caused by duplications (at the gene, chromosomal arm or full chromosome level). Since amplification of these regions are likely contributing to the disease state, it is possible that they are chromatin active (and accessible regions) that can pile up more signal. However, since the ATAC-seq processing pipeline removes PCR duplicate fragments, the magnitude of this effect would be reduced. A higher number of input reads produces more peaks, which in turn, can mean that a greater proportion of the genome is marked as having open chromatin.

Surprisingly, the number of single ends used to call peaks in each sample does not correlate with the assigned fraction produced (Figure 3-1 E). If chromatin accessible reads were randomly placed, samples with higher number of peaks would be expected to have higher assigned fraction, however, as it is seen between MM and PC, this is not the case and likely points at a genuine sample-specific distribution of open chromatin. As mentioned before, since MM CL have a significantly lower sequencing depth, it is difficult to compare the assigned fraction with the other groups of samples.

3.3.2. Consensus chromatin accessible peaks for primary PC and MM

All the samples were subjected to ATAC-seq read adapter removal, quality check and filtering as explained in the Materials and Methods chapter, section 2.3.1. Then mapped to the human genome, and sample peaks were called (as explained in section 2.4.2).

Consensus chromatin accessible peaks for primary PC and MM (Figure 2-3) were obtained as specified in the “Balanced consensus chromatin accessible peaks for PC and MM” in the Materials and Methods chapter. To obtain a combined set of peaks for downstream uses, first consensus peak sets were obtained separately for primary PC and MM conditions by down-sampling reads in each sample and pooling reads together and calling peaks on the joint read sets. The reason for having two separate pools was to account for the fact that there were an uneven number of samples for PC and MM and pooling all reads together and calling peaks might "hide" PC peak signal behind background MM signal. The down sampled filtered shifted tags per sample used was 28,231,242. Consensus peaks were called by merging all the samples reads first and calling peaks instead of calling sample peaks and then merging, to solve the data snooping issue associated where reads are used twice if sample peaks are called (Lun and Smyth, 2014).

The number of balanced consensus peaks found for PC was 188,065 and for MM 306,709. These were merged (any regions within two peaks of 200bp or less are also considered peak regions) and a joint set of 330,500 consensus peaks were found, referred to as primary MM and PC consensus peaks. These are areas of high chromatin accessibility in at least one of the conditions (MM or PC) considered using the same sample sequencing depth.

3.3.3. DAMM regions

Within consensus chromatin accessible peaks, areas where chromatin is significantly more open in PC compared with MM: a proxy for PC enhancers being inactivated in MM, areas which become more open in MM versus PC (enhancer activation in cancer) and areas which are open in both with not much change (enhancers in both conditions) were obtained. All but

the latter regions were attained and studied. The process followed is detailed in the Materials and Methods chapter, sections 2.4.9.1 and 2.4.13.1.

Several factors complicate this analysis, the five PC samples belonged to three patients and for two of them, there were positive and negative CD19 receptor status on the cell surface samples (Table 2-1). Additionally, sequencing of samples was done in several batches and some samples had sequencing material from more than one batch. Ideally, condition, batch, CD19 status and donor id effects would have been modelled together. However a full model including all these factors simultaneously could not be constructed because the factors were confounding. Thus, three models were constructed, each accounting for a different confounder, and the results combined in a conservative fashion. Each test can be seen in the section 2.4.13.1 of the Materials and Methods.

Since only 33 primary MM and PC were used from the starting samples, the experimental design and its corresponding batches were affected. For some batches, only a single sample passed quality control. For these samples it is not possible to estimate a batch level effect. These samples were put in a single batch (singleton batch). Since the idea is to model the data using additive effects for the covariates, by doing this the effects of the singleton batch are the same for all conditions. For example, taking only the covariates “batch” and “condition”, the singleton batch (batch_singletons) effect will be the weighted average change from the reference batch for all conditions. This is the weighted effect (depending on the number of samples in each subgroup) of:

- Mean of samples in condition PC (for the reference batch level) – mean of samples PC (batch_singletons)
- Mean of samples in condition MM (for the reference batch level) – mean of samples MM (batch_singletons)

If the change for both comparisons above is in the same direction, for example, all conditions having higher average in the batch_singletons, it will be modelled by the singleton batch (batch_singletons) effect, more likely, it will cancel out. When performing this analysis, the assumption is that the batch_singletons effect in the samples in this batch is the same. Moreover, this approach allows significant effects in other batches to be modelled independently of the batch_singletons effect. The MM and PC primary samples tested and the covariates can be seen in Table 2-1.

The results from the tests performed show the number of statistically significant (FDR < 0.05 and absolute $\log_2$FoldChange greater or equal to 1) consensus peak regions for each test:

- 25,565 MM vs. average PC CD19.
- 21,092 MM and PC donor average.
- 24,278 batch and condition significance over batch alone.

The final set of regions was the intersection of these three sets. 18,339 final regions complying with all the above criteria were found, these regions are referred to as DAMM regions (Figure 2-3). There are 365 regions which have more chromatin accessibility in PC and 17,974 regions more accessible in MM. Despite the fact that the healthy state has more regions of DNA marked as accessible chromatin in individual samples (Figure 3-1 A), when observing the consensus regions, it seems that most of the regions are more accessible in the cancer state. This points to a general opening of chromatin and possible activation of enhancers in cancer, however, it has to be taken into account that despite PC having more accessible-marked regions per sample on average, there are nearly six times more MM samples generating consensus accessible regions, which are mostly formed by MM accessible regions. Hence, these results should be taken with caution. Also, these results are consistent with what has been suggested before: it is possible that the ATAC-seq signal has a more diffused spread in PC samples.

The table with all the regions and the corresponding tests results can be found in MM_vs_PC_supervised_analysis/MM_vs_PC_all_DE_ATAC_regions_all_cond.tsv.gz

The table with only the statistically significant regions can be found in MM_vs_PC_supervised_analysis/MM_vs_PC_sign_DE_ATAC_regions_all_cond.tsv.gz

3.3.4. MMPC enhancers

Since the objective of this work was to locate DNA enhancer regions regulating genes, it is important to distinguish promoters from enhancers (and keep only the later). With this in mind, all RNA-seq from PC and MM samples (including cell lines, with the exception of around 60% of the sequencing material from the cell line RSJJN3.1 which wasn't included due to initial mislabeling) was used to assemble a transcript set. Start locations of multi-exon transcripts from the assembly were used as potential promoters, to which were added start sites of annotated multi-exon transcripts (see Materials and Methods chapter, section 2.3.5). These regions were filtered out from the DAMM regions (see Materials and Methods chapter, section 2.5.1.2). Single exon transcripts may be eRNA (Ding et al., 2018) and we wanted to include these regions in the analysis. This left approximately half of the starting regions (9,527), referred to as "MMPC enhancers" (Figure 2-3). Full results can be found in: MM_vs_PC_supervised_analysis/MM_vs_PC_sign_DE_ATAC_regions_all_cond_no_TSS.tsv.gz

3.3.5. MM enhancers

Once the MMPC enhancers are found, the MM exclusive enhancers (referred to as "MM enhancers") are determined as it is explained in the section 2.4.14.2 of the Materials and Methods chapter and Figure 2-3. Out of the 9,527 MMPC enhancers shared between conditions only 1,462 regions are thought to be activated exclusively in the malignant state. The table with the MM enhancers can be found in:

MM_vs_PC_supervised_analysis/MM_vs_PC_sign_OE_ATAC_regions_all_cond_no_TSS.tsv.gz

3.3.6. DEMM genes

Once the MMPC and MM candidate enhancers had been identified, I sought to find a matching set of genes which are significantly changed expression (DEMM genes). Testing of the expression changes was performed similarly to the ATAC-seq analysis described above, including the same tests (see Materials and Methods, section 2.3.6.1). The primary samples tested and the covariates can be seen in (Table 2-1). Significant genes were those with FDR < 0.05 and absolute $\log_2$FoldChange greater or equal to 1.5 for each test:

- 2,660 MM vs. average PC CD19.
- 2,575 MM and PC donor average.
- 1,214 batch and condition significant over batch alone.

To obtain genes complying with all three criteria, the significant genes for each test were intersected. 806 final genes complying with all the above criteria were found. Full results can be found in MM_vs_PC_supervised_analysis/MM_vs_PC_all_DE_genes_all_cond.tsv.gz

The table with only the statistically significant genes can be found in MM_vs_PC_supervised_analysis/MM_vs_PC_sign_DE_genes_all_cond.tsv.gz

To find which cell processes are affected in terms of gene expression in the PC to MM transition, a gene ontology analysis was performed testing the category enrichment in the 806 DEMM genes compared with the 57,992 genes which are quantified (specified in the Materials and Methods chapter, section 2.9.1). In total 202 categories were overrepresented (FDR 0.1) in the DEMM genes. The table is filtered to show only Molecular Function and Biological Process categories with fold enrichment equal or greater to 2, only some categories are shown in Table 3-1, the full table with all the enriched categories can be seen in:

MM_vs_PC_supervised_analysis/GO_pan_MM_vs_PC_DE_genes_Wallenius.xlsx

Some categories such as regulation of cell migration, regulation of cell motility, negative regulation of extrinsic apoptotic signalling pathway, blood vessel morphogenesis, protein

binding involved in protein folding, extracellular matrix structural constituent and angiogenesis are general cancer categories. While, some are MM specific such as positive regulation of interleukin-8 production which may contribute in MM metastasis, cell proliferation and angiogenesis (summarized in Aggarwal et al., 2006). Other categories are typical of PCs and the immune response such as activation of immune response, biological adhesion, antimicrobial humoral response and response to bacterium.

KEGG analysis for these genes shows only one over-represented category for Systemic lupus, the literature already shows connections (albeit rare) between MM and Lupus (Bila et al., 2007; Choi et al., 2010).

Term	Ontology	FDR	Fold
biological adhesion	BP	0.00032	2
regulation of cell migration	BP	0.00484	2
extracellular matrix structural constituent	MF	0.00553	4
regulation of cell motility	BP	0.01150	2
regulation of immune response	BP	0.01150	2
negative regulation of extrinsic apoptotic signalling pathway	BP	0.01263	5
positive regulation of interleukin-8 production	BP	0.02117	7
antimicrobial humoral response	BP	0.02359	4
activation of immune response	BP	0.02747	2
blood vessel morphogenesis	BP	0.04532	2
response to bacterium	BP	0.04777	2
protein binding involved in protein folding	MF	0.04802	10
angiogenesis	BP	0.06619	2

Table 3-1: Gene Ontology Categories using Wallenius approximation for DEMM genes between MM vs. PC.

BP: Biological Process, MF: Molecular Function. Fold: Fold enrichment for the category.

3.3.7. OEMM genes

After determining the deregulated genes between the cancer and healthy condition, the gene activation program required for Myelomagenesis is determined. OEMM genes were identified as explained in the Materials and Methods chapter, section 2.3.6.5. 548 OEMM genes were identified, they can be found in:

MM_vs_PC_supervised_analysis/MM_vs_PC_OE_RNA.gz

To find which gene categories are enriched in these genes and find MM associated biological processes and functions, gene ontology enrichment was performed (see section 2.9.2 in Materials and Methods chapter). 221 categories from 548 genes are found, see:

MM_vs_PC_supervised_analysis/GO_pan_MM_vs_PC_OE_genes_Wallenius.xlsx

The results show new categories emerging compared with DEMM genes enrichment. Some are more general, such as DNA and chromatin remodeling is present (DNA packaging, chromatin assembly, chromatin assembly and chromatin disassembly categories). Other categories are more MM specific, one example is response to Bone morphogenetic proteins (BMP) and cellular response to BMP. Additionally, positive regulation of cell migration (DEMM genes only have regulation of cell migration) may be cancer related. Another relevant pathway with enriched genes is ossification, which is the process of cartilage to bone formation by osteoblasts and can be thought of the reverse of osteolysis.

Some ontology categories are enriched both in OEMM and DEMM genes, for example nucleosome organization/assembly, which can influence enhancer state. General cancer categories include negative regulation of extrinsic apoptotic signalling pathway and negative regulation of apoptotic signalling pathway which contribute in the Myelomagenesis state (OEMM genes also include the negative regulation of extrinsic apoptotic signalling pathway in absence of ligand category). Another class which is also found is negative regulation of megakaryocyte differentiation (negative regulation of cell differentiation and negative regulation of myeloid cell differentiation only enriched in OEMM genes). It is known that until recently PC has been considered terminally differentiated but new research is emerging with new differentiation states within PC with relevant prognostic implications (Paiva et al., 2017), it is therefore possible that MM potentially acquires gene expression profiles similar to other cell types.

3.3.8. MMPC enhancers regulating DEMM protein coding genes

To identify MMPC enhancers regions that may regulate DEMM genes, I linked regions and genes based on a maximum genomic linear distance of 1Mb. It has been previously observed that 90% of promoter interacting regions were found to be within 1Mb (Javierre et al., 2016), so this distance threshold should cover the great majority of genuine interactions. Not all enhancer – gene pairs found using this method will be biologically meaningful interactions, but it is a good starting point to generate a set of candidates for further study.

Pairing MMPC enhancers with all DEMM protein coding genes within 1Mb leads to 124,728 region - pairs (see section 2.5.1.2). Of these 2,698 pairs involve simultaneously MMPC enhancers and DEMM protein coding genes. These 2,698 interactions are therefore referred to as “MMPC enhancers regulating DEMM protein coding genes”.

Examples of candidate interactions between MMPC enhancers and DEMM genes are shown in the context of B-cell line interactions (Figure 3-2, Figure 3-3 and Figure 3-4). As can be seen in

Figure 3-2, ABCA1, a gene involved in ATP-binding cassette (ABC) which transports various molecules across membranes, has interactions in B-cells with the proposed enhancer (chr9:104,968,040-104,968,931), marked in red. ABCA1 is very expressed in PC and MM with the latter having a 14-fold higher expression, SNPs in this gene in MM patients have been associated with Thalidomide-Related Neuropathy (Johnson et al., 2011). The enhancer thought to interact with ABCA1 is accessible in PC but more accessible in MM. BMP4 is a Bone Morphogenetic Protein anticorrelated with proliferation and apoptosis in MM cultures (Fukuda et al., 2006) and correlated with resistance to Bortezomib treatment in Myeloma (Grčević et al., 2010). With nearly 100 times more expression in MM, it is thought to interact with the region chr14:53,968,790-53,969,329 (having nearly quadruple the accessibility in MM compared to PC) and having a high enrichment of contacts in B-cells (Figure 3-3). Finally, HGF, a gene thought to be regulated in MM by H3K27Ac-marked enhancers (Jin et al., 2018) is also present in my study with multiple related enhancers. As can be seen in Figure 3-4, the HGF promoter (significantly overexpressed in MM) has visible interactions (marked in red) with one of the enhancer regions delineated in my study: chr7:81,886,300-81,888,146, which is significantly over accessible in MM. This gene regulates cell growth, motility and morphogenesis, in MM it has been targeted previously (Rao et al., 2018) and when bound to the surface of extracellular vesicles derived from Myeloma, it can activate HGF/c-Met signaling of osteoblast-like cells (Strømme et al., 2019). Furthermore, interactions of MM cells with bone marrow stromal cells expressing CXCL12 can cause HGF OE among other genes, involved in angiogenesis and osteoclastogenesis (Ullah, 2019).

3.3.8.1. MMPC enhancers regulating DEMM protein coding genes reproducible in MM CL

To improve the inference of candidate enhancers, information about open chromatin regions and DE genes from MM CLs was cross checked with MMPC enhancers regulating DEMM protein coding genes. The aim is to identify reproducible interactions from the primary sample analysis which may also be present in the MM CLs to test them in the CLs.

To obtain DE genes between MM CLs and PC, RNA-seq data was used as specified in the Materials and Methods chapter (section 2.3.6.3). Since, only 1/5 PC samples have an overlapping batch with MM CL samples, DE genes are directly obtained.

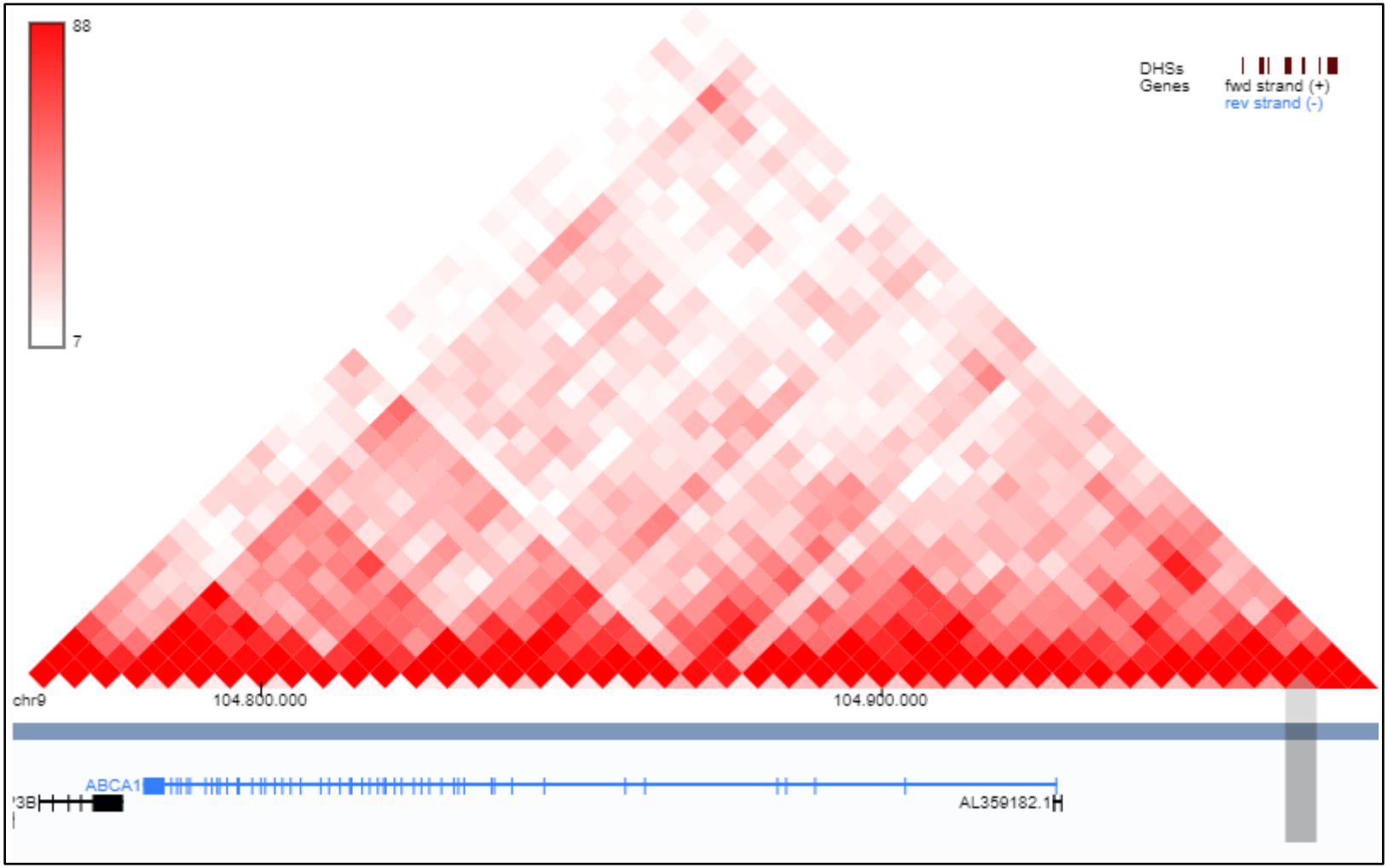


Figure 3-2: chr9:104,968,040-104,968,931 interactions with ABCA1 in B cell line (GM12878).

Hi-C raw data on hg38 assembly at 5Kb resolution (Rao et al., 2014), location of the chr9:104,968,040-104,968,931 enhancer marked with a grey column (bottom part). Visualization obtained using the Feng Yue lab at Northwestern University Genome Browser (Yue, n.d.). *Scale of interactions shown on the top left.*

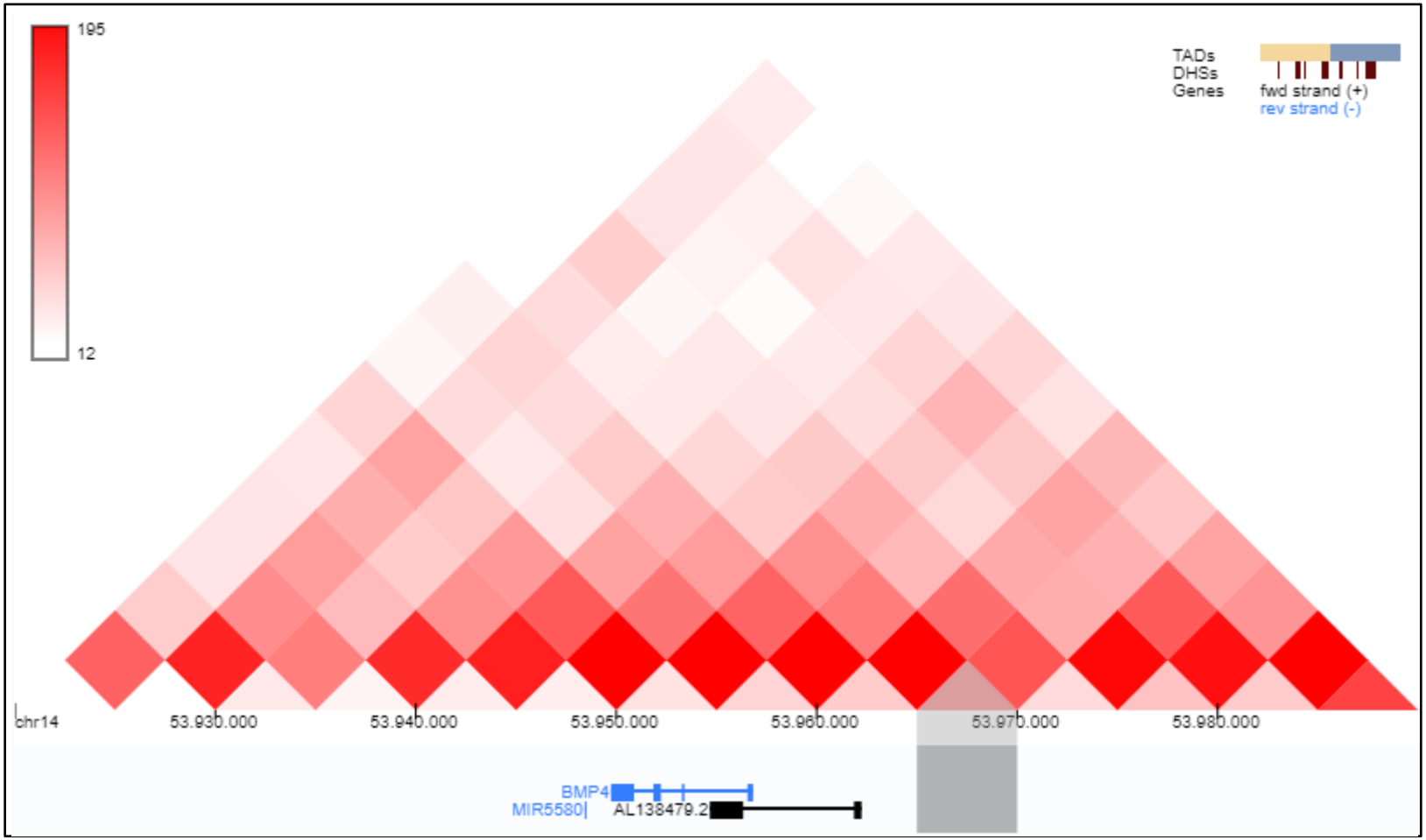


Figure 3-3: chr14:53,968,790-53,969,329 interactions with BMP4 in B cell line (GM12878).

Hi-C raw data on hg38 assembly at 5Kb resolution (Rao et al., 2014), location of the chr14:53,968,790-53,969,329 enhancer marked with a grey column (bottom part). Scale of interactions shown on the top left.

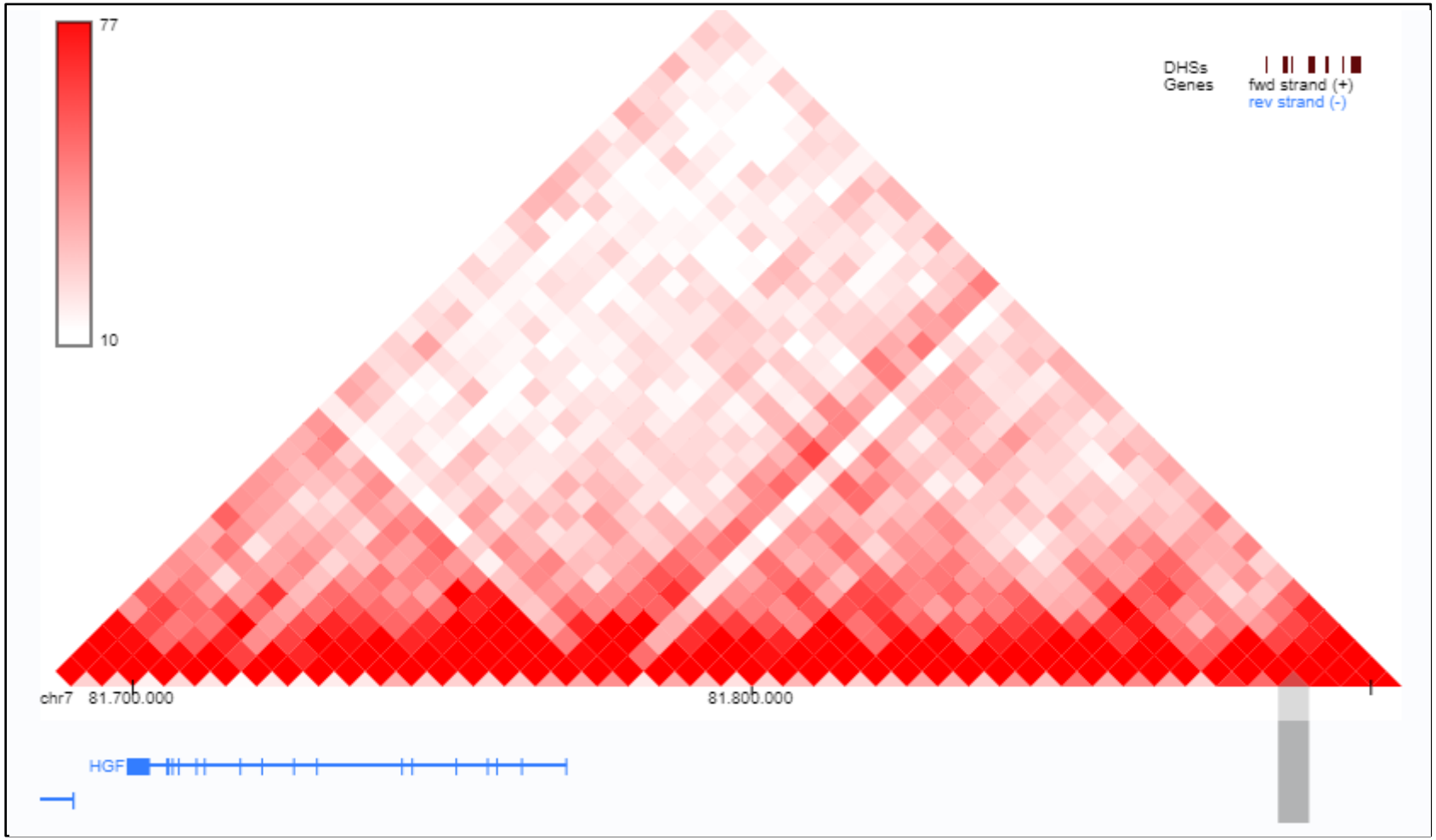


Figure 3-4: chr7:81,886,300-81,888,146 interactions with HGF in B cell line (GM12878).

Hi-C raw data on hg38 assembly at 5Kb resolution (Rao et al., 2014), location of the chr7:81,886,300-81,888,146 enhancer marked with a grey column (bottom part). Scale of interactions shown on the top left.

The results from the test performed show that there are 8,463 genes significantly DE (FDR < 0.05 and absolute log_2FoldChange greater or equal to 1.5). The table showing these genes can be found in:

MM_vs_PC_supervised_analysis/MM_CL_vs_PC_DE_genes.tsv.gz

As mentioned previously, MM CL samples' ATAC-seq material is sequenced at a significantly lower depth, not complying with ATAC-seq guidelines (Ackermann et al., 2016) and therefore accessibility differential analysis between MM CLs and PC is not performed. To compensate for this, while still finding accessible chromatin regions shared between MM patient samples and MM CL samples, individual MM CL sample chromatin accessible peaks were obtained (following the method in section 2.4.2 of the Materials and Methods chapter). MMPC enhancers regulating DEMM protein coding genes were marked with overlapping sample MM CL accessible chromatin regions and DE genes between MM CLs and PC.

The 2,698 MMPC enhancers regulating DEMM protein coding genes were annotated with the information from the MM CL samples. 75 region – gene pairs are found to be reproducible interactions in PC vs. MM CL. These final regions have the following characteristics:

- All primary samples ATAC and RNA tests (MM vs. average CD19 PC, MM vs. PC donor average and condition accounting for batch) significant (FDR 0.05) and with absolute log_2FoldChange equal or greater than 2.
- Containing a gene DE in Cell line MM vs. primary PC.
- Having 3 or more cell lines with overlapping chromatin accessible peaks.

These regions are referred to as "MMPC enhancers regulating DEMM protein coding genes reproducible in MM CL". A full list of these candidate enhancers is contained in: MM_vs_PC_supervised_analysis/MM_vs_PC_all_DE_ATAC_DE_RNA_1Mb.xlsx (those reproducible in cell lines are marked "Strong Candidate").

The MMPC enhancers regulating DEMM protein coding genes were studied to find clues to the chromatin state of these potential enhancer regions on other cell types (section 2.5.3 in Materials and Methods chapter). This will hopefully provide insights of whether these regulatory regions are novel MM enhancers, or are active in other cell types while also becoming active in MM. For this, the chromatin state segmentations (see section 1.4 for more details) using different histone modifications (Albrecht et al., 2016) and ChromHMM (Ernst and Kellis, 2017) were used (Figure 3-5**,** Figure 3-6 and Figure 3-7).

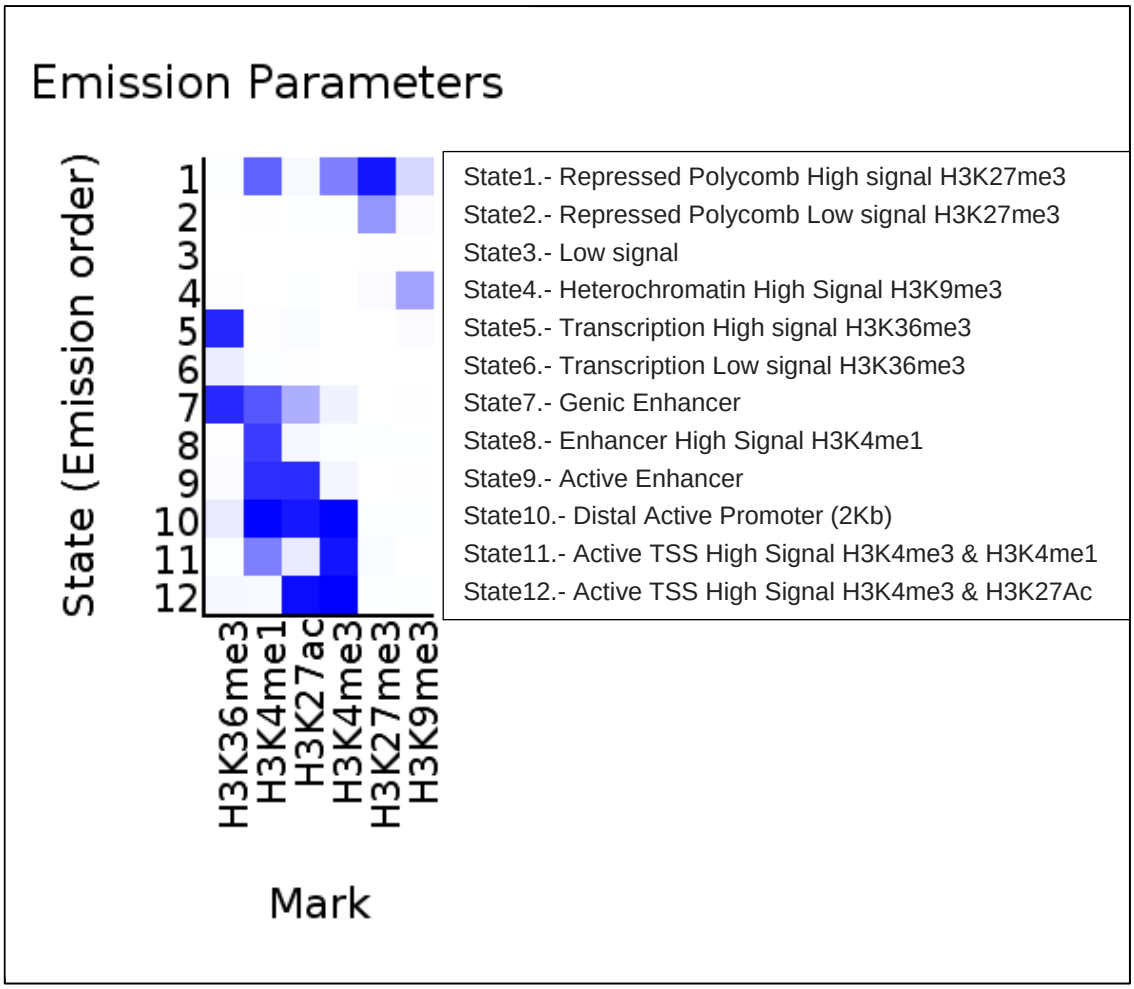


Figure 3-5: Relationship between each histone modification signal and the state assigned by ChromHMM. From (Albrecht et al., 2016) *and* (Blueprint_project, 2016).

It must be noted that state 12 can be considered a possible active enhancer state in this study. This is because, state 12 is considered an active TSS, but TSS are removed from the regions under study and recent findings (Henriques et al., 2018) have found at least some enhancers to be marked by H3K4me3 instead of H3K4me1, perhaps as a consequence of multiple rounds of transcription generating eRNA (Soares et al., 2017). Arguably, state 10 (considered distal active promoters) can also be considered an active enhancer state since it contains H3K27ac, H3K4me3 and H3K4me1 signals, although it has been shown that H3K4me1 "was anticorrelated with polymerase density" (Wissink et al., 2019).

From the 2,698 MMPC enhancers regulating DEMM protein coding genes, 1,959 unique regions (33 of them more accessible in PC) are used overlapping 11,220 200bp genome windows which can be seen in:

MM_vs_PC_supervised_analysis/MM_vs_PC_all_DE_ATAC_DE_RNA_1Mb_enh_other_cells.gz

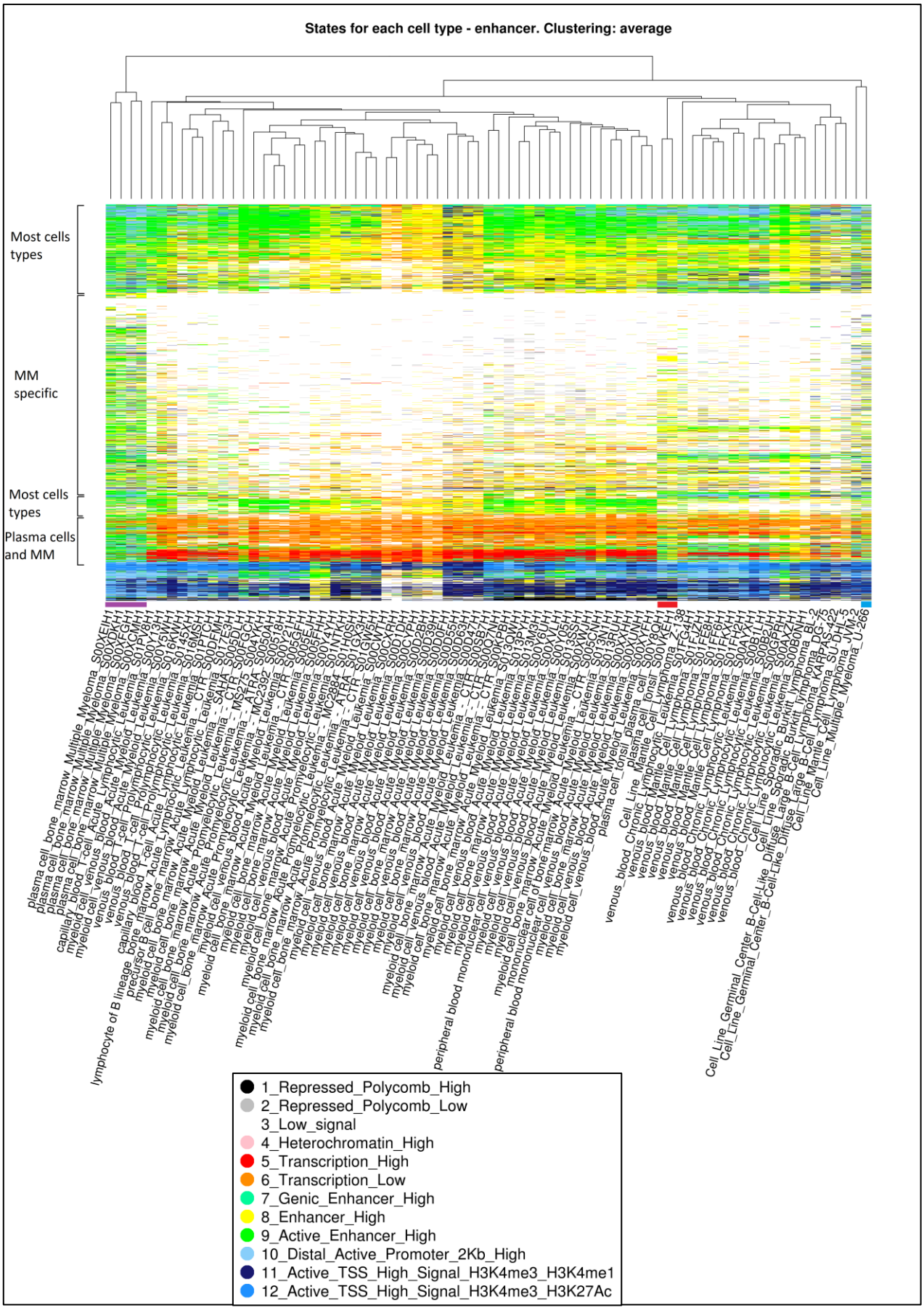


Figure 3-6: MMPC enhancers regulating DEMM protein coding genes in disease cell types.

Each row represents a 200bp window containing a MMPC enhancer regulating a DEMM protein coding gene. Each column reflects a Blueprint disease cell line also including MM primary samples (marked in purple between the cell line label and the heatmap), plasma cells in red and MM CL in blue. Each cell in the heatmap contains the assigned chromatin state for the 200bp window and cell type. 200bp enhancer regions and disease cell types are hierarchically clustered by Gower distance and average linkage. Different types of enhancers shown on the left.

The cell types are separated into disease (Figure 3-6) and healthy (Figure 3-7) cell types, both plots include the PC, MM tissues and MM U-266 CL for reference. In general, the disease heat map contains an enrichment of the active enhancer state in the majority of regions only for the MM tissue (Figure 3-6 columns marked in purple between the cell line label and the heatmap have predominantly a green state), which confirms that the regions being surveyed have the potential to be genuine enhancers. Within this heatmap, very distinct groups are formed, one of them is the "MM specific" group, these are considered novel in MM compared with other cell types and have high chromatin accessibility, H3K4me1 and H3K27ac. Figure 3-6 also contains "plasma cells and MM" candidate enhancers which might be tissue specific and important in Plasma cell biology and another set of regions ("Most cell types") were those that tend to be acetylated in MM (and considered active: green or cyan) and either active or with the potential for activation on other cell lines (yellow). Finally, there is a subgroup of enhancers (Figure 3-6 bottom rows, not labelled) which are related with active TSS states, since state 12 can also be considered an enhancer state, it is possible that this group also comprises active enhancers across different cell types.

The heat map showing healthy cells (Figure 3-7) has less distinct blocks of regions. Additionally, there are two sets of regions which are common to the B-cell lineage, including PCs and MM enhancers (marked as "MM, Plasma and B-cell") which are active in these cell types. The bottom set common to B-cell lineage enhancers are only marked as enhancers in the B-cell lineage (as well as PC and MM). Some regions are marked as "MM specific" (novel) enhancers while a block of regions seems to be either active or with the potential for activation in non-MM cell types. Enhancers labelled as "other cell types" in Figure 3-7 have predominant enhancer state throughout many cell types. A subgroup of these regions also has the active enhancer state in neutrophils, as has been previously found, MM cells from a patient resembled morphological characteristics of mature neutrophils (Wei Wang and Shimin Hu, 2015). As it is the case with the healthy cell types heat map above, there are H3K4me3 enriched regions which can also be considered enhancers (blue dominant rows at the bottom).

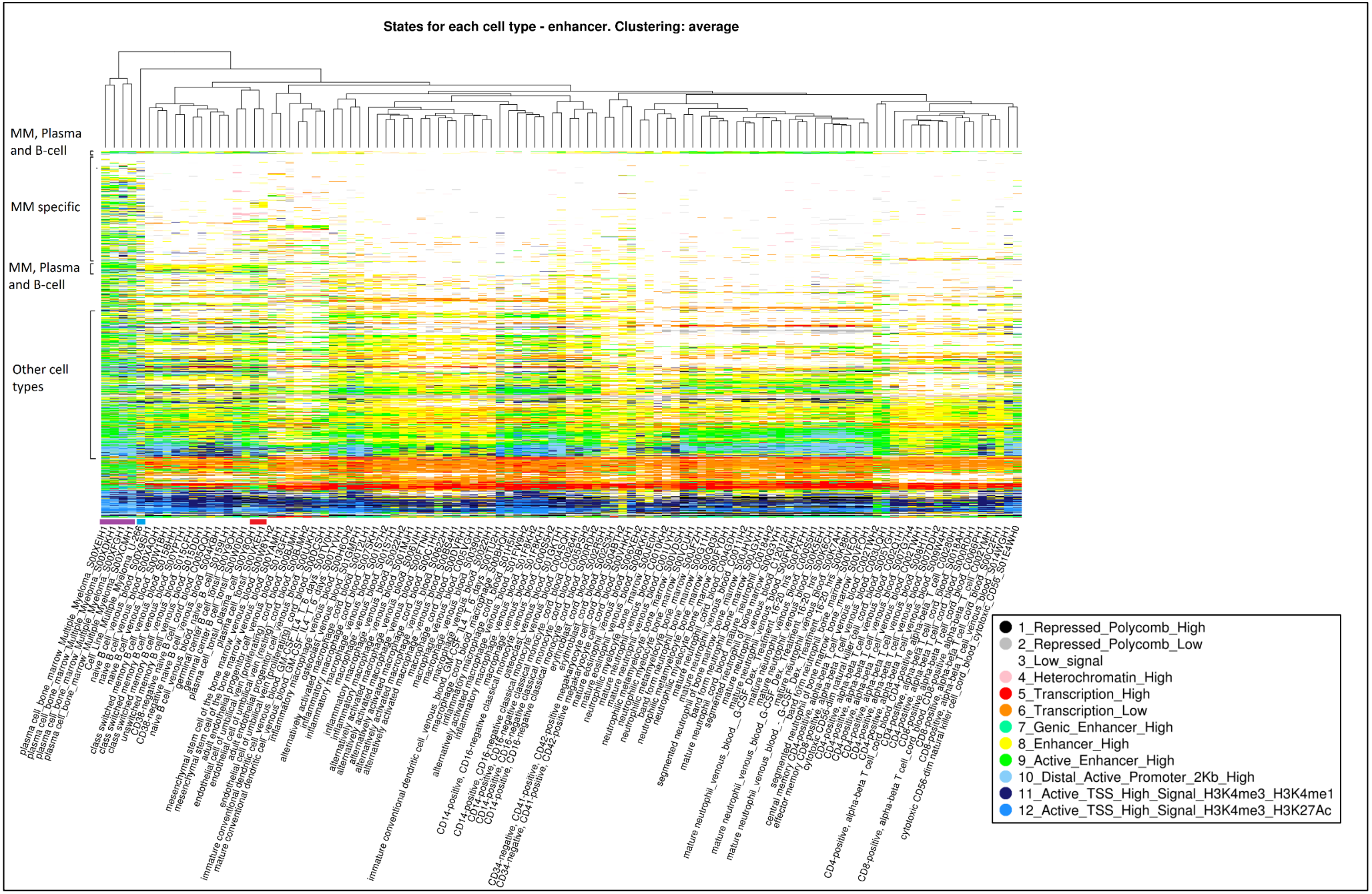


Figure 3-7: MMPC enhancers regulating DEMM protein coding genes in healthy cell types.

Each row represents a 200bp window containing a MMPC enhancer regulating a DEMM protein coding gene. Each column reflects a Blueprint healthy cell line also including MM primary samples (marked in purple between the cell line label and the heatmap), plasma cells in red and MM CL in blue. Each cell in the heatmap contains the assigned chromatin state for the 200bp window and cell type. 200bp enhancer regions and healthy cell types are hierarchically clustered by Gower distance and average linkage. Different types of enhancers shown on the left.

3.3.9. MM enhancers near OEMM protein coding genes

Now that I have obtained a list of putative PC and MM enhancers and possibly related altered genes and determined how the chromatin state in other cell types correlate, an important piece of knowledge to determine is which enhancers are becoming active in cancer in terms of chromatin accessibility and their target OEMM genes.

For this, the table produced in the last section containing the 2,698 MMPC enhancers regulating DEMM protein coding genes was further filtered removing PC more accessible and expressed regions and genes respectively (see section 2.5.1.3 in Materials and Methods). 481 regions were identified that are more open in MM and are within 1Mb of an upregulated gene. 311 of these interactions are not overlapping PC open chromatin areas, which points at these subset of enhancers being active in MM and inactive in PC. These are referred to as "MM enhancers regulating OEMM protein coding genes".

The table containing these details can be observed in:

MM_vs_PC_supervised_analysis/MM_vs_PC_OE_ATAC_OE_RNA_1Mb.gz

These regions, which represent potential MM specific enhancers, together with nearby OEMM genes were studied in detail. One of the four MM interactions with no accessibility in PC and accessible chromatin in all five MM CLs studied and having gene overexpression in them (thereby having a high chance of being reproducible in all MM CLs studied) is shown in Figure 3-8. This interaction involves a MM enhancer (chr7:124,059,178-124,059,929, marked in black below the signal tracks on the left) it is thought to regulate the GPR37 gene (Figure 3-8 right-side panel) at a distance of around 700Kb and increase significantly its expression. As can be seen, the chromatin accessibility is significantly open at the enhancer for most of the MM and MM CLs samples while it is likely in the heterochromatin state for the PC samples. Additionally, the GPR37 promoter, while being somewhat accessible in PC samples, it is more accessible in MM and CLs. The GPR37 gene has been previously linked with the proliferation of MM cells (Huang et al., 2014).

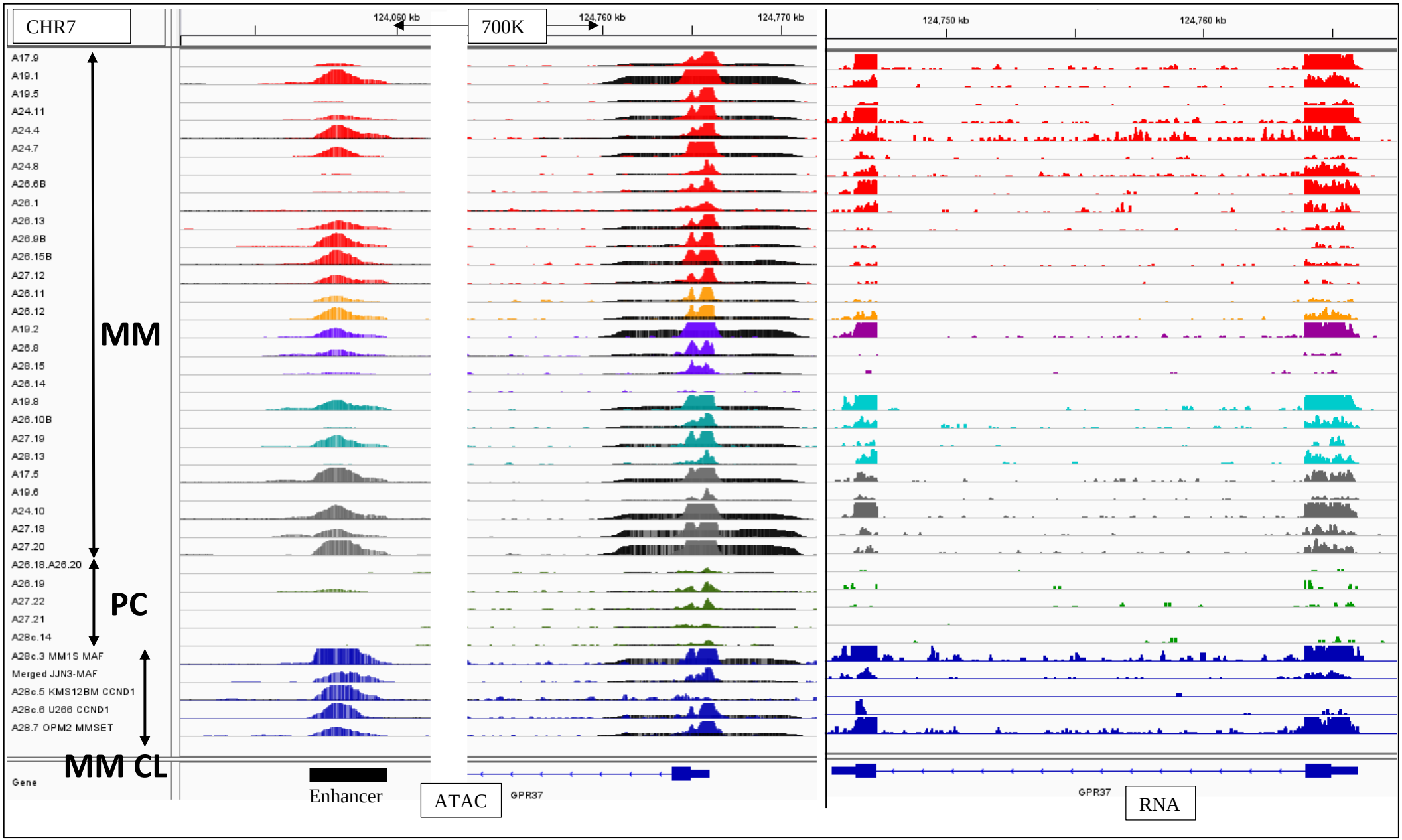


Figure 3-8: Samples chromatin accessibility and gene expression for a candidate enhancer and GPR37 gene

Left: ATAC-seq tracks show the different chromatin accessibility profiles of the different samples (and whether they are primary MM or PC or MM CLs) in different colours depending on the different subgroups: HD in red, MAF in orange, CCND1 in purple, MMSET in cyan, cytogenetically non-annotated samples in grey, PC samples in green and MM CL in dark blue. The baseline calculated chromatin accessibility signal (noise) is overlaid in black for all samples. On the left section, the enhancer at location chr7:124,059,178-124,059,929 (marked in black below the signal tracks), on the centre, the chromatin accessibility for the GPR37 gene. The scale for each track is 0-1.5 (fragment pileup normalized for each sample per million reads).

Right: RNA-seq signal corresponding to each of the ATAC-seq tracks (each RNA sample is horizontally aligned with its patient's ATAC sample, ATAC-RNA sample correspondence can be seen in Table 2-1), the colours for the RNA-seq signal are the same as for the ATAC-seq signal. The scale for each track is 0.1 (reads mapping normalized by sample million reads mapped).

It is important to validate whether these candidate enhancers are thought to regulate the target genes through methods such as 3C or CRISPR-dCas9 repression in MM primary samples.

3.3.10. Chromatin accessible regions in the vicinity of MM and PC genes

Previous reports suggest that enhancers can interact with multiple promoters. Studies have found that 90% of these interactions occur within the 1Mb range (Javierre et al., 2016). In general enhancers are more likely to interact with their target genes with a frequency inversely proportional to the distance separating both (Corces et al., 2018). In this section I will investigate the relationship between genes and the number of putative enhancers within 1Mb. If these enhancers are regulatory, genes thought to be regulated should be expected to have a greater number of candidate enhancers within regulatory range than any gene (which may or not be regulated).

In particular, the number of MMPC enhancers within 1Mb of all DEMM genes (includes DEMM genes independently of being within 1Mb of a MMPC) and genes in general is compared. Similarly, it is also observed whether there is an enrichment in the number of MM enhancers in the regulatory range of all OEMM genes (includes OEMM genes independently of being within 1Mb of a MM enhancer) compared with genes in general. The presence of MMPC and MM enhancers are studied in the context of protein coding and non-protein coding genes as specified in section 2.5.2 from the Materials and Methods. Furthermore, the proportion of genes in these categories with one or more enhancer(s) within 1Mb are observed.

Overall, a greater proportion of all DEMM protein coding genes have more MMPC enhancers within the regulatory range where most enhancer – promoter interactions occur (Javierre et al., 2016) compared with protein coding genes (Figure 3-9 A). This points to the idea that in

general, DEMM protein coding genes have the potential for a greater regulatory influence from MMPC enhancers than regular protein coding genes. For non-protein coding genes, there is a higher proportion of DEMM genes with no MMPC enhancers within regulatory range compared to non-protein coding genes in general (Figure 3-9 B). There is however, a greater ratio of non-protein coding genes DEMM genes with eight or more MMPC enhancers compared to non-protein coding genes in general.

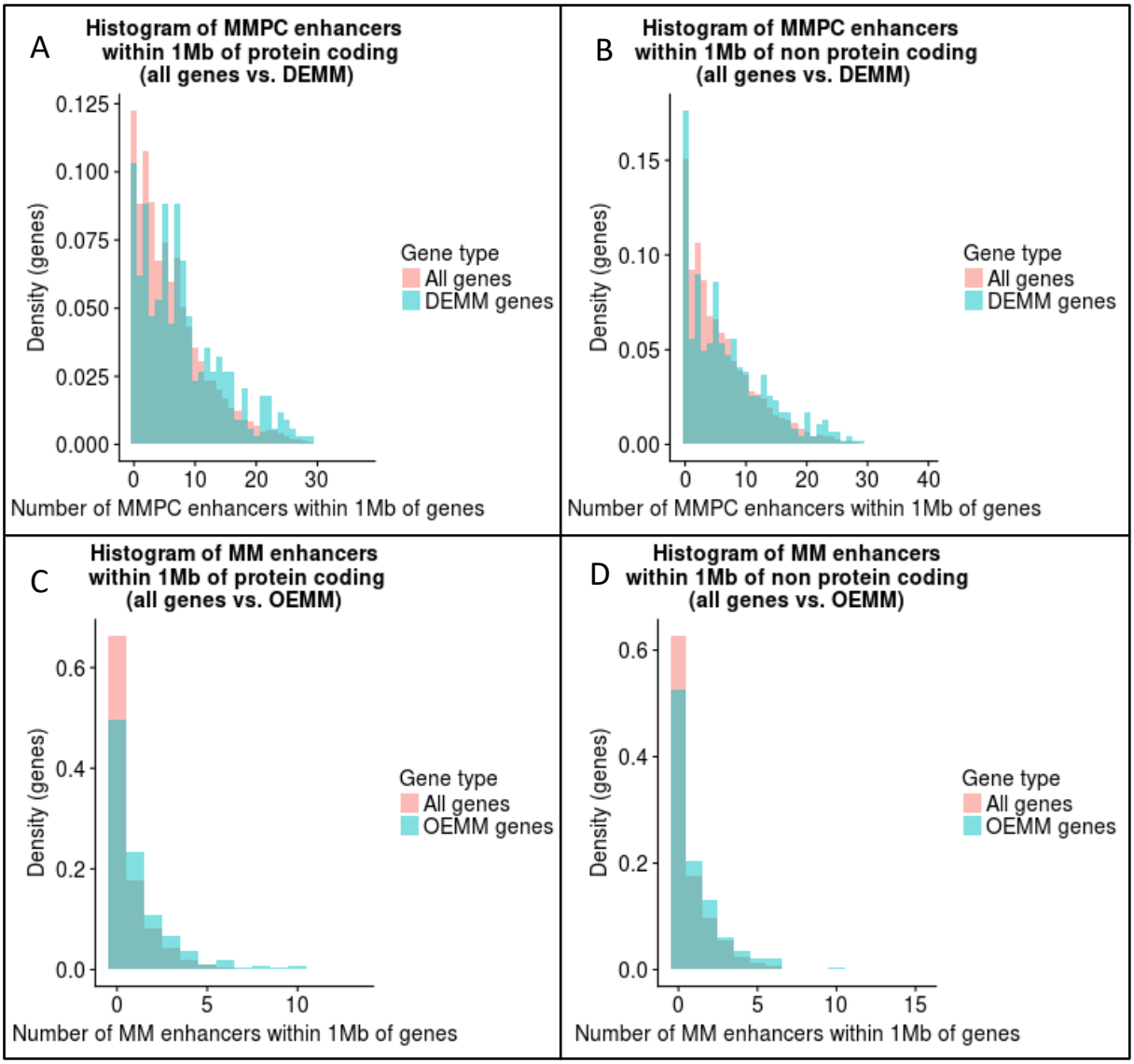


Figure 3-9: Histograms showing different sets of candidate enhancer regions intersecting different gene sets.

A: Number of MMPC enhancers within 1Mb of all DEMM protein coding genes (whether within 1Mb of a MMPC enhancer or not) and protein coding genes in general.

B: Number of MMPC enhancers within 1Mb of all DEMM non-protein coding genes (whether within 1Mb of a MMPC enhancer or not) and non-protein coding genes in general.

C: Number of MM enhancers within 1Mb of all OEMM protein coding genes (whether within 1Mb of a MM enhancer or not) and protein coding genes in general.

D: Number of MM enhancers within 1Mb of all OEMM non-protein coding genes (whether within 1Mb of a MM enhancer or not) and non-protein coding genes in general.

The great majority of DEMM protein coding (90%) and DEMM non-protein coding genes (82%) have at least one MMPC enhancer in the selected regulatory range, with no significant difference compared with the ratio from all protein coding (88%) and non-protein coding genes (85%) containing one or multiple MMPC enhancers respectively (Figure 3-10 “Coding, MMPC” and “Non-coding, MMPC” categories).

When considering MM enhancers with OEMM genes and all genes, the OEMM genes tend to be enriched for having more MM enhancers nearby compared with genes in general, whether protein coding or non-protein coding (Figure 3-9 C and D respectively). While only around one third of all protein coding genes contain nearby MM enhancers, around half of all the OEMM protein coding genes do (Figure 3-10 “Coding, MM” category). Similarly, 37% of all non-protein coding genes have nearby MM enhancers but 48% of all non-protein coding OEMM genes have (Figure 3-10 “Non-coding, MM” category). When it comes to genes with one or more MM enhancer(s) in range, the proportion from all OEMM genes is statistically significantly higher, both for protein coding and non-protein coding genes.

There is a statistically significant increase in the proportions of MMPC enhancers near DEMM genes for protein coding compared with non-protein coding genes (Figure 3-10 “Coding, MMPC” and Non-coding MMPC” categories) but not between protein coding and non-coding proportions of OEMM genes with MM enhancers nearby (Figure 3-10 “Coding, MM” and Non-coding MM” categories).

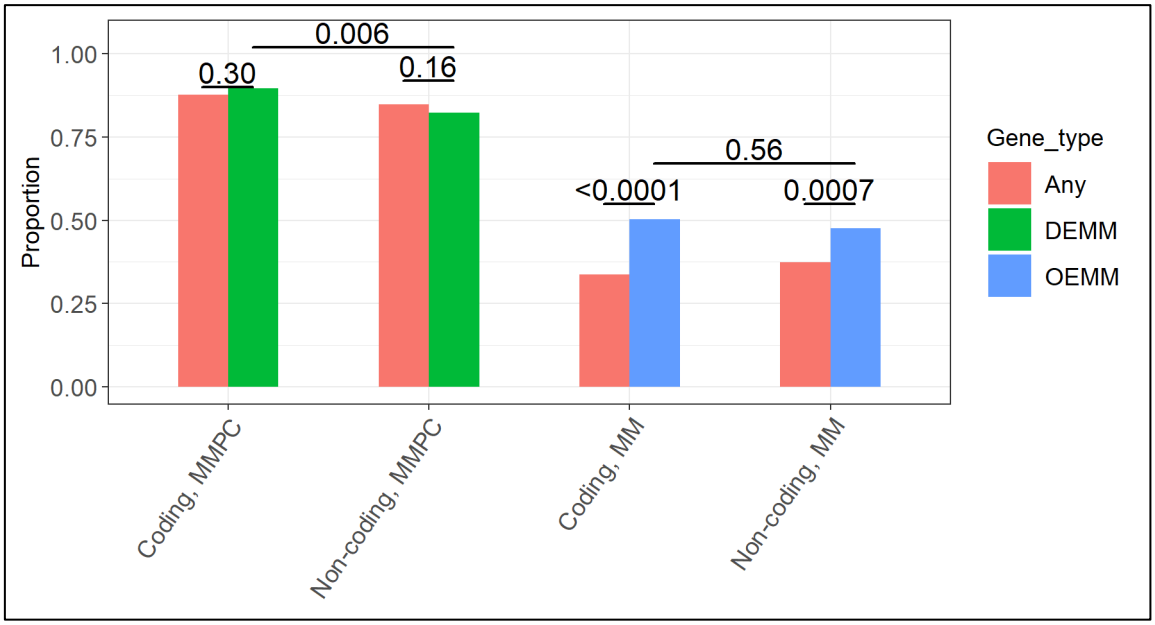


Figure 3-10: Proportion of genes with MMPC and MM enhancers.

The proportion of any gene (pink bars), DEMM genes (green) and OEMM (blue) containing at least 1 enhancer (MMPC or MM) within 1Mb are shown. Comparisons are divided into different categories for protein coding and non-protein coding genes (see column labels, down).

Pairwise Pearson's chi-squared test using Benjamini & Hochberg (1995) correction p-adjusted values between the groups shown above.

Taken together, a similar proportion of DEMM genes are in the range of influence of MMPC enhancers compared with any random gene. There is, however, a statistically greater proportion of OEMM genes containing MM enhancers in the regulatory space than what would be expected by chance.

Presumably, in this analysis, MMPC enhancers regulating DEMM genes capture two effects: MM activation and deactivation of enhancer – promoter interactions compared to PC. However, since these two effects are considered simultaneously, a DEMM gene that is downregulated in MM quantifies all MMPC enhancers within 1Mb (including MM deactivated but also activated enhancers with respect to PC). The inverse effect will also be quantified (PC activated MMPC enhancers near DEMM genes downregulated in PC). On the other hand, when only considering MM enhancers within the regulatory range of OEMM genes, only MM activated interactions are considered and the differences between the number of MM enhancers near OEMM and regular genes is significant. This is to be expected in a scenario

where MMPC enhancers and particularly MM enhancers are thought to regulate DEMM genes and OEMM genes respectively.

3.3.11. TF binding in MM enhancers near OEMM protein coding genes

Another important question to address is which TF (or TFs) are critical for enhancer activation in MM as compared with the PC counterparts provides critical knowledge on the mechanism which favors the cancer state. To identify potential regulators, motif enrichment was performed on the 311 MM enhancers regulating OEMM protein coding genes, by getting the unique regions (see Materials and Methods, section 2.6.1). This set of regions is thought to be most representative of the MM enhancers because it takes into account regions which are significantly more chromatin accessible in MM compared with PC and it doesn't include TSS or PC chromatin accessible areas. Additionally, only the regions being within 1Mb of a OEMM gene are considered.

Enriched motifs from MEME and Dreme were clustered together to group similar motifs, and a reference motif selected to represent the cluster. Significance was measured at FDR 0.05 using the E-value metric, which, according to the Meme Tools developers is an "estimate of the number of motifs (with the same width and number of occurrences) that would have equal or higher log likelihood ratio if the input sequences had been generated randomly according to the (0-order portion of the) background model" (Bailey and Noble, n.d.).

Each of the novel motifs is compared to a database of known Transcription Factor binding motifs (also filtered at E-value<0.05), full results in Table 3-2 and: MM_vs_PC_supervised_analysis/pan_MM_enhancers_all_sign_motifs.xlsx

De novo motif	E-value between seed and TF motif	TF name (species)	Matching TF consensus motif	Seed enrichment E-value in enhancers
AAAAGAAAAAAAAAAAAAGAAA	5.33E-11	SOC1 (Arabidopsis thaliana)	AAAAAAAAAAAAAAAAAAAAA	1.00E-73
AAAAGAAAAAAAAAAAAAGAAA	0.000813161	IRF4 (Mus musculus)	AAAAAAGAAAATGAAA	1.00E-73
AAAAGAAAAAAAAAAAAAGAAA	0.00127638	CPEB1 (Homo sapiens)	AAAAAAAAAAATAAAAA	1.00E-73
AAAAGAAAAAAAAAAAAAGAAA	0.00183016	BCL6 (Mus musculus)	AGGAGAGAAGGGGAAGGGAAGAAAGGGAAA	1.00E-73
AAAAGAAAAAAAAAAAAAGAAA	0.0115145	STAT1 (Mus musculus)	AAGAAAGAGAAACTGAAAG	1.00E-73

De novo motif	E-value between seed and TF motif	TF name (species)	Matching TF consensus motif	Seed enrichment E-value in enhancers
AAAAGAAAAAAAAAAAAAGAAA	0.0250331	AZF1 (Saccharomyces cerevisiae)	AAAAAGAAA	1.00E-73
GGTCAGGAGTTCGAGACCAGCCTGGCCA	0.0110846	RXRG (Mus musculus)	GAGTTCAAGGTCAGCCT	9.40E-54
GGTCAGGAGTTCGAGACCAGCCTGGCCA	0.0120605	RXRA (Mus musculus)	AGGATCAGGAGTTCAAGGTCAG	9.40E-54
GCCACTAGATGGCAGT	1.40E-08	CTCF (Homo sapiens)	TGGCCACCAGGGGGCGCTA	1.90E-53
GCCACTAGATGGCAGT	3.60E-08	CTCFL (Homo sapiens)	TGGCCACCAGGGGGCGCTA	1.90E-53
TGTGTGTGTGTGTG	0.00093051	DAF-12 (Caenorhabditis elegans)	GTGTGTGTGTGCGTG	0.0057
AAATGAAA	0.0130044	IRF8 (Homo sapiens)	GAGAAAGTGAAACTG	6.20E-07
AAATAAAT	0.0077805	AHCTF1 (Mus musculus)	AAATAAAT	7.80E-05
ATTTGCAT	0.0226467	POU2F1 (Homo sapiens)	ATTTGCATA	0.0017
ATTTGCAT	0.0403733	POU2F2 (Homo sapiens)	TTCATTTGCATAT	0.0017

Table 3-2: Significant motifs and associated TFs for the MM enhancers near OEMM genes.

De novo motif: The pattern significantly enriched in the MM enhancers near OEMM genes regions.

E-value between seed and TF motif: E-value between the pattern significantly enriched and the TF motif from the database.

Matching TF consensus motif: The TF motif from the database.matching the pattern significantly enriched in the MM enhancers near OEMM genes regions.

Seed enrichment E-value in enhancers: E-value for the pattern significantly enriched in the MM enhancers near OEMM genes regions.

Many of the enriched motifs in the enhancer regions are of low complexity, but some are not. One such motif matches to CTCF (Figure 3-11). CTCF is expressed in both PC and MM (at similar levels). This, in conjunction with the fact that these regions are chromatin accessible in MM and not in PC (since PC chromatin accessible peaks have been removed) could point to a possible chromatin mechanism where the CTCF sites in the enhancer regions are becoming more accessible through interactions with transcription activators, as it is the case with GATA-1 and CTCF regions in Erythroid cells (Kang et al., 2017). This increased binding of CTCF to DNA

through chromatin accessibility in MM might create new CTCF boundaries, for example through insulation (S. Kim et al., 2015) placing enhancers in a position where they are regulating new genes compared with PC. Mutations can also influence the binding affinity of CTCF in these enhancers and disrupt enhancer – promoter interactions (Katainen et al., 2015).

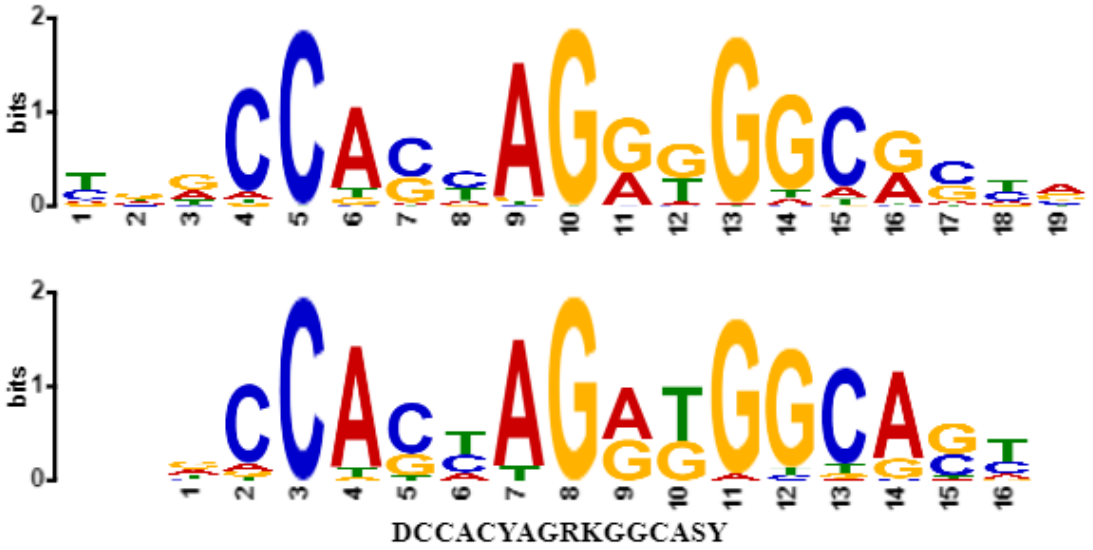


Figure 3-11: CTCF motif.

Top: the consensus logo for the TF CTCF. Bottom: the consensus motif found to be enriched in the MMPC enhancers.

The low complexity *de novo* motif "AAAAGAAAAAAAAAAAAAGAAA" matches to many TF binding sequences. One of them is IRF4: an essential gene in the development of PCs and their immune function (Klein et al., 2006; Mittrücker et al., 2017). It is also essential for MM, IRF4 repression can be toxic for cell lines inducing downregulation of more than 300 genes, forming a positive feedback loop with MYC expression and deregulating genes involved in cell cycle regulation, metabolism and energy or cell death regulation among others (Shaffer et al., 2008). For these reasons, IRF4 is considered a therapeutic target for MM (Agnarelli et al., 2018). The results show that there is an enrichment for candidate enhancer regions containing two motifs which each match the *Mus musculus* IRF4 motif, which is very similar to *Homo sapiens* IRF8 and IRF4 motifs as can be verified in the HOCOMOCO database (Kulakovskiy et al., 2016). Additionally, as it is expected IRF4 is very expressed both in MM and PC (in the top 1% of genes) with nearly double expression in MM compared with PC.

CPEB1 is another TF found to have the previously mentioned associated motif significantly enriched. CPEB1 is expressed in both the healthy and cancer condition but not OE in MM and it is sometimes methylated and down regulated in MM (Heller et al., 2008). AHCTF1 is involved in mitotic and cell cycle processes, very highly expressed both in MM and PC and is down regulated in response to treatment (Hernández-García et al., 2017). POU2F1 and POU2F2 are highly expressed in PC and MM, POU2F2 has a two-fold expression in MM vs. PC and it has been shown to be an important expressed gene in maintaining PC identity (Nagy et al., 2002; Radomska et al., 1994).

This shows that the MM enhancer regions thought to regulate OEMM genes are enriched for motifs associated with TFs that have previously been well established as important in the PC and MM literature. This reinforces the idea that these regions might be genuine enhancers and also active.

3.3.12. Protein coding promoter accessibility and gene expression in MM and PC

Enhancer chromatin accessibility and state has been studied in detail in this thesis, in the context of gene transcription. Transcription initiates at promoters and it is important to determine how changes in promoter accessibility affect expression of genes, inducing transcription or repression.

The great majority of protein coding genes are accessible at the promoter, both in MM and PC, with more genes lying in open chromatin in MM compared with PC, likely pointing at a general opening of chromatin in MM (see Table 3-3, Materials and Methods, “Relationship between protein coding promoter accessibility and gene expression in MM and PC”). Highly expressed genes were separated from less expressed genes using a gene Transcripts Per Million (TPM) average per condition (MM or PC) threshold of 5 (Table 3-4). In both MM and PC around one third of the genes are highly expressed. There is a large difference (close to 50% of all assayed promoters) between the number of promoters being accessible and those highly expressing genes point at promoter open chromatin not being sufficient for high gene expression, perhaps other mechanisms such as acetylation are involved. This effect occurs both in MM and PC.

	TSS in open chromatin (% TSS open/total)	TSS in closed chromatin (% TSS closed/total)
MM	16,301 (82%)	3,656 (18%)
PC	14,512 (73%)	5,445 (27%)

Table 3-3: Promoter accessibility of protein coding genes in MM and PC.

Percentage of protein coding genes overlapping a MM or PC consensus peak from the total protein coding genes quantified and annotated.

	Lowly expressed genes (% lowly expressed/total)	Highly expressed genes (% highly expressed/total)
MM	13,020 (65%)	6,937 (35%)
PC	14,004 (70%)	5,953 (30%)

Table 3-4: Promoter expression of protein coding genes in MM and PC.

Percentage of protein coding genes with high expression (TPM of at least 5) and low expression from the total protein coding genes quantified and annotated.

As can be seen in Table 3-5 and Figure 3-12, the proportion of low expressed genes in open chromatin is very similar in MM than in PC and being in both cases the great majority of low expressed genes. This suggests open chromatin at the TSS not being sufficient, as a generic mechanism, for gene expression in general. Additionally, most of the highly expressed genes in either condition have their promoter in accessible chromatin for both MM and PC (Table 3-5 and Figure 3-12) which can mean that open chromatin is a necessary component for most of the genes to achieve gene expression in both conditions. There are, however, a number of highly expressed genes which are not in open chromatin, which could point at compensatory mechanisms occurring in the transcription of these genes, or the fact that the method I used is not 100% sensitive at detecting open chromatin at promoters. Another possibility is that alternative promoters are used for the expression of these genes.

From all the genes in open chromatin, about two thirds are lowly expressed in both conditions, with roughly equal proportions for both conditions. This means that for a high proportion of genes, open chromatin by itself does not result in higher gene expression: additional mechanisms must be required. This suggests that open chromatin is not sufficient for high-expression without indicating whether its necessary. From all the genes in closed chromatin, about one sixth of genes are highly expressed in both conditions with no difference among the conditions. This means that for a proportion of genes, closed chromatin by itself does not result in gene repression, requiring supplementary mechanisms.

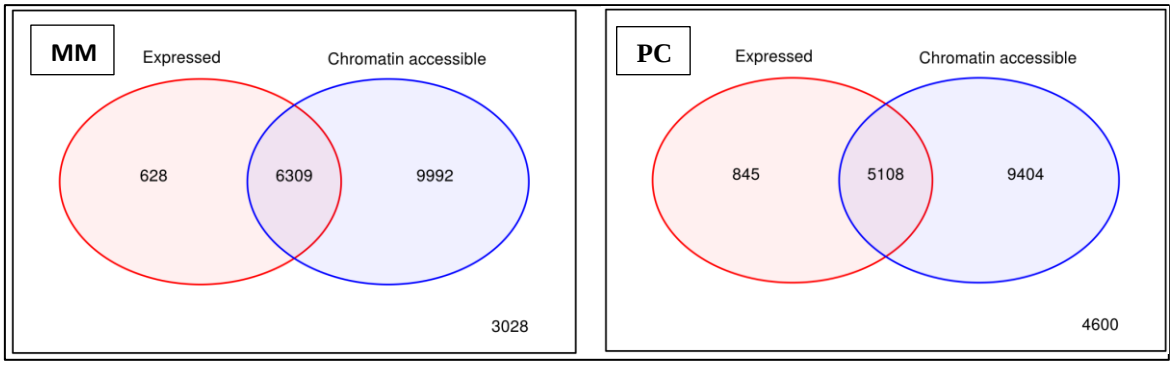


Figure 3-12: Gene counts for each category.

Expressed genes (red) are defined as having an average of 5 TPM or more in each condition. Chromatin accessible promoters (blue) overlapping condition specific consensus peaks.

	MM	PC
Low expressed genes in open chromatin (% Low expressed genes in open chromatin/Total low expressed genes)	9,992 (77%)	9,404 (67%)
Low expressed genes in closed chromatin (% Low expressed genes in closed chromatin/Total low expressed genes)	3,028 (23%)	4,600 (33%)
Highly expressed genes in open chromatin (% Highly expressed genes in open chromatin/Total high expressed genes)	6,309 (91%)	5,108 (86%)
Highly expressed genes in closed chromatin (% Highly expressed genes in closed chromatin/Total high expressed genes)	628 (9%)	845 (14%)
Open chromatin low expressed genes (Open chromatin low expressed genes/Total genes open)	9,992 (61%)	9,404 (65%)
Open chromatin highly expressed genes (Open chromatin highly expressed genes/Total genes open)	6,309 (39%)	5,108 (35%)
Closed chromatin low expressed genes (Closed chromatin low expressed genes/Total genes open)	3028 (83%)	4600 (84%)
Closed chromatin highly expressed genes (Closed chromatin highly expressed genes/Total genes open)	628 (17%)	845 (16%)

Table 3-5: Categories of protein coding promoters based on chromatin accessibility and gene expression.

Percentages for different categories. Total protein coding genes quantified and annotated: 19,957.

Criteria: Low expressed genes are protein coding genes with TPM of less than 5, while high expressed genes have TPM of more than 5. Open chromatin genes for a condition are protein coding genes where the promoter is overlapping a condition consensus peak, while closed chromatin are non-overlapping promoters.

Importantly, chromatin accessibility of the 264 protein coding MM genes which are up regulated with respect to PC are studied in detail (Figure 3-13). A third of these genes (86) have their promoters in a non-accessible chromatin PC state; most of these are also closed in MM (48 closed vs. 38 accessible in MM). It is possible that these OEMM genes that have the promoters in the non-accessible chromatin state are not expressed: with the expression going from very low in PC to low in MM.

As it is expected, for the remaining two thirds (178) which have their promoters accessible in PC, the majority (173 promoters) are accessible, while 5 are not accessible in the cancerous state. Surprisingly, since most of the OEMM genes have the promoter already in open chromatin in the PC state, an extra mechanism is required for these genes to be OE in MM either in the promoter (for example recruitment of a TF) or elsewhere (for example activation of a nearby enhancer) or both.

In terms of protein coding genes, around 20% (53/264) of the OEMM genes are in non-accessible chromatin in MM and 33% (86/264) have non-open chromatin in PC compared with 18% of all protein coding genes with non-accessible chromatin in MM and 27% in PC (Figure 3-14 A). As can be seen in Figure 3-14 A, the differences between the proportion of genes with low accessibility at the TSS from all OEMM and from any gene are non-significant for PC and MM. This indicates that whether a gene is OE in the cancerous condition does not affect its possibility of being chromatin accessible compared with any random gene. It is possible that there is no clear relationship between a promoter's accessibility and its overexpression in different conditions, another possibility could be that this exposes the inherent sensitivity of the ATAC-seq assay: whether all chromatin accessible regions are recalled. The great majority of OEMM with accessible promoter in MM also have an accessible promoter in PC, suggesting that a chromatin accessible promoter might be necessary but not sufficient for gene expression or overexpression, something which has already been suggested before (Klemm et al., 2019).

Finally, there are 1.2% (178/14,512) MM up regulated genes from all accessible promoters in PC compared to 1.5% (86/5,445) for non-accessible promoters (Figure 3-14 B). 1.3% (264/19,957) genes are generally up regulated in MM and the differences in this proportions are not significant (Figure 3-14 B). This indicates that having open chromatin in the PC state, does not make a gene more likely to become up regulated in the cancerous state, actually, a slightly higher proportion of genes in non-accessible chromatin in PC become up regulated in MM. The percentages, are also consistent with the general chance of any gene (whether

having open or closed TSS) to be up regulated in MM. Therefore, overexpression of a gene in the malignant state does not seem dependent on the healthy state open or closed chromatin.

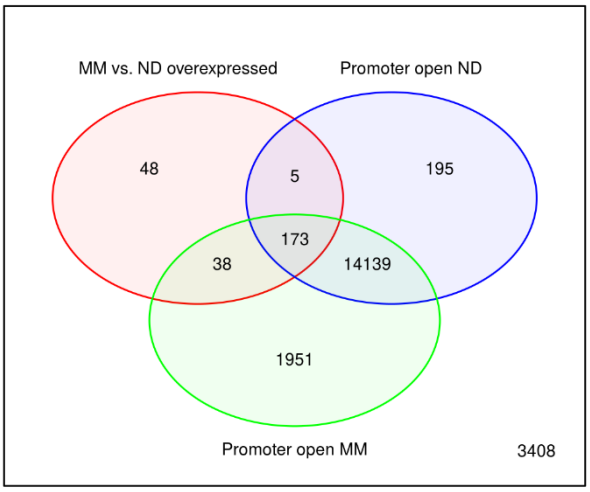


Figure 3-13: MM OE promoter state of protein coding genes in detail. (ND: PC).

OEMM genes in red, chromatin accessible promoters overlapping condition specific consensus peaks: PC peaks (blue), MM peaks (green).

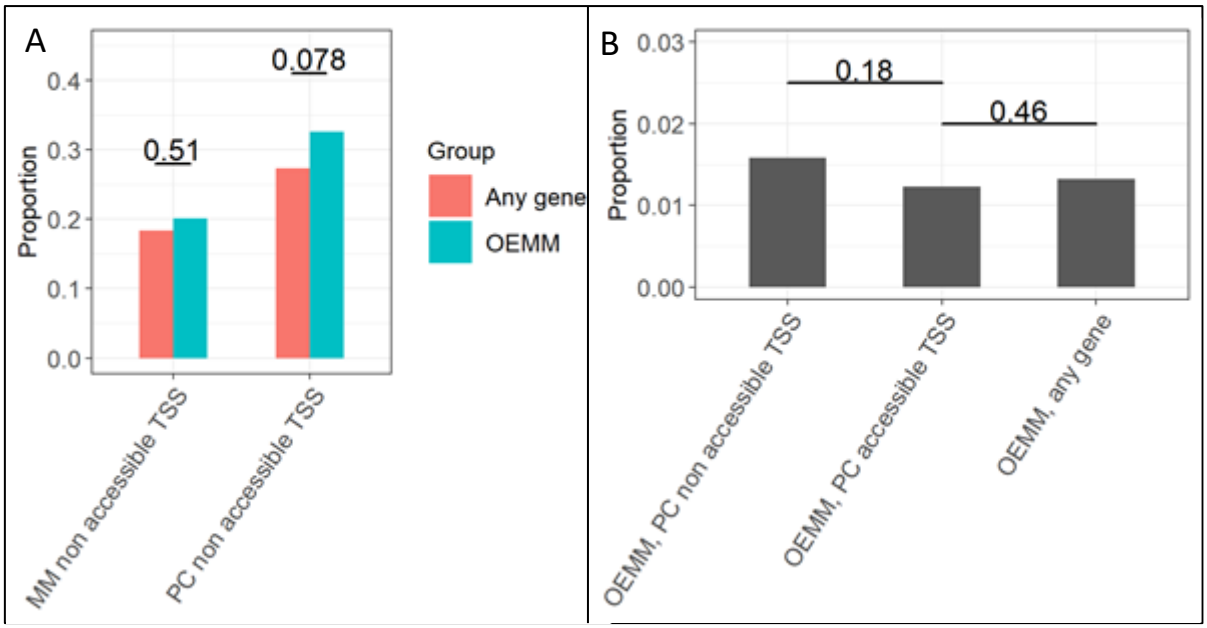


Figure 3-14: OEMM genes and TSS accessibility effects. (ND: PC).

A: Proportion of MM non-accessible TSS from all protein coding genes (pink) and proportion of MM non-accessible TSS from all OEMM protein coding genes (cyan) marked by the group of bars with label "MM non accessible TSS". Proportion of PC non-accessible TSS from all protein coding genes (pink) and proportion of PC non-accessible TSS from all OEMM protein coding genes (cyan) marked by the group of bars with label "PC non accessible TSS".

B: Left to right categories: proportion of OEMM genes from all non-accessible protein coding TSS in PC, proportion of OEMM genes from all accessible protein coding TSS in PC, proportion of OEMM genes from all protein coding TSS.

Pairwise Pearson's chi-squared test using Benjamini & Hochberg (1995) correction p-adjusted values between the groups shown above.

3.4. Discussion

In this chapter I have delineated the regulatory network of 28 MM primary samples compared with 5 PC samples (coming from 3 individual patients) from bone marrow. To my knowledge, only one previous study used *in vivo* primary MM and PC samples to analyze interactions between enhancers and promoters. This was done by obtaining recurrent mutated regions and associated genes with altered expression and linking them through B-cell promoter Capture Hi-C data (Hoang et al., 2018). Additionally, *in vitro* differentiated PCs were used to establish enhancers based on histone acetylation (Jin et al., 2018). Y. Jin et al. discovered around 20,000 regions with differential histone acetylation between PC and MM, considered enhancers. Unfortunately, to my knowledge, no comprehensive list of these enhancer regions was published to be compared with my results. In the study presented here 18,339 regions were considered to be DA, but importantly only 9,527 of these regions were considered MMPC enhancers after removing TSS (except for e-RNAs). Y. Jin et al. employed H3K27Ac to establish the candidate enhancer sets and ATAC-seq data was also used to obtain correlations with H3K27Ac signals. Despite the fact that they were correlated, the correlation is far from perfect and it is reasonable to assume that the *in vitro* differentiated cells do not fully resemble the PCs in this study. Furthermore, considering also a previous study (Kleftogiannis et al., 2015) showing that enhancers sets obtained by different methods (and from the associated data types) yield “significantly different enhancer predictions” it is very possible that the candidate regions from Y. Jin et al. obtained from acetylation data will differ from chromatin accessible data derived enhancers.

The study by Hoang and colleagues proposed 114 recurrent mutated regions intervening with 271 altered genes in MM (through B-cell interactions). When considering these mutated regions as possible enhancers, only 9 MMPC enhancers overlapping were found. These are: chr9:37,197,100-37,199,238, chr9:37,390,761-37,391,506, chr9:37,300,722-37,303,075, chr9:37,370,895-37,372,317, chr1:178,531,785-178,532,684, chr9:37,408,907-37,409,738, chr15:59,543,131-59,545,052, chr6:142,706,264-142,706,564 and chr14:35,883,745-35,883,958. The latter two being the only ones overlapping the 1,462 MM enhancers

elucidated in this thesis, hence, pointing at novel MM regulatory candidates provided to be tested.

806 DEMM and 548 OEMM genes were also found to be deregulated in the MM transition in this work. From these, the genes contained in the mentioned interactions (Hoang et al., 2018) overlap only: MIR3153, CKS2, SNORA63, CYP2R1 and COBLL1 (the latter only in the DEMM genes). Once again, pointing at different mechanisms: mutations or enhancer activation targeting different pathways of gene deregulation.

Y. Jin et al. also performed gene ontology and pathway analysis using these genes associated with the candidate activated MM vs. PC enhancers (Jin et al., 2018). In this study, MM and PC DEMM as well as OEMM genes were surveyed and functions regarding PCs and MM have been found. To my knowledge there is no comprehensive list of DE genes between MM and PC published by Y. Jin et al. but examples of genes that are very highly expressed in MM compared with PC (recapitulating previously found evidence by Y. Jin et al.) are NCAM1, MLLT3 or CDH2. Examples of enriched functions in DEMM genes include activation of immune response and response to bacterium. OEMM genes are enriched for novel pathways such as negative regulation of extrinsic apoptotic signalling pathway or negative regulation of megakaryocyte differentiation. Some pathways corresponding to OEMM genes are also found enriched in genes associated with regions having heterochromatin state in B-Lymphocytes but higher accessibility in MM compared to PC (Jin et al., 2018). Examples include: positive regulation of cell migration (with positive regulation of blood vessel endothelial cell migration being found by Y. Jin et al.), within the neuron differentiation category, neuron recognition subcategory found by Y. Jin et al., angiogenesis (sprouting angiogenesis found by Y. Jin et al.) or homophilic cell adhesion via plasma membrane adhesion molecules (homophilic cell adhesion found by Y. Jin et al.). Other specific categories are found to be enriched in the OEMM genes regarding BMPs which have bone formation inducing properties among other roles (Katagiri and Watabe, 2016). BMP2 is known to have apoptosis-independent effects affecting proliferation and cell cycle arrest in MM (Lagler et al., 2017), while BMP9 also has anti-myeloma activity (Olsen et al., 2014). Another category enriched is cell migration, which in MM is known to be induced by various pathways including SH3GL3 overexpression in CD138-negative myeloma cells (Chen et al., 2016), SH3GL3 is not significantly expressed/OE in MM so it seems like another mechanism might be involved. This pathway could be perhaps through CD40 activation (Tai et al., 2003) or high GJA1/CX43 expression promoted by SRC3 expressed in bone marrow stromal cells (Jin et al., 2017). Furthermore, ossification is also an enriched pathway. Osteolysis (the reverse of ossification) can be caused by MM through osteoclasts

differentiation and activity (Mansour et al., 2017) and maybe through other ways, at the same time osteoblasts can slow MM growth (Reagan et al., 2015). It seems that osteoblasts, osteoclasts and MM are closely related.

Applying the knowledge that 90% of enhancer - promoter interactions occur within the 1Mb range with a median linear distance between the elements of 331Kb (Javierre et al., 2016) and in favour of covering a greater proportion of the interactions, an extended threshold of 1Mb has been considered to increase the sensitivity. To account for a higher Type I error rate, only interactions having the same direction ATAC-seq and RNA-seq signals changes have been considered. The result is a list of 1,959 MMPC enhancers generating a set of 2,698 pairs of MMPC enhancers regulating DEMM genes, which were further filtered to 311 MM only enhancers up regulating genes. To my knowledge, Jin *et al.* didn't publish a list of candidate interactions regarding enhancers and target genes, from the 12 examples given regarding MM-related genes, only the gene HGF is recapitulated in my study. This gene has a high significance in MM and signalling with other cells in the bone marrow, promoting angiogenesis and osteoclastogenesis (Ullah, 2019). An example candidate enhancer possibly activing this gene in the cancer state is shown with over accessible MM chromatin and relevant Hi-C interactions in B-cells (of which PCs are a subset). Furthermore, other possible interactions are shown for other DEMM genes such as BMP4 and ABCA1. If the TAD boundaries are maintained, these interactions could reveal possible enabling mechanisms.

The B-cell interactions on the basis of recurrent mutations in candidate enhancers being associated with target gene expression changes proposed by Hoang *et al.* resulted in only 6 interactions proposed as enriched in 1% - 6% of the MM cases considered. These involved chr11:14,579,387-14,583,849 with CALCB, chr2:165,615,060-165,624,028 with COBLL1, chr17:46,094,139-46,103,073 and HOXB3, chr3:186,739,608-186,745,052 and ST6GAL1, chr9:37,375,172-37,395,282 with PAX5 and chr1:16,944,603-16,958,779 with ATP13A2. Additionally, two MM subgroup associated interactions were suggested: chr3:187,635,970-187,636,359 related with TPRG1 in a very low MM samples percentage (2% of the HD and 3% of MYC translocated subtype samples) and chr3:186,739,608-186,745,052 was found related to the ST6GAL1 gene in 4/109 (4%) of MYC translocated samples (Hoang et al., 2018). Despite not overlapping, the present study has a general MM vs. PC differential potential interaction with the COBLL1 gene less than 4Kb downstream of the reported one: chr2:165,626,541-165,627,945.

Jin *et al.* studied the chromatin state of enhancer regions determined for MM in GM12878 (normal B-cell line). The study found a high enrichment of MM accessibility in the heterochromatin state (determined by ChromHMM using histone modifications) of GM12878 (B cell healthy cell line) combined with a moderate correlation of MM H3K27Ac with accesibility at heterochromatin state regions of GM12878 (Jin et al., 2018). Together this suggests that there may be a significant proportion of inactive enhancers in B-cells becoming active in MM and perhaps regulating target genes. In this chapter, I established the predicted chromatin state of the putative MM and PC enhancer regions in 173 other cell types available in both healthy and cancer states. As can be seen in Figure 3-7, the putative enhancers labelled "MM specific" comprise around one third of all the MMPC enhancers thought to regulate DEMM protein coding genes and are mainly assigned "low signal" state in B-cells while mainly having an active enhancer state in MM and other cell types (such as endothelial) for some regions. Despite there being heterochromatin and repressed states, it is possible that the low signal state represents inactivation in most of these regions and cell types except MM, perhaps recapitulating the findings by Y. Jin et al. Another study profiling histone modifications and gene expression at 16 developmental stages of hematopoietic commitment reflected that 90% of the 48,415 hematopoietic candidate enhancers altered their state during hematopoiesis and were similar within a major lineage: myeloid, lymphoid, and erythroid, with enhancer activation occurring early in the lineage and repression more progressively (Lara-Astiaso et al., 2014). The fact that different cell types are clustering together based on chromatin state (enhancer state) is also seen in the subset of all hematopoietic candidate enhancers presented in Figure 3-6 and Figure 3-7 for disease and healthy cell types respectively. Within this cell type clustering, some enhancers in various hematopoietic cell types, such as macrophage or neutrophils within the myeloid lineage or T-cells and precursor B-cells, have various regulatory programs comprising of activation of enhancers that are also active in malignant differentiated PCs. In this regard, neutrophils are found to have activated a set of enhancers also active in tonsil PC and MM samples.

Activated MM enhancers regulating OEMM genes have also been found to be enriched in binding motifs for proteins that have previously been associated with myeloma activation (superenhancer associated TFs by Jin *et al.*) such as the IRF family, POU2F2 or CTCF (Jin et al., 2018) and novel motifs such as CPEB1, BCL6, STAT1 or AHCTF1, POU2F1 and CTCFL. Interestingly, BCL6 which is key for B-cell germinal center formation and differentiation into PCs (Fukuda et al., 1997) has bounded sites with affinity to STAT TFs which are also found to be enriched at the MM enhancers (Dent et al., 1997). Moreover, BRD4, a TF found to bind at

enhancers in MM CLs (Fulciniti et al., 2018), was not found in my study, however, it was not found either in the study by Jin and colleagues done on primary samples. Perhaps it is more detectable in the MM CLs. Together these results increase confidence in the enhancer regions proposed and suggest mechanisms of enhancer activation which can be targeted. Another important control for the proposed regions in this analysis is the fact that there is a clear enrichment in the number of MM enhancers found within 1Mb of OEMM genes compared with any gene. This is likely in a scenario in which the regions are genuinely regulating the OEMM genes.

Importantly, since a differentiation has been made in the candidate regions to retain only enhancers and not promoters, the relationship between chromatin accessibility in terms of the promoter and gene expression is tested. General promoter accessibility is a required characteristic of expressed genes but not sufficient, perhaps requiring further mechanisms, a finding recapitulated (Klemm et al., 2019). Additionally, it is also found that an accessible promoter state in the PC does not increase the chances of high gene expression of that gene in the MM state. The relationship between promoter accessibility and gene expression has been previously studied. One study comparing breast cancer cell lines treated with BMP4 and vehicle control determined that in general chromatin openness at the TSS correlated with its expression but with the caveat that transcription variation was not fully explained by it, with different TF binding partially or increased intron accessibility (perhaps enhancers) explaining it (Ampuja et al., 2017). Coherently with this hypothesis, a study profiling chromatin accessibility and gene expression during hematopoiesis, showed enrichment of accessibility at “active promoters” (Lara-Astiaso et al., 2014). Moreover, another paper involving progenitor cells from human cord blood found that assigning the top 2,000 most variable accessibility peaks to the nearest TSS reproduced transcriptomic-like clusters reflecting the different blood lineages, thereby establishing that promoter accessibility and expression might be linked but also found that accessibility was coherent with gene expression only in some lineage transitions (Zheng et al., 2018). In an analysis in human cortical neurogenesis, highly significant correlations were found between TSS accessibility and expression in neurodevelopment and highly DE genes between the two main human neocortex regions, but less so at genome-wide protein-coding genes and lncRNA (de la Torre-Ubieta et al., 2018). Apart from different TF binding and enhancers other factors proposed to influence gene expression are DNA methylation, histone modifications and miRNA (summarized in de la Torre-Ubieta et al., 2018).

Finally, a recent paper, concluded that genes with a significant correlation between promoter accessibility and gene expression in the mid-gestation placenta (both variables being high)

were likely related to housekeeping functions while high expression but medium – low accessibility marked tissue-specific genes, replicating these findings in other mouse cell types such as alpha, beta, embryonic stem or hematopoietic (Starks et al., 2019). Furthermore, genes with a medium – low expression and high accessibility at the promoter were found to be actively suppressed in general, suggesting that repressed differentiation programs (such as the neuronal in the mid-gestation placenta cells) occur through this mechanism, mediated at least in part, by repressive histone modifications at such gene promoters (Starks et al., 2019).

Together, these results extend and complement the current knowledge of MM enhancer biology, activation TFs required and enhancer evolution through different hematopoietic cell types, allowing to test the findings and advancing the therapeutical aspect.

4. Chapter 4: MM subgroups

4.1. Acronyms and Abbreviations used in the chapter

Acronym	Definition
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
BM	Bone Marrow
bp	Base pair
BP	Biological Process
ChIP-seq	Chromatin Immunoprecipitation sequencing
CL	Cell Line
CRISPRi	CRISPR interference
DA	Differentially Accessible
DAMM	Differentially Accessible MM
DASMM	Differentially Accessible Subgroup MM
DE	Differentially expressed
DESMM	Differentially Expressed Subgroup MM
eRNA	Enhancer RNA
FDR	False Discovery Rate
GO	Gene Ontology
H3ac	Acetylation of histone H3
H3K27ac	Acetylation of histone H3 lysine 27
HD	Hyperdiploid
IgH	Immunoglobulin Heavy Chain
Kb	Kilobase
KEGG	Kyoto Encyclopedia of Genes and Genomes
log_2foldchange	Log (base 2) fold change
LRT	Log Ratio Test
Mb	Megabase
MF	Molecular Function
miRNA	Micro RNA
MM	Multiple Myeloma
MMPC	Multiple Myeloma and Plasma Cell
mRNA	Messenger RNA
ND	Normal Donor
OE	Overexpressed
OESMM	Over Expressed Subgroup MM
PC	Plasma Cell (used interchangeably with ND)
rLog	Regularized Log
RNA-seq	RNA sequencing
SMM	Subgroup MM
TF	Transcription Factor
TPM	Transcripts per Million
TSS	Transcription Start Site
UTR	Untranslated Region

4.2. Introduction

As mentioned in the Introduction, sub-classification of cancer within a particular originating tissue cell type is very important for the patient from multiple points of view, the most important ones: diagnosis, prognosis and treatment strategies. MM tends to involve presence of chromosomal translocations in the IgH locus, nearly in all cases derailing the expression of either CCND1, FGFR3 and NSD2 (used interchangeably with MMSET in this work in correspondence with oncologists), MAF, MAFB, or CCND3. Hiperdiploidy (HD) or amplification/deletion of chromosome arms (Perrot et al., 2019) are also driver events accounting for a large proportion of the cases (Bergsagel and Kuehl, 2005) and are used to sub classify MM into different subgroups.

Characterization of MM subtypes has been previously performed first using cytogenetic information to elucidate the primary oncogenic drivers and then using the available data in a supervised way as explained in section 1.7.2. Initially MM subgroups were determined by unsupervised methods on gene expression microarray data, leading to two basic classification systems: translocation/cyclin D (TC) and University of Arkansas for Medical Science (UAMS) classification. As its name implies, the TC classification is based on IgH translocation partners and cyclin D expression. Subgroups are determined using absolute expression of the 4 typical translocation partners NSD2/MMSET or FGFR3, high MAF or MAFB, high CCND3 and CCND1 with manual expression boundaries for assigning samples to groups (Bergsagel et al., 2005). Additionally, 4 subgroups with relative enrichment of expression in MM compared to bone marrow PCs: D1 only (CCND1), D2 only (CCND2), D1 and D2 and finally, none of D1 or D2 (Bergsagel et al., 2005) for 8 subgroups in total.

The UAMS classification used a cohort of 414 newly diagnosed MM patients creating 7 groups, 4 of them corresponding to the cytogenetic classification in this chapter: MMSET [t(4;14)], MAF [t(14;16)] (also including MAFB [t(14;20)]), CCND1 [t(11;14)] and HD/HY, but extending it to CCND3 [t(6;14)], proliferation associated genes and low percentage bone disease (Zhan et al., 2006).

The UAMS classification was later extended by a study involving 320 newly diagnosed MM patients that found a small percentage of bone disease samples clustering with the MAF/MAFB translocated cluster and 4 additional groups: one with a myeloid signature, another with high expression in the nuclear factor kappa-light-chain-enhancer of activated B cells (NF-κB) genes, a new group with overexpression of cancer testis antigens without overexpression of proliferation genes and finally one with upregulation of the protein tyrosine phosphatase

PTP4A3, the receptor PTPRZ1 and the suppressor of cytokine signaling SOCS3 (Broyl et al., 2010). Furthermore, gene expression data has been used to infer cytogenetic abnormalities with some success (Zhou et al., 2012).

Other methods have been used for sub classification of MM samples, such as genomic features like DNA mutations (Kuijjer et al., 2018), epigenetic such as DNA methylation (Kumar Mishra and Guda, 2017) or through transcriptomic data as it is the case of gene expression panels (Golub et al., 1999). Some studies have studied the general MM enhancers (Hoang et al., 2018; Jin et al., 2018) or their activator proteins (Fulciniti et al., 2018), however, the MM enhancers and regulating genes in the context of MM subgroups remains unknown.

As mentioned in section 3.3.1, the samples contain cytogenetic information regarding the most common driver initiation events in MM: translocations of the IgH locus with MMSET/NSD2, MAF and CCND1 genes and the Hyperdiploidy (HD) status.

4.2.1. Chapter Aims

One of the aims of this chapter is to identify the chromatin and gene expression landscape, including the activated enhancers in each subgroup (Figure 2-4). As well, to detect to what extent these features are shared between or are exclusive to each subtype. Moreover, the determined chromatin/enhancer and gene expression activation profiles are validated to decide its usefulness in assigning subtype to the available samples (including non-cytogenetically annotated).

Additionally, mechanistic ways in which the enhancer – promoter candidate interactions might be driving the cancer phenotype in a given subtype are examined. In particular the TF network regulating the set of subtype specific enhancers is determined. Furthermore, it is also observed whether the activity of TFs correlates with altered gene expression or chromatin accessibility. Finally, the gene-associated pathways leading to oncogenesis for each MM subgroup are uncovered.

4.3. Results

4.3.1. Subgroup specific quality control statistics

A similar criteria for quality control of primary PC and MM samples and MM cell lines has been used as in chapter 3. Starting with the 38 paired ATAC – RNA samples used in chapter 3, only samples with cytogenetic information were selected resulting in 33 paired ATAC – RNA were used (28 primary samples and 5 MM CL samples). The 33 samples consisted of: 5 primary PC

samples comprising of 3 donors (with $CD19^+$ and $CD19^-$ variant samples for two of them and a technical replicate for one of these) and 23 cytogenetically annotated MM samples comprising of 13 Hyperdiploid, 4 containing t(11;14) IGH/CCND1, 2 containing t(14;16) IGH/MAF, 4 containing t(4;14) MMSET/IGH and 5 MM cell lines (2 t(14;16) IGH/MAF, 2 t(11;14) IGH/CCND1 and 1 t(4;14) MMSET), see Table 2-1. From the 38 samples used in chapter 3, non-cytogenetically annotated samples were excluded in the subgroup quality control statistics since it can't be determined whether they may include known translocations and belong to an existing subgroup, but were subject to the sample quality threshold where used. For the 33 samples, a total of 2,695,129,775 RNA-seq read pairs were generated with an average mapping rate of 82% when quantifying reads in transcripts, generating an average of 67,170,007 mapped reads per sample. 2,803,428,347 ATAC-seq read pairs were sequenced and 1,963,890,986 total unique read single ends (average 59,511,848 per sample) were input into the peak caller (MM subgroup, PC and MM cell line sample single ends entering peak calling can be seen in Figure 4-1 A) to generate a total of 2,015,593 and 2,291,890 sample narrow and broad peaks respectively. The average sample assigned fraction is 20%. Assigned fractions for each sample grouped by subgroup can be seen in Figure 4-1 B and the number of peaks per sample per subgroup in Figure 4-1 C. Additionally, the subgroup sample RNA-seq mapped pairs in Figure 4-1 D. The table with all the samples and details can be seen at:

MM_vs_PC_supervised_analysis/ATAC_and_RNA-seq_stats.xlsx

In general, it can be seen that the median sample number of peaks generated (Figure 4-1 C) is highest in the subgroups with the highest median single ends per sample (Figure 4-1 A). The clear exceptions to this are the MAF subgroup and the MM CL with around 37M and 17M sample median single ends and 75K and 37K sample median peaks respectively. This is to be expected in samples with exceptional proportion of reads in accessible regions, which is the case for both (Figure 4-1 B): sample median 24% for MAF and 20% for the MM CLs. As it was explored in Chapter 3, the number of single ends entering the peak calling is not correlated with the assigned fraction (not shown). Additionally, higher number of single ends produce greater number of called peaks (not shown). Also, as Figure 4-1 E shows, the number of called chromatin accessible peaks is only weakly but significantly correlated with the assigned fraction.

The RNA-seq mapped pairs (Figure 4-1 D) tended to be above 60M for all primary sample subgroups. As is the case with ATAC-seq, cell line samples with significantly less number mapped reads were allowed in the analysis.

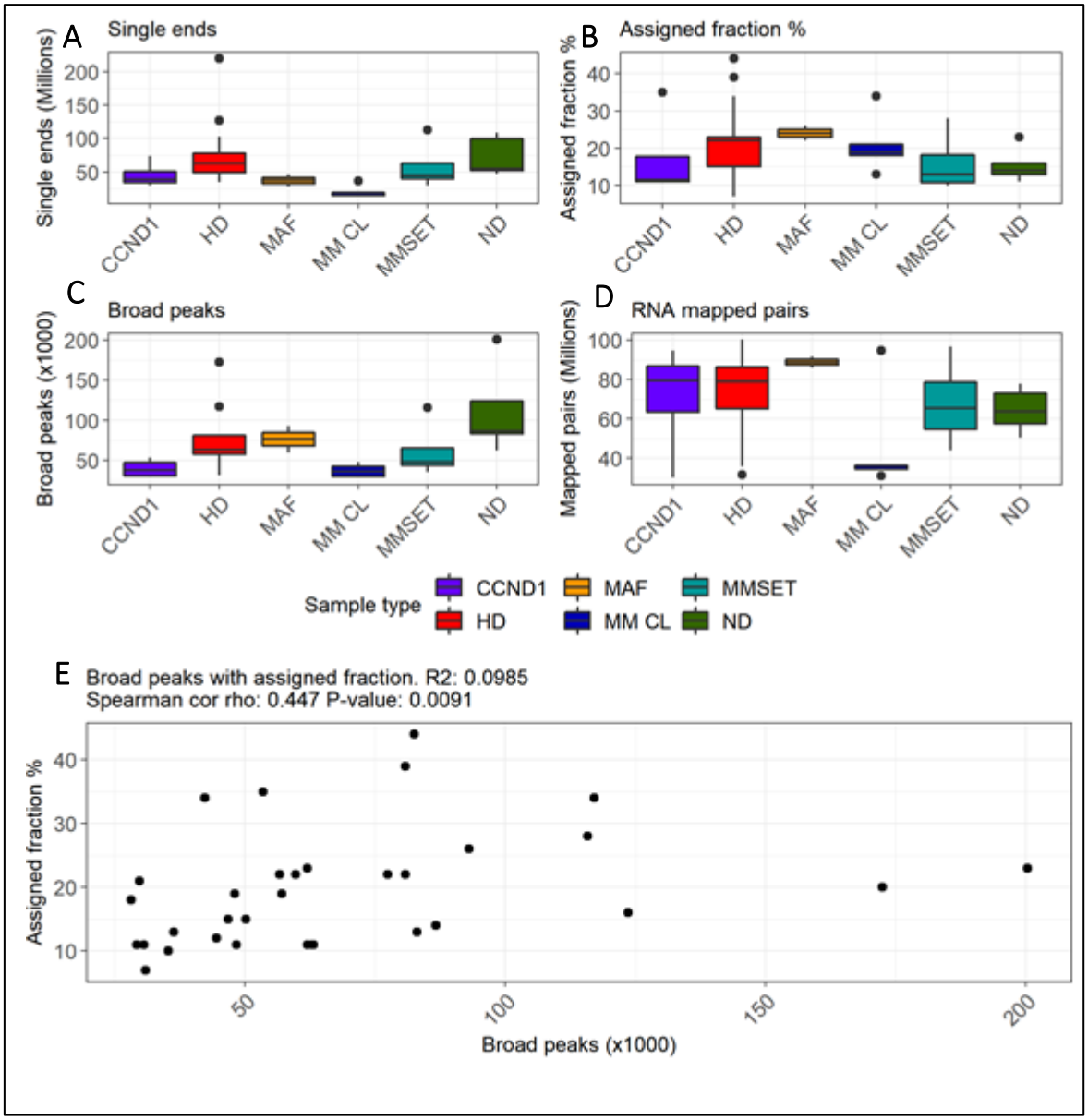


Figure 4-1: Quality control metrics for the different subgroups

A-D: Distribution of ATAC-seq statistics for different groups of all samples used in the study including all cytogenetic annotated MM subgroups for primary samples based on primary initiating genetic event (CCND1, HD, MAF, MMSET) as well as PC (ND) and MM CL (Table 2-1). A) Number of filtered reads used for peak calling. B) Assigned fraction. C) Broad peaks called by MACS2, after filtering. D) RNA-seq mapped read pairs. E) Relationship between the assigned fraction and the broad peaks called by MACS2, after filtering.

4.3.2. Consensus peaks for PC and MM subgroups

Since the purpose of the analysis is to obtain MM subgroup differences, it is important to consider all subgroup specific features. The number of samples in every subgroup is unbalanced, therefore, similarly to how consensus peaks for primary PC and MM samples were

produced, first consensus peak sets were obtained for the different subgroups and then merged (as described in the Materials and Methods).

As was the case with the PC and MM samples (see Chapter 3), the number of down sampled filtered shifted single ends per sample used was 28,231,242. The number of balanced consensus peaks found for each subgroup can be seen in Table 4-1, in general, it can be seen that the called subgroup consensus peaks may be entering saturation as the number of samples increases in the HD group. The PC group, despite having 5 samples still has a very high peaks per sample ratio and can likely benefit from the incorporation of more samples and higher sequencing. PC chromatin accessibility and RNA-seq expression were observed in terms of CD19 status and donor id (not shown) with the highest variance being due to donor id status, it is possible that these two covariates were producing the heterogeneity in the PC samples, leading to the high peak to sample ratio. Despite having 13 samples, the HD group is still creating a considerable peak set in terms of peaks per sample, this can be due to the high variability between these samples due to different copies of chromosomal arms.

Subgroups	Number of samples	Subgroup peaks (balanced)	Peaks/sample
MAF	2	104,903	52,452
HD	13	224,475	17,267
MMSET	4	97,631	24,408
CCND1	4	73,899	18,475
PC	5	188,261	37,652

Table 4-1: Consensus peaks per sample for each subgroup.

Subgroups: MM subgroups based on cytogenetics and PC. Subgroup peaks (balanced): PC and MM subgroup consensus peaks produced at equal sequencing depth per sample.

All the subgroup balanced consensus peaks were merged and a joint set of 295,238 consensus peaks were found, referred to as subgroup MM and PC consensus peaks. These are areas of high chromatin accessibility in at least one of the samples considered using the same sample sequencing depth.

From the 295,238 total regions, 45,322 regions are only accessible in PC samples (Figure 4-2 A white area, bottom right). As can be seen in Figure 4-2 A, there are 44,296 regions accessible in

all MM subgroups and 5,062 of them not intersecting with PC (Figure 4-2 B). There are therefore 39,234 (44,296 – 5,062) regions common to all MM subgroups and PC regions.

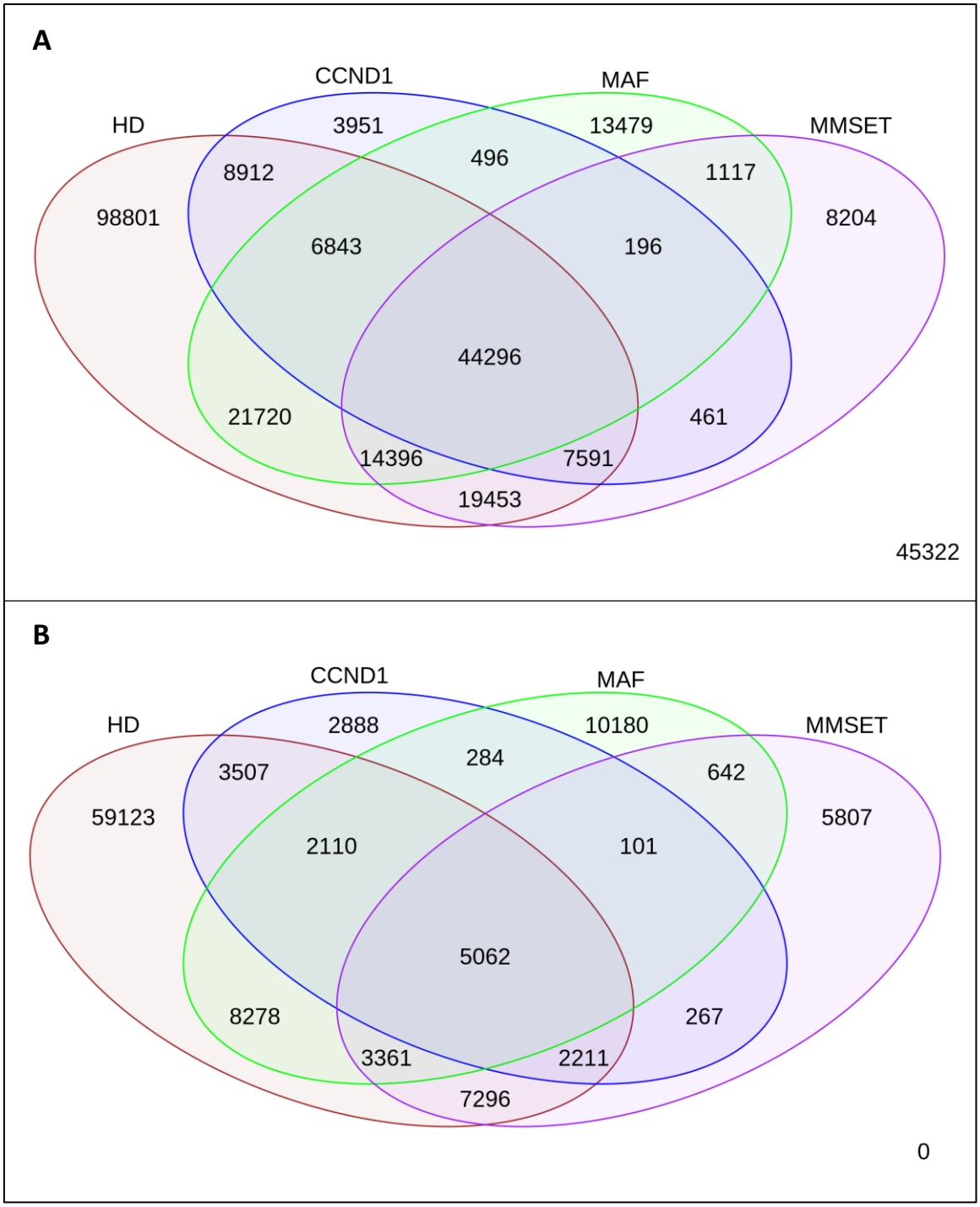


Figure 4-2: Consensus peak regions overlap for each subgroup.

Consensus peak regions for each MM subgroup: HD (red), CCND1 (blue), MAF (green), MMSET (purple) and PC (in white area, bottom right) overlapping the consensus peaks for PC and MM subgroups.

A: Chromatin accessible regions for all subgroups including PC regions. B: Excluding PC.

111,117 peaks remained after removing PC chromatin accessible regions (Figure 4-2 B). The largest set of subgroup specific chromatin accessible areas occurs in HD, with more than half of the total (59,123), and MAF (10,180). A very high proportion of MM only regions are common to both of them: 8,278, meaning that one quarter of MAF accessible chromatin is also accessible in HD. HD and MMSET share 7,296 MM only peaks, a third of MMSET peaks are also found in HD. Since the HD subgroup has many samples and more reads can pile to the same regions due to Hyperdiploidy, this can contribute to creating a larger set of consensus peaks. It is therefore to be expected that the chromatin accessibility profile for HD will partially recapitulate other subgroups. Furthermore, 5,062 MM only chromatin accessible peaks are common to all MM subgroups.

The annotations of the consensus peak regions and the different subgroup specific subsets were studied (see Materials and Methods): the distributions are generally similar to each other and to a random set of genomic regions (Figure 4-3 and Figure 4-4). Most of the regions are introns (between 37% and 45% of the total), followed by intergenic which is highly underrepresented in consensus peaks for PC and MM subgroups and the subset of peaks for all subgroups (18% to 28%) compared with randomly generated genomic regions representative per chromosome (Figure 4-3 "RANDOM_ALL" category in cyan and Figure 4-4 D with 40%).

There is an enrichment of open chromatin in PC and MM subgroups in the proportion of promoters (9-16%) compared to random regions (5%); coding sequences (Figure 4-3 labelled "cds" in green and Figure 4-4 C): 10-13% vs. 5%. Also, the 5' UTR, non-coding parts of the mRNA involved in translational regulation, is enriched for accessible chromatin (Figure 4-3 labelled "5UTRs" in dark yellow and Figure 4-4 B) vs. random background: 5-12% vs. 2%. These enrichments can be explained because promoters can extend from 1kbp upstream of the TSS to 1kb downstream, therefore a portion of the 5'UTR (and perhaps the CDS) annotated sites might in fact be promoters. Also, unannotated TSS involved in MM and PC might overlap exons (CDS) on the annotation, so these CDS might in fact be acting as promoters of expressed genes. As it was seen in Chapter 3, around half of the DAMM regions were annotated or unannotated TSS and promoter accessibility tends to be necessary for gene expression. Finally, the 3' UTR ratio is even throughout the different MM subgroups, PC and random background (3-4%), this is marked in Figure 4-3 labelled "3UTRs" in light red and Figure 4-4 A.

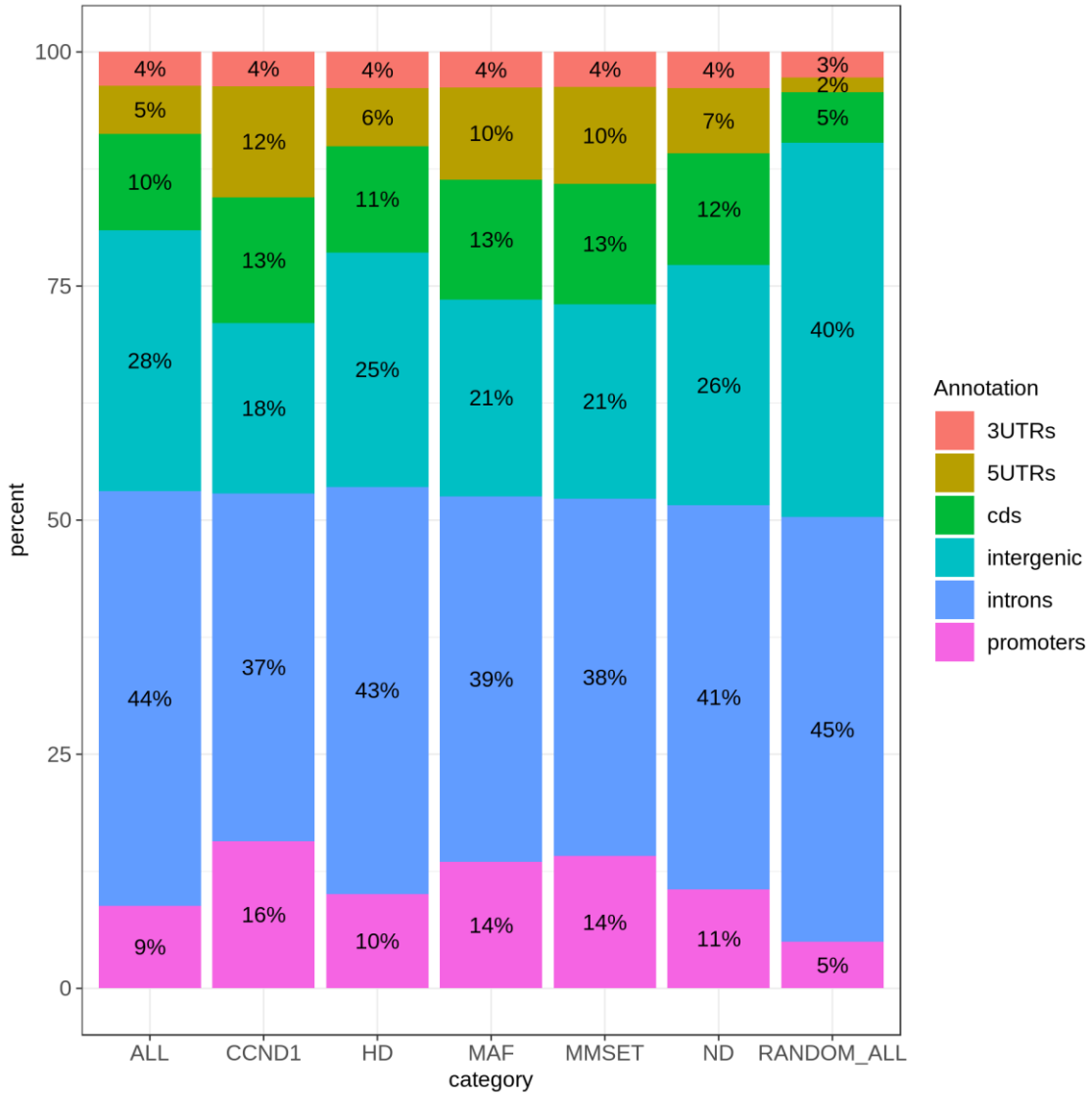


Figure 4-3: Annotation of consensus peak regions. (ND: PC)

Genomic annotation of the consensus peak regions, one region can overlap multiple genomic categories on both strands but each genomic category was counted only once per region. ALL: All consensus peak regions. MM subgroup regions: CCND1, HD, MAF, MMSET and PC. RANDOM_ALL: A random generation of sequences simulating a sample equal to all consensus peak regions per chromosome. UTR: Untranslated Region. 3UTRs: 3 prime end UTR, 5UTRs: 5 prime end UTR. CDS: coding sequence.

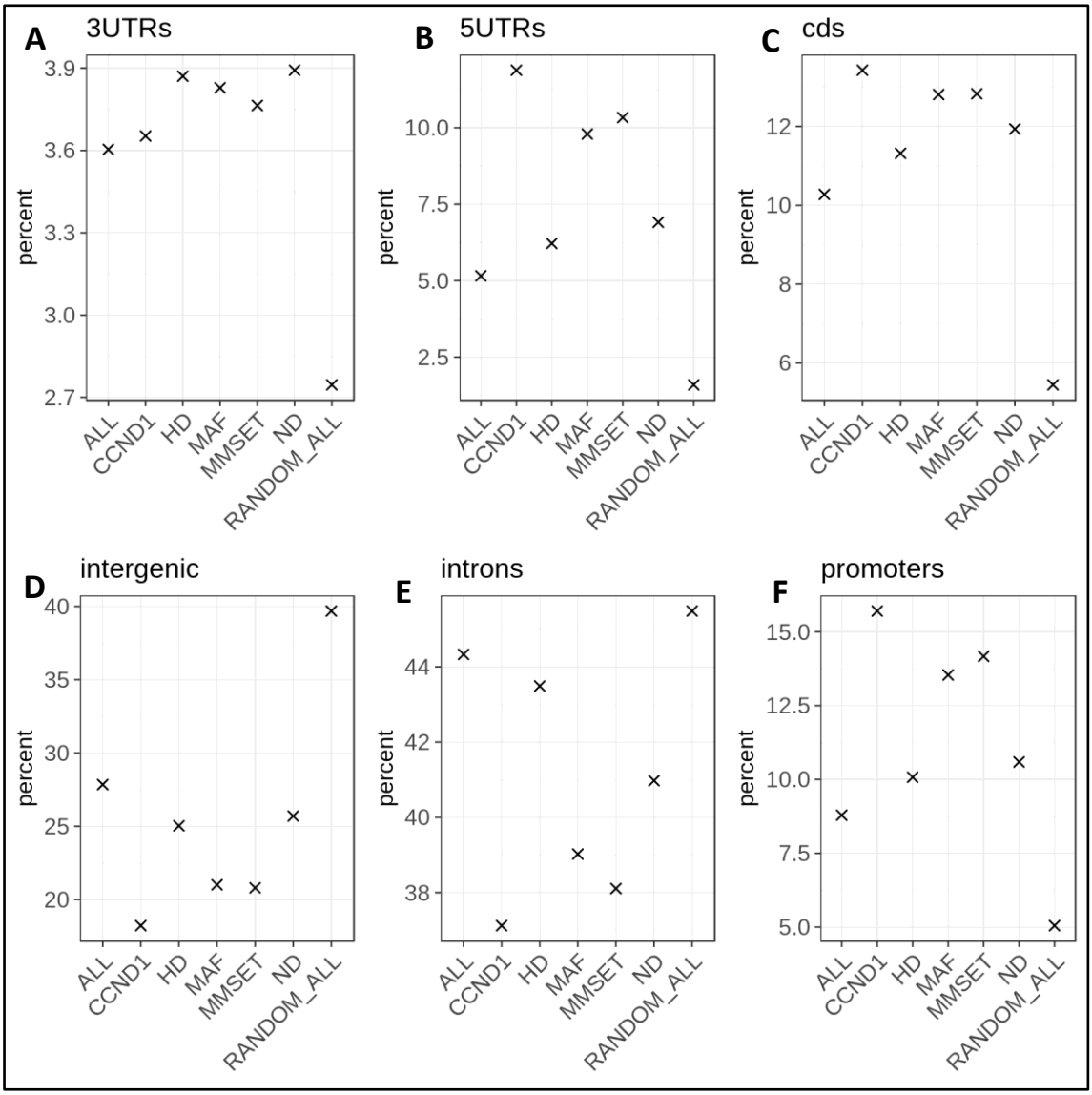


Figure 4-4: Ratio scatterplots showing annotation of consensus peaks for PC and MM subgroups. (ND: PC)

Proportion of consensus peak regions corresponding to each genomic annotation: A: 3 prime end UTR, B: 5 prime end UTR, C: coding sequence, D: intergenic, E: introns, F: promoters. One region can overlap multiple genomic categories on both strands but each genomic category was counted only once per region. The groups on the x-axis are: All consensus peak regions (ALL). Regions overlapping different subgroups: CCND1, HD, MAF, MMSET and PC ("ND" label). A random generation of sequences simulating a sample equal to all consensus peak regions per chromosome (RANDOM_ALL).

The chromatin accessibility profiles were obtained as specified in the Materials and Methods section and can be seen in Figure 4-5. In general the cancer state is characterized by general opening of chromatin (enrichment of regions in Figure 4-5 in the first column above the 0 log_2Foldchange for each subgroup). CCND1 and HD samples have a large proportion of regions opening up even more than already open regions in PC (regions in the top right quadrant in the

first column in Figure 4-5). MMSET has an enrichment of regions becoming accessible which are in inactive chromatin in PC (regions in the top left quadrant in the first column in Figure 4-5). MAF seems to have more even distribution in this regard. Finally, the subgroup samples are very correlated in terms of the change in fold accessibility compared to normal. This phenomenon however, can be a result of spurious correlation of ratios where even if two MM subgroups are not correlated with each other, they may be individually correlated with a third.

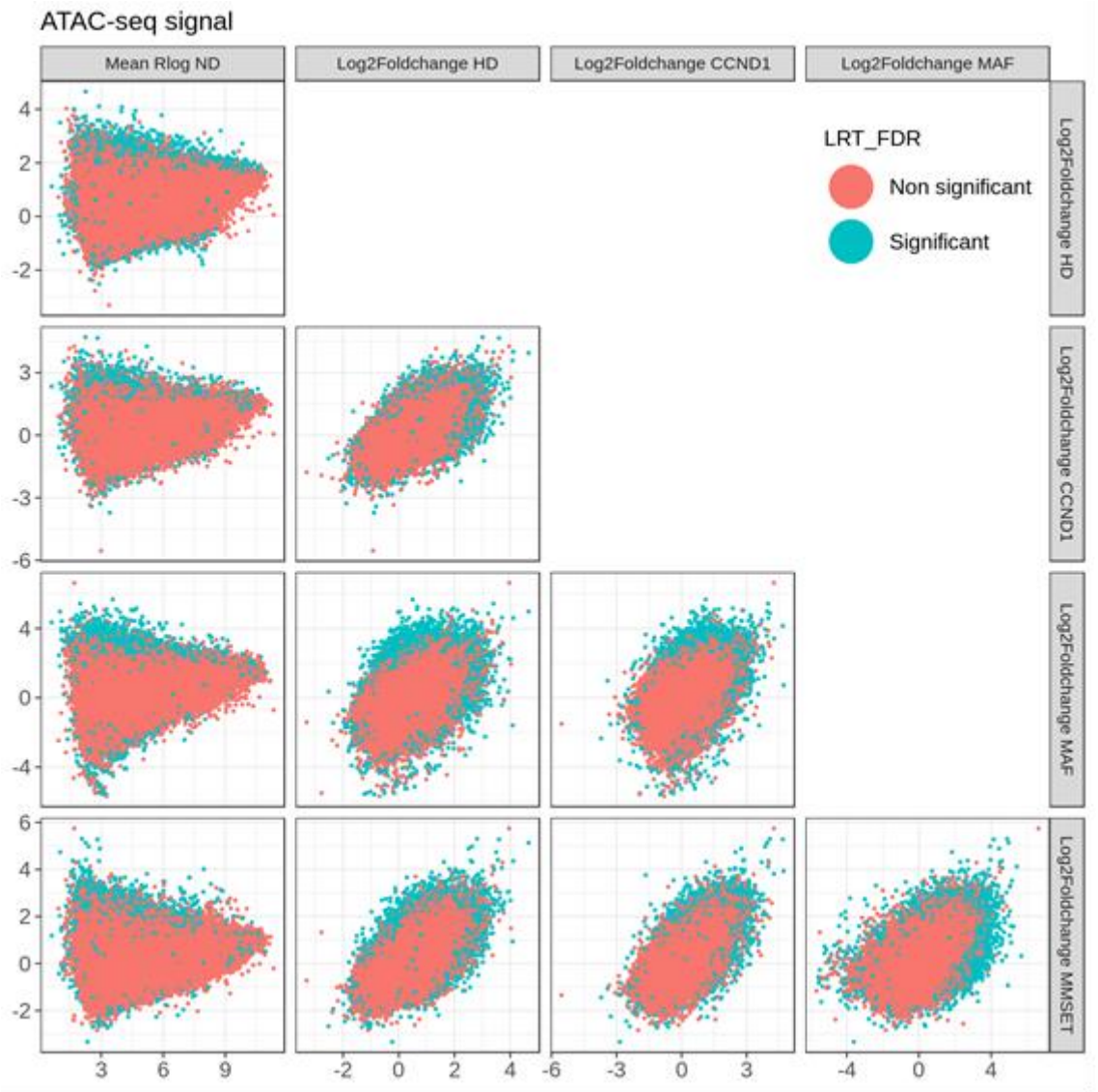


Figure 4-5: Subgroup chromatin accessibility profiles. (ND: PC)

Details for the subgroup MM and PC consensus peak regions. The first column shows the average Rlog (normalized) chromatin accessibility for the PC samples. The rest of the columns and rows show the Log_2fold change in chromatin accessibility signal between the samples of each subgroup specified and PC samples. Distinction is made between signals with significant Log Ratio Test (LRT) where the effect of the subgroup accounting for batch is significant.

4.3.3. DASMM regions

Using the consensus peaks for PC and MM subgroups as features, regions with significantly different chromatin accessibility between any MM subgroup and PC state were determined: a proxy for enhancers in each condition. The process followed is detailed in the Materials and Methods chapter. Outlier detection by DESeq2 (based on Cooks distance for subgroups having 3 or more samples) was disabled for this analysis. This is because the HD group samples are now isolated in a condition and since they are very heterogeneous, due to the fact that different sets of chromatin accessible regions can be found in multiple copies in different samples, they can genuinely have extreme signal at certain regions.

As was the case with the MM vs. PC analysis, several covariates convolute the analysis of these regions. As mentioned previously (and can be seen in Table 2-1), the 5 PC samples involved corresponded to 3 donors with different CD19 status for two of the donors and one of them having an additional replicate which was merged with its corresponding sample. Additionally, sample batches exist. A model testing for MM subgroup vs. PC accounting for batch was performed using an LRT which fitted the data using a "reduced model" with only batch and compared it to a "full model" including batch and condition effects (where each subgroup was a "condition"). If the condition effect was decided to be significant enough at explaining the data, a region's chromatin accessibility variation among different conditions was assumed to be influenced by it. As it was done in Chapter 3 for the analogous analysis, samples belonging to singleton batches were assigned to the same batch. Since the MM samples were subdivided into distinct subgroups, subgroups containing few samples can't be compared with PC while taking into account the PC donor or CD19 status because of confounding variables. Individual tests comparing each subgroup to PC average patient id and PC average CD19 status were not performed because the variables are confounding.

The results from the test performed showed there are 8,911 statistically significant DA (FDR < 0.05 and absolute $\log_2$FoldChange greater or equal to 1) consensus peak regions (DASMM regions). They can be seen in:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_sign_DE_ATAC_regions.tsv.gz

An analysis of the different types of genomic regions distributions can be seen in Figure 4-6 and Figure 4-7. Qualitatively, the results are very similar to the consensus peak region distributions (Figure 4-3 and Figure 4-4). Introns make up between 42-46% of the distributions (Figure 4-6 blue and Figure 4-7 E). DASMM regions lying in intergenic parts of the genome being particularly downrepresented in all MM subgroups and the complete set of DASMM

regions (Figure 4-6 cyan “intergenic label” and Figure 4-7 D) compared with its representative counterpart random set of regions (Figure 4-6 and Figure 4-7 D “RANDOM_ALL” label). Also, similarly as with the consensus peak regions, promoters (Figure 4-6 in pink and Figure 4-7 F), coding sequences (Figure 4-6 labelled “cds” in green and Figure 4-7 C) and 5’ UTRs (Figure 4-6 labelled “5UTR” in dark yellow and Figure 4-7 B) are also qualitatively overrepresented in all DASMM regions and MM subgroup specific ones compared with random sets of regions. This is to be expected because the DASMM regions are a subset of the consensus peak regions including TSS.

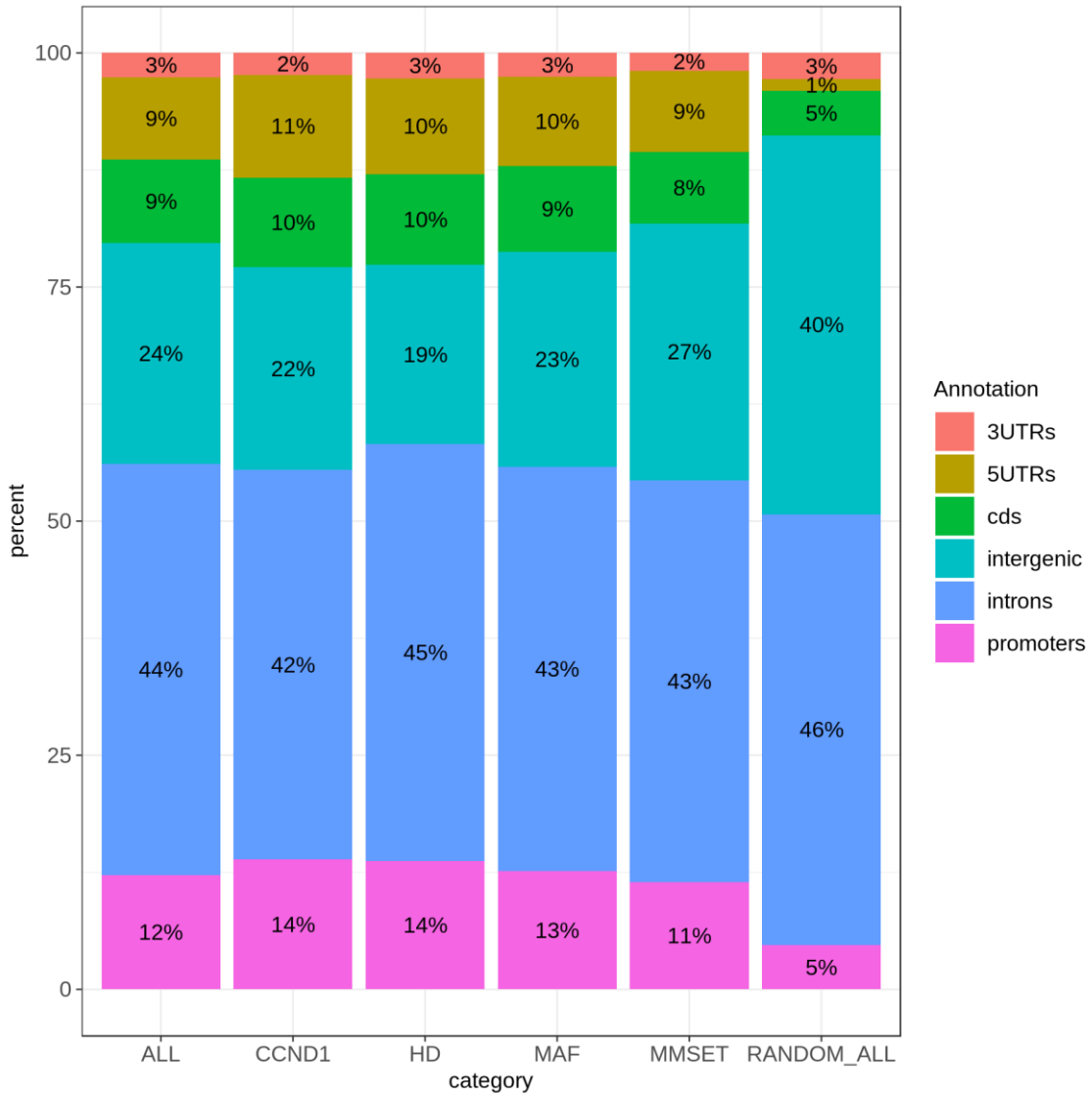


Figure 4-6: Annotation of DASMM regions.

Genomic annotation of the DASMM regions, one region can overlap multiple genomic categories on both strands but each genomic category was counted only once per region. ALL: DASMM regions DA in any MM subgroup vs. PC.

MM subgroup regions (DA in a particular subgroup vs. PC, can also be DA in other subgroups): CCND1, HD, MAF, MMSET. RANDOM_ALL: A random generation of sequences simulating a sample equal to the "ALL" group per chromosome. UTR: Untranslated Region. 3UTRs: 3 prime end UTR, 5UTRs: 5 prime end UTR. CDS: coding sequence.

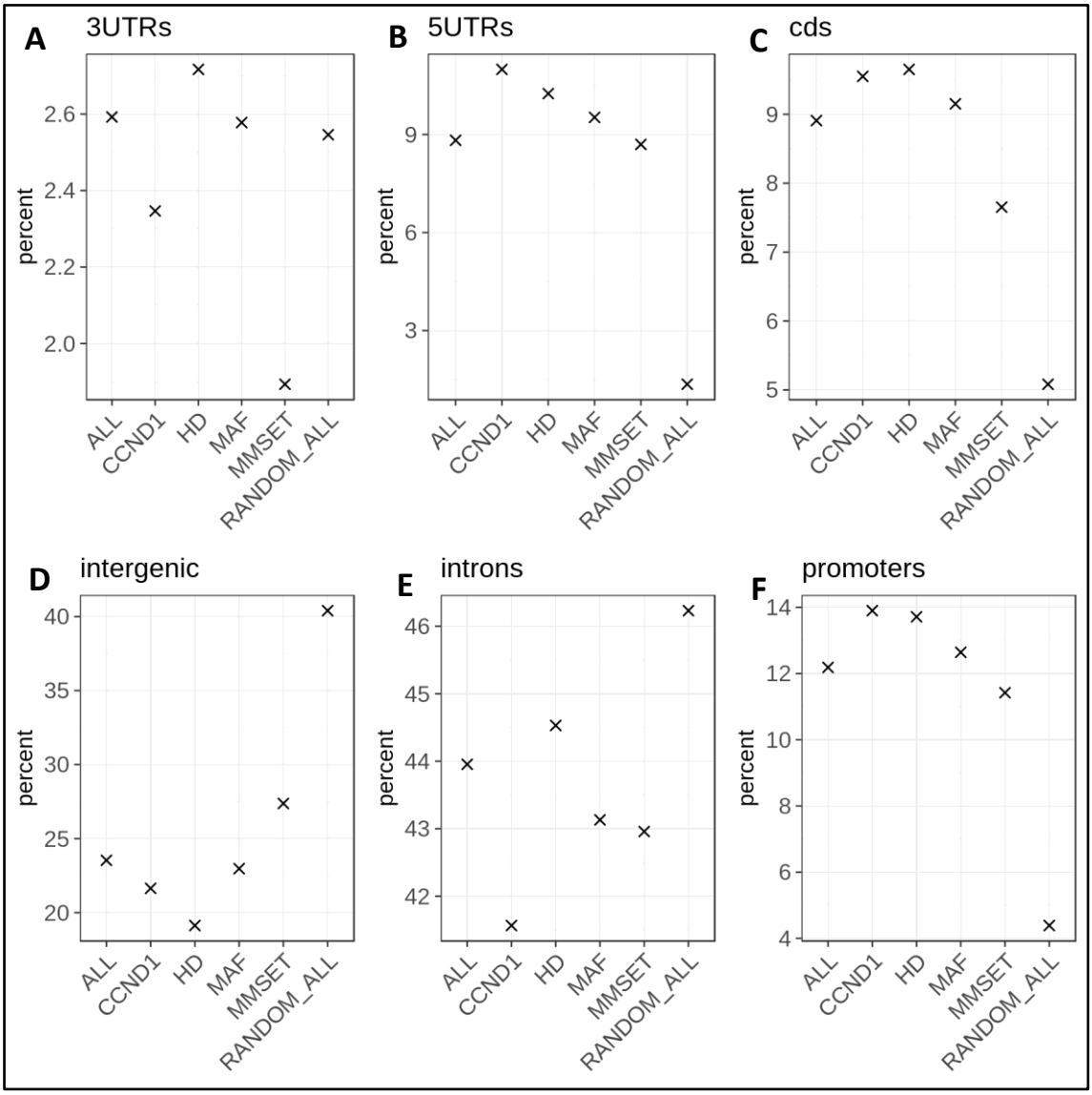


Figure 4-7: Ratio scatterplots showing annotation of DASMM regions. (ND: PC)

Proportion of DASMM regions corresponding to each genomic annotation: A: 3 prime end UTR, B: 5 prime end UTR, C: coding sequence, D: intergenic, E: introns, F: promoters. One region can overlap multiple genomic categories on both strands but each genomic category was counted only once per region. The groups on the x-axis are ALL: DASMM regions DA in any MM subgroup vs. PC. MM subgroup regions; MM subgroup DASMM regions (DA in a particular subgroup vs. PC, can also be DA in other subgroups): CCND1, HD, MAF and MMSET. A random generation of sequences simulating a sample equal to all DASMM regions per chromosome (RANDOM_ALL).

4.3.4. Removing TSS from the DASMM regions

As in the pan MM vs. PC analysis, annotated and unannotated TSS were removed from DASMM regions (see Materials and Methods chapter), while allowing unannotated TSS from novel single exon transcripts, which may be eRNA transcripts (Ding et al., 2018). From the 8,911 starting regions, only 6,897 remained after this step, these regions are referred to as DASMM enhancers. The table with the regions can be found in:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_sign_DE_ATAC_regions_no_TSS.tsv.gz

Figure 4-8 shows how they are distributed among the different subgroups (compared with PC).

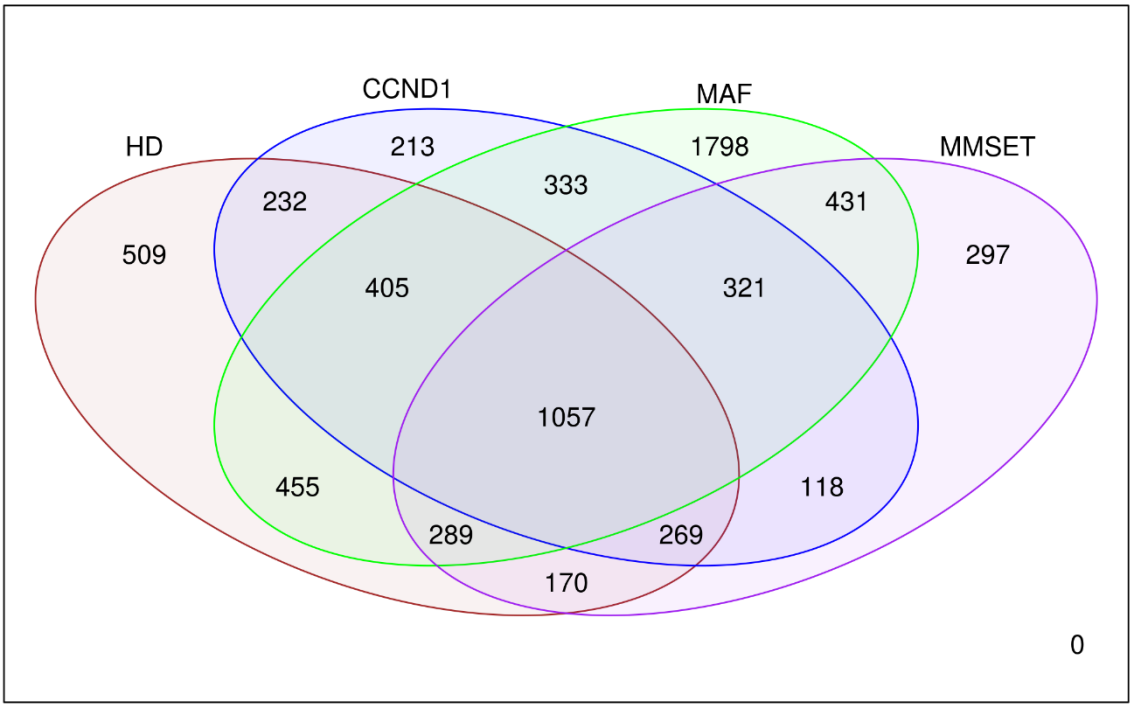


Figure 4-8: DASMM enhancers overlap for each subgroup.

DASMM enhancers for each MM subgroup: HD (red), CCND1 (blue), MAF (green) and MMSET (purple) overlapping the DASMM enhancers.

As can be seen qualitatively, about a quarter (1,798) of all the DASMM enhancers are MAF specific which makes this subgroup the most different from the healthy state in terms of chromatin accessibility. It must be noted that there are only 2 MAF samples, so the effect of one of them having a strong signal could contribute strongly to this effect. Figure 4-2 A, shows that about 5% of all consensus peaks (13,479 out of 295,238) for PC and MM subgroups are MAF and PC specific so there is a 5-fold enrichment when it comes to the proportion of

DASMM enhancers found compared with the proportion of simultaneously accessible areas for MAF and PC. The HD group follows with 509 specific genomic regions with altered chromatin state, while about 15% of all the regions (1,057) are common DASMM enhancers between all MM subgroups and PC.

4.3.5. SMM regions

In order to identify *de novo* enhancer candidates, it is important to determine regions that are more chromatin accessible in the cancer state (for each subgroup) and are not already accessible in the healthy state (see Materials and Methods): the SMM regions. The regions can be seen in:

MM_vs_PC_supervised_analysis\MM_subgroup_vs_PC_ATAC_OE_regions_no_ND_peaks.bed.gz

There are 3,107 regions which are significantly more accessible in at least one MM subgroup than in PC, the distribution of the regions can be seen in Figure 4-9 and Figure 4-10. Around half (50-60%) of the SMM regions (MM subgroup specific SMM regions, any SMM region and random set of regions) lie in introns and between 28-44% are now intergenic (Figure 4-9 blue and cyan label and Figure 4-10 E and D respectively). In general, the intergenic category is now at qualitatively very similar levels as the background randomly generated sequences and not underrepresented compared to random sequences as was the case in consensus peak regions (Figure 4-3 cyan label and Figure 4-4 D) and DASMM regions (Figure 4-6 cyan label and Figure 4-7 D). All the categories of SMM regions have similar profiles (Figure 4-9), except the HD state, having a qualitatively enrichment in the proportion of introns and a decrease in the intergenic category with respect to the rest (Figure 4-9 blue and cyan and Figure 4-10 E and D respectively). Maybe pointing at amplification in the copy number of enhancers as a consequence of Hyperdiploidy generating more accessibility in each enhancer.

There is an enrichment in promoters (pink), coding sequences (labelled “cds” in green) and 5’ UTR regions (labelled “5UTRs” in dark yellow) overlapping subgroup regions compared with randomly sampled regions that was seen in previous cases such as with the consensus peaks (Figure 4-3 labels corresponding to mentioned colours and Figure 4-4 F, C and B respectively). This enrichment has been qualitatively eliminated (Figure 4-9 labels corresponding to mentioned colours and Figure 4-10 F, C and B respectively). Perhaps this is a consequence of chromatin accessible areas for PC samples being removed, particularly promoters (and regions which are adjacent to them and might be annotated as promoters). It is possible that a large proportion of promoters correspond to genes which may be expressed in MM while still being

accessible in the PC state, although this may not result in PC gene expression (which was shown in chapter 3), leading to a set of regions more deterministic to MM gene expression.

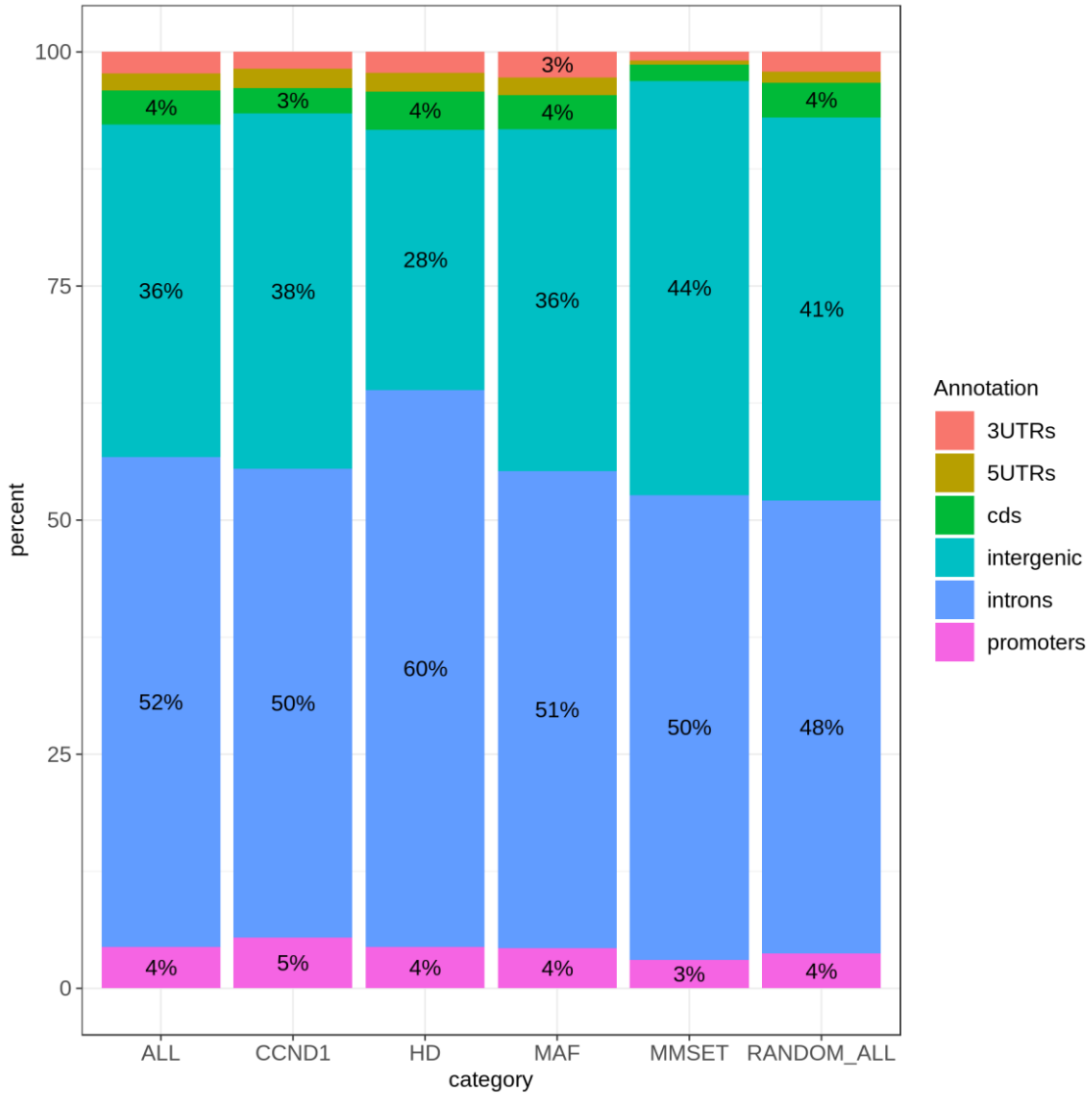


Figure 4-9: Annotation of Distribution of the SMM regions.

Genomic annotation of the SMM regions, one region can overlap multiple genomic categories on both strands but each genomic category was counted only once per region. Labels for percentages of 2% or less are omitted.

ALL: SMM regions over accessible in any MM subgroup vs. PC. MM subgroup SMM regions (over accessible in a particular subgroup vs. PC, can also be over accessible in other subgroups): CCND1, HD, MAF, MMSET. RANDOM_ALL: A random generation of sequences simulating a sample equal to the "ALL" group per chromosome. UTR: Untranslated Region. 3UTRs: 3 prime end UTR, 5UTRs: 5 prime end UTR. CDS: coding sequence.

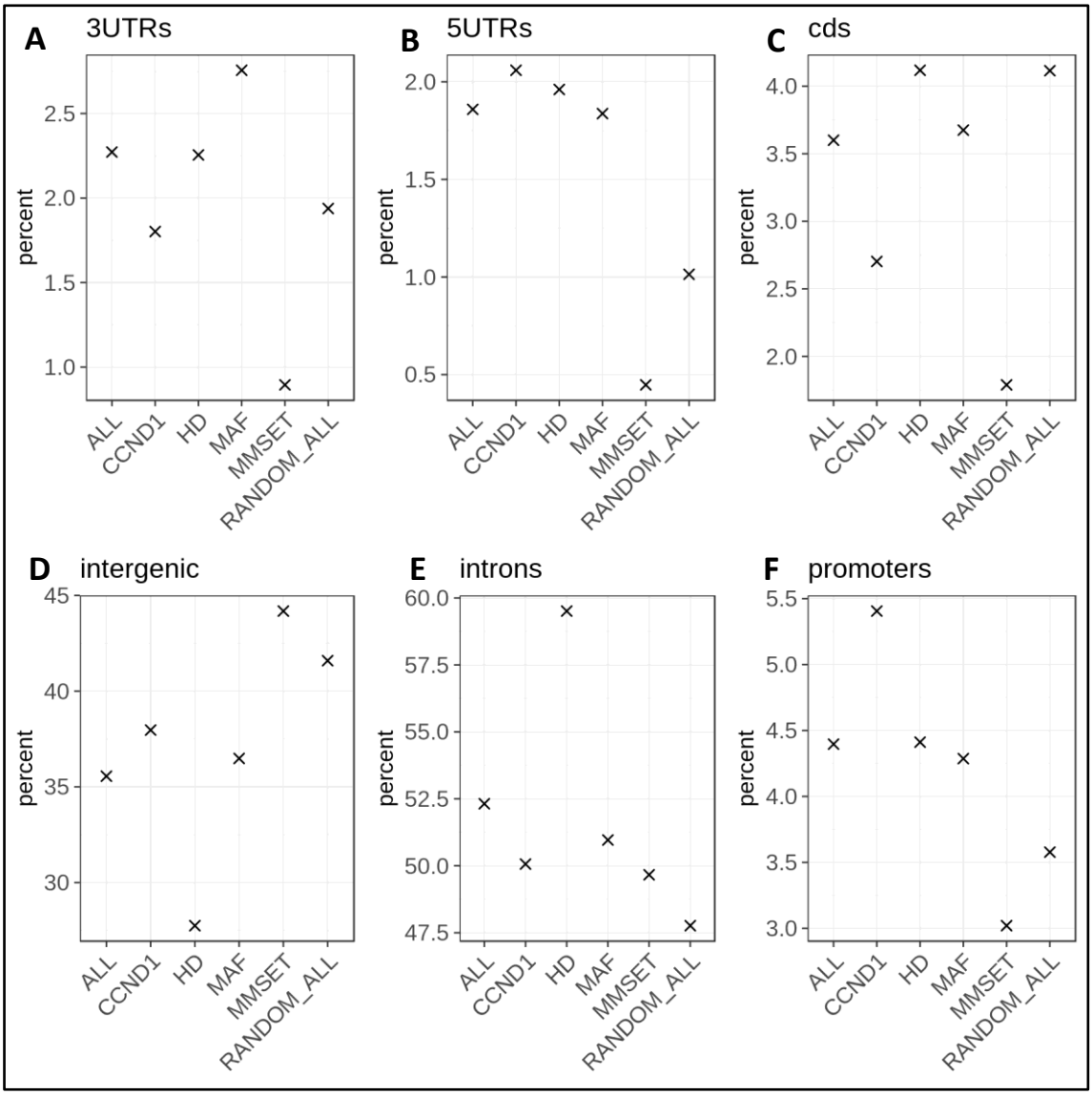


Figure 4-10: Ratio scatterplots showing annotation of SMM regions.

Proportion of SMM regions corresponding to each genomic annotation: A: 3 prime end UTR, B: 5 prime end UTR, C: coding sequence, D: intergenic, E: introns, F: promoters. One region can overlap multiple genomic categories on both strands but each genomic category was counted only once per region. The groups on the x-axis are ALL: SMM regions over accessible in any MM subgroup vs. PC. MM subgroup regions (over accessible in a particular MM subgroup vs. PC, can also be over accessible in other subgroups); CCND1, HD, MAF MMSET. A random generation of sequences simulating a sample equal to all SMM regions per chromosome (RANDOM_ALL).

4.3.6. SMM enhancers

SMM regions will contain both cis-regulator regions and transcription start sites. To identify enhancer candidates, the TSS should be removed. 2,801 regions are considered chromatin accessible exclusively in MM subgroups vs. PC, referred to as SMM enhancers and can be seen in Figure 4-11 and:

MM_vs_PC_supervised_analysis\MM_subgroup_vs_PC_ATAC_OE_regions_no_ND_peaks_no_TSS.bed.gz

All MM subgroups have a statistically significant difference in the proportion of subgroup exclusive regions when transitioning from consensus peaks for PC and MM subgroups (Figure 4-12 pink bars within each subgroup) to DASMM enhancers (Figure 4-12 green) and from DASMM enhancers to SMM enhancers (Figure 4-12 blue). The proportion of MM subgroup exclusive regions increases through these transitions, except in the case of the HD subgroup with a significant decrease in the ratio of HD consensus peaks to HD DASMM enhancers. The relative conversion ratio of SMM to DASMM enhancers is similar for each subgroup (between 63% and 73%) but the overall conversion rate is 41% and 13% for regions common to all subgroups (Table 4-2). While 1,057/6,897 (15%) of DASMM enhancers are DA in all subgroups compared with the normal state, only 134/2,801 (5%) are over accessible in all subgroups compared to the PC. This points at a specialization in the chromatin profile in terms of subgroups: the regions representing candidate enhancers, which are inactive in the normal donor state and thought to become functional in the cancer state, tend to be subgroup specific. Additionally, a high share of the consensus peaks for PC and MM subgroups are coming from the HD subgroup, something to be expected due to the high sample numbers, as already observed in Table 4-1.

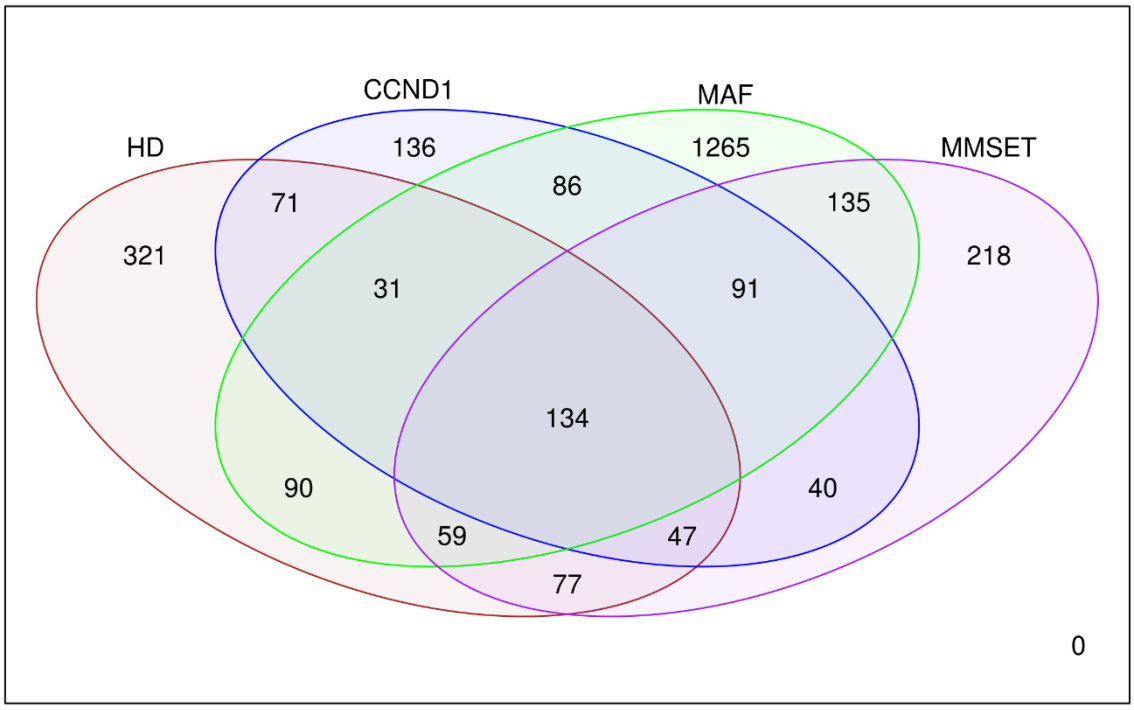


Figure 4-11: SMM enhancers overlap for each subgroup.

SMM enhancers for each MM subgroup: HD (red), CCND1 (blue), MAF (green) and MMSET (purple) overlapping the SMM enhancers.

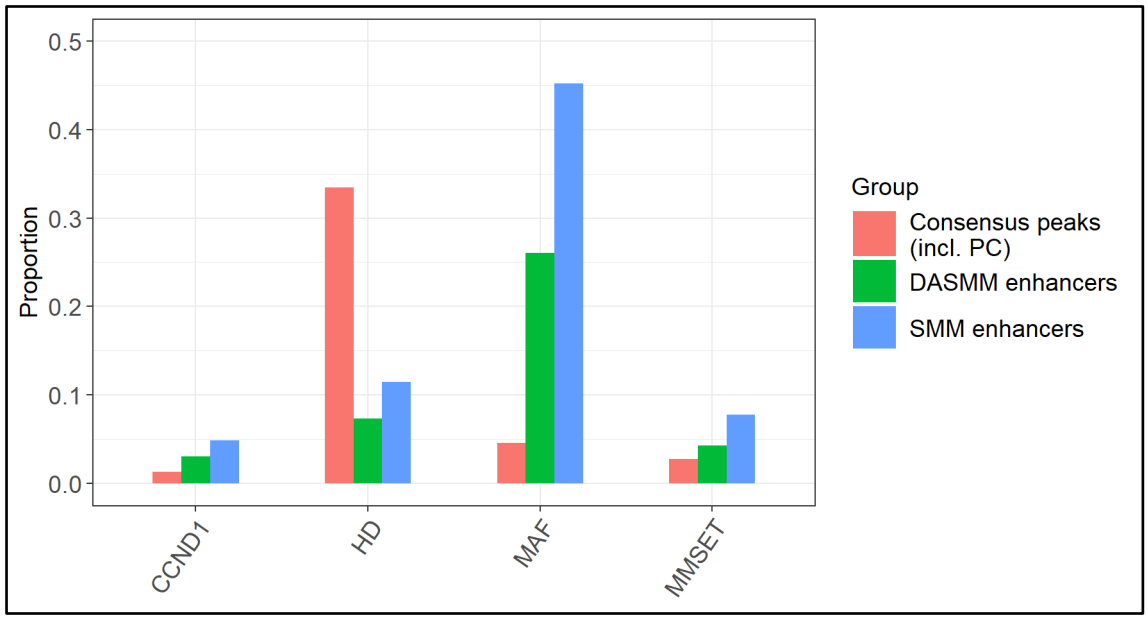


Figure 4-12: Proportion of exclusive MM subgroup regions from total.

For each MM subgroup (CCND1, HD, MAF and MMSET), the proportion of MM subgroup exclusive regions from all consensus peaks for PC and MM subgroups (pink bars, includes PC regions overlapping), DASMM enhancers (green) and SMM enhancers (blue) are shown. The pairwise Pearson's chi-squared test using Benjamini & Hochberg (1995) correction p-adjusted values between proportions within MM subgroups are all statistically significant (p-adjusted values less than 0.0001 for all cases).

	HD	CCND1	MAF	MMSET	Common to all subgroups	Total
SMM enhancers	321	136	1,265	218	134	2,801
DASMM enhancers	509	213	1,798	297	1,057	6,897
Ratio (activated/differential)	63%	64%	70%	73%	13%	41%

Table 4-2: Ratios of SMM and DASMM enhancers.

Ratios of regions exclusive to each subgroup (significant when compared to PC) are included: HD, CCND1, MAF and MMSET. Also regions common to all subgroups vs. PC and total regions.

Nearly half (1,265/2,801) of the SMM enhancers becoming active are exclusive to the MAF subgroup (Figure 4-11 and Figure 4-12 MAF subgroup blue bar), a very high proportion of the total enhancers found. This represents a high and statistically significant enrichment compared

with about a quarter: 1,798 out of 6,897 all the DASMM enhancers which are DA in the MAF vs. PC subgroup (Figure 4-12 MAF subgroup green bar). This points to the fact that MAF is the group deviating most in terms of chromatin accessibility profile of putative enhancer sites from the healthy state, with a general euchromatin state.

4.3.7. Transcriptomic profiles

Once the enhancer regions with a change in chromatin accessibility between healthy and subgroup MM state had been identified, I sought to study the MM subgroup transcriptome. RNA-seq profiles were obtained as stated in the Materials and Methods section and can be seen in Figure 4-13. In general all subgroups display a similar profile compared with PC (Figure 4-13 first column) with a proportion of the genes which become OE in MM (perhaps corresponding to oncogenes) and genes becoming suppressed in the cancer state (which could be tumor suppressing genes) with variable baseline expression. The MAF subgroup has the highest fold variation compared to PC in both extremes while the MMSET has the lowest. The relevant genes for each subgroup are marked: for each subgroup, the gene affected by the IgH enhancer translocation and CCND2. CCND2 expression is high and OE in the HD, MAF and MMSET subgroups and low in CCND1, pointing at a dichotomy in CCND1-CCND2 expression. The log fold change of this gene with respect to PC is comparable to the IgH enhancer translocated gene target overexpression for each subgroup. Additionally, as it is expected for the CCND1 subgroup, CCND1 is very highly OE compared with PC expression. There is also a correlation between the subgroups and their gene change with respect to PC (Figure 4-13 second, third and fourth column), as in the case for accessibility, this could be due to spurious correlation of ratios.

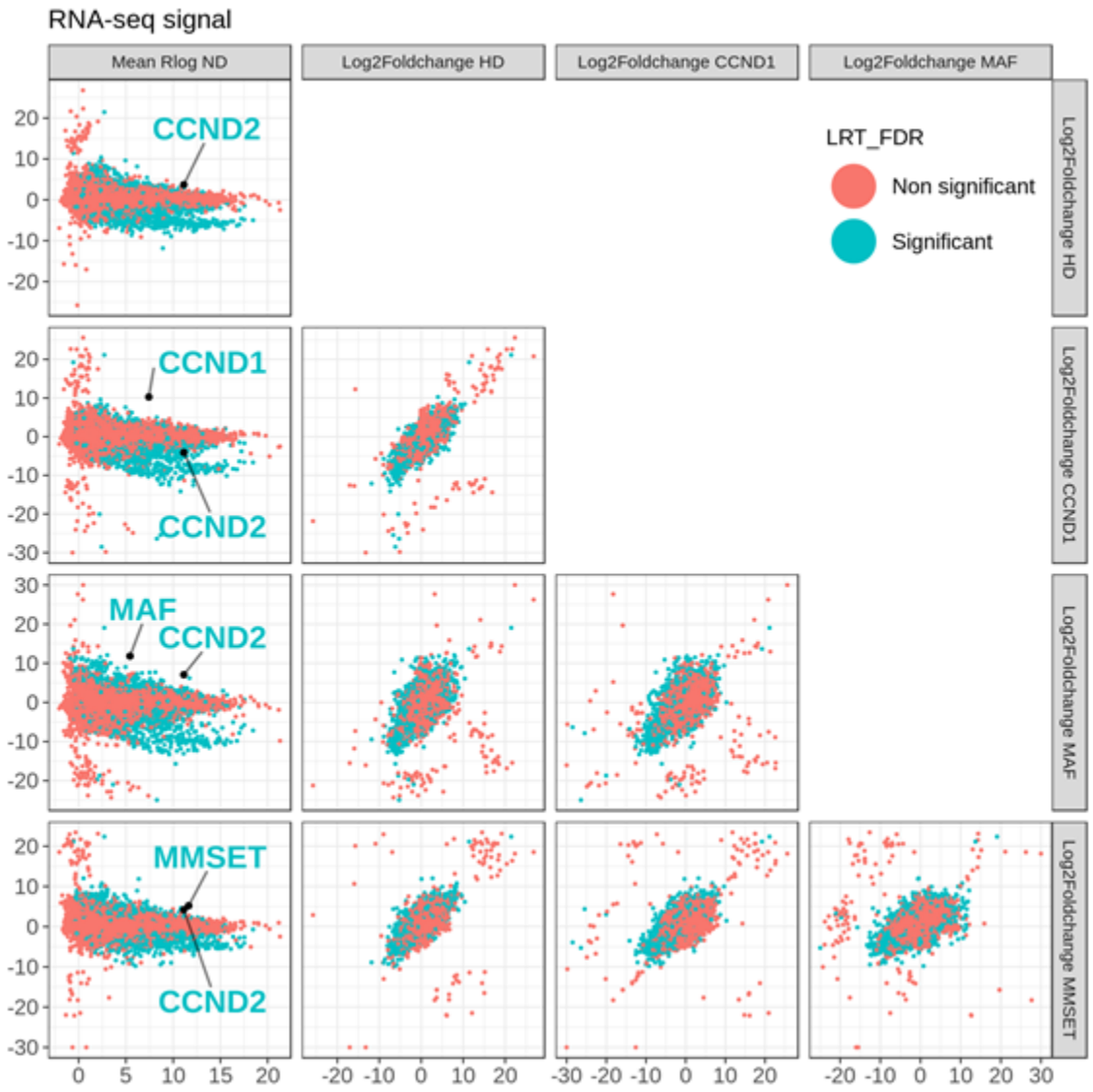


Figure 4-13: Transcriptomic profiles of the different MM subgroups compared with PC. (ND: PC)

Details for all the quantified genes. The first column shows the average Rlog (normalized) gene expression for the PC samples. The rest of the columns and rows show the Log_2fold gene expression between the samples of each subgroup specified and PC samples. Distinction is made between signals with significant Log Ratio Test (LRT) where the effect of the subgroup accounting for batch is significant. Relevant genes are marked for each subgroup.

4.3.8. DESMM genes

After the general gene expression profiles were studied, I sought to find a set of genes with significantly changed expression with respect to PC. Testing of the expression changes was performed similarly to the ATAC-seq analysis described above (but with outlier detection since DESeq2 was initially developed for RNA-seq analysis and it is more robust), (see Materials and Methods, section 2.3.6.2). The samples tested and the covariates can be seen in Table 2-1. 2,749 differentially significant genes, referred to as DESMM genes can be viewed in:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_sign_DE_genes_all_cond.tsv.gz

The number of DESMM genes in each subgroup vs. PC comparison can be seen in Figure 4-14. About a quarter of these genes are DE in all subgroups compared with PC. The MAF subgroup has the highest number of subgroup specific DESMM (298) followed by the MMSET group with 198, it is important to note that these two subgroups also have a significant number of group specific peak regions not accessible in PC, which could be regulating these genes (Figure 4-2 B). There is also a high overlap of DE genes between the MAF, HD and CCND1 group, perhaps showing a proportion of similar pathways deviating from their PC counterparts. The table with all the DE details for all the genes can be viewed in:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_all_DE_genes_all_cond.tsv.gz

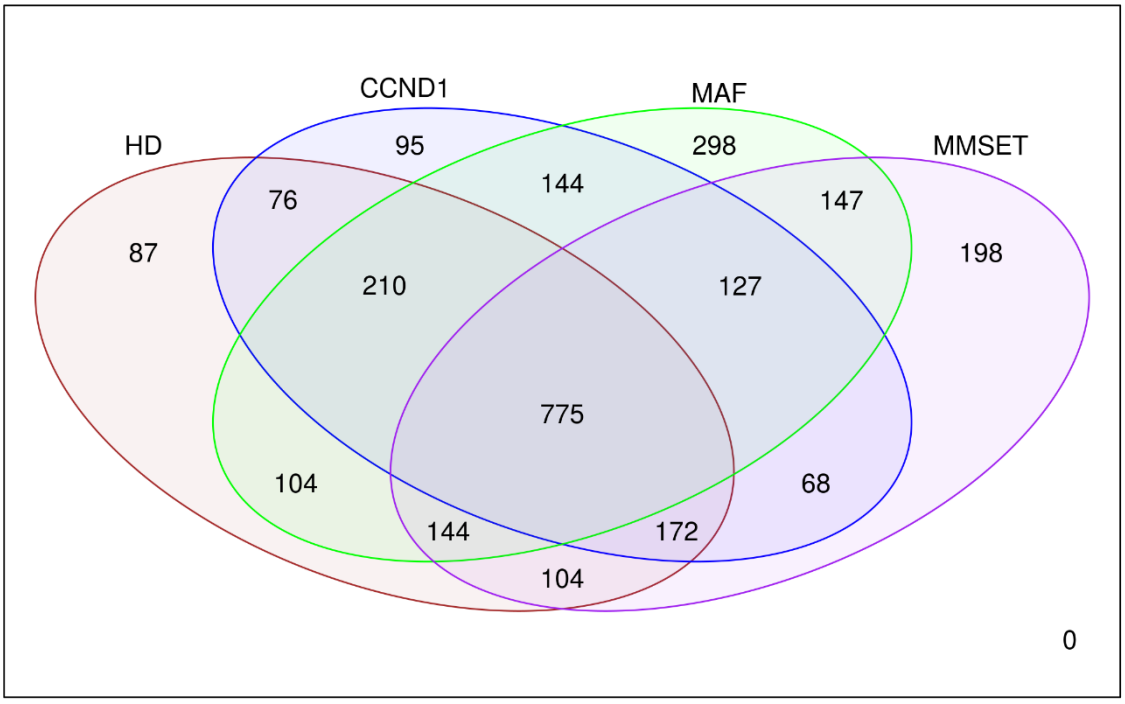


Figure 4-14: DESMM genes overlapping each subgroup.

DESMM genes overlapping each MM subgroup: HD (red), CCND1 (blue), MAF (green) and MMSET (purple).

To observe which cell processes are affected in terms of gene expression in the PC to MM subgroup transition, gene ontology analysis was performed testing the category enrichment in all the 2,749 DESMM genes (and subgroups specific DESMM) compared with the 57,992 genes which were quantified as specified in the Materials and Methods chapter. The results show

1,158 Biological Process, 105 Molecular Function and 105 Cell Cycle categories with overrepresentation in at least one of the subgroups or all the DESMM genes and can be viewed in:

MM_vs_PC_supervised_analysis/GO_subgroup_MM_vs_PC_DE_genes_Wallenius.xlsx

165 Biological Process and Molecular Function categories are overrepresented in all subgroups, some are expected, novel and general to gene expression such as “proximal promoter DNA-binding transcription activator activity, RNA polymerase II-specific” or plasma cell processes: “humoral immune response mediated by circulating immunoglobulin”, “positive regulation of leukocyte activation”, “killing of cells of other organism” or “ERK1 and ERK2 cascade” which is involved in differentiation of PCs (Yasuda et al., 2011). Also, the term “regulation of ossification” (process of cartilage to bone formation) is overrepresented in all subgroups except HD. Some terms associated with cancer could be found to be enriched, for example, “angiogenesis” (formation of new blood vessels commonly found in cancer) were found enriched in MAF and MMSET subgroups and previously found enriched in MM (Jin et al., 2018), “vasculogenesis” in MAF, “cell migration” and “regulation of B cell activation” in all but CCND1 and also in the literature (Hoang et al., 2018; Jin et al., 2018). “Chemokine production” is overrepresented in all subgroups, as an example, CX3CL1 is a chemokine which is found to be involved in MM angiogenesis (Marchica et al., 2019). Another pathway found enriched is cAMP (cyclic adenosine monophosphate) mediated signalling in all subgroups but MMSET subgroup and also found in the literature (Jin et al., 2018).

4.3.9. OESMM genes

To find a set of genes that could be regulated by subtype specific enhancers, DESMM were filtered to get only genes OE in at least one subgroup (see Materials and methods chapter). 1,664 genes fitting these criteria were found (referred to as OESMM genes). These can be found in:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_OE_RNA.gz

As Figure 4-15 shows, there are 240 genes which are OE in all subgroups vs. PC. All MM subgroups have a number of exclusive genes OE with respect to normal state which is either similar to the DESMM genes (Figure 4-14) in the case of the HD (72 OE, 87 DE) and CCND1 (73 compared to 95) subgroups or higher: MAF (346 versus 298) and MMSET (344 and 198). Accordingly, the proportions of MM subgroup exclusive genes from the total reflect an increase in DEMM compared to OEMM genes for each subgroup (Figure 4-16 pink and cyan bars respectively), being statistically significant for MAF and MMSET.

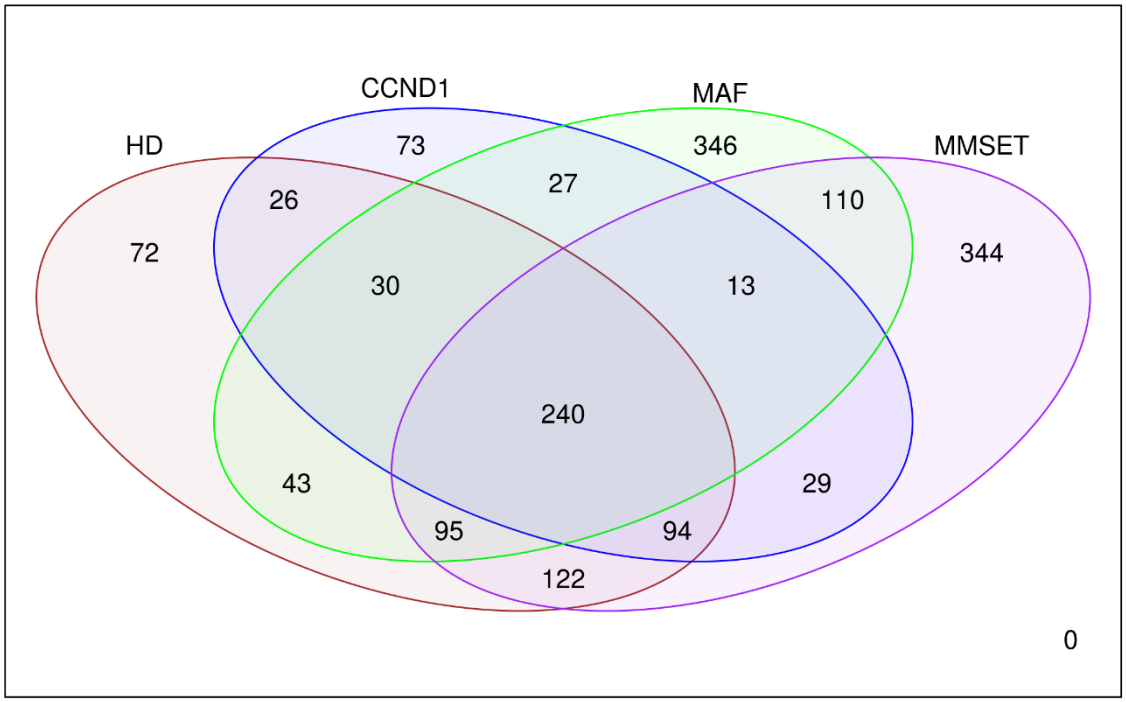


Figure 4-15: OESMM genes overlapping each subgroup.

OESMM genes overlapping each MM subgroup: HD (red), CCND1 (blue), MAF (green) and MMSET (purple).

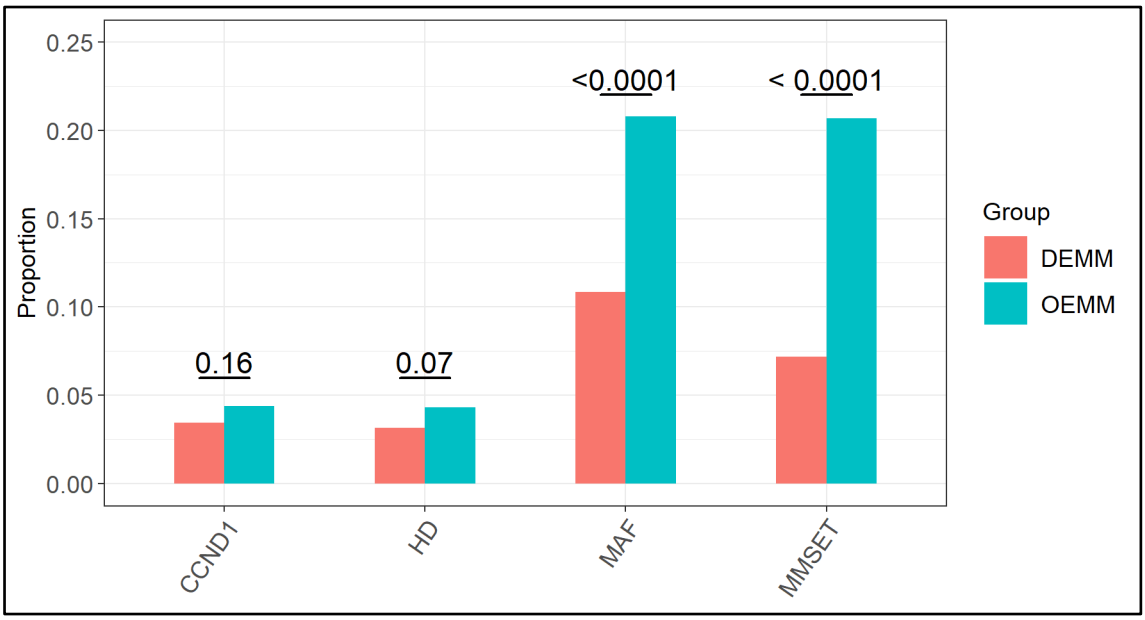


Figure 4-16: Proportions of MM subgroup exclusive DEMM and OEMM genes.

For each MM subgroup (CCND1, HD, MAF and MMSET), the proportion of subgroup exclusive genes from all DEMM genes (pink bars), OEMM genes (cyan) are shown. The pairwise Pearson's chi-squared test using Benjamini & Hochberg (1995) correction p-adjusted values between proportions within MM subgroups are shown above the bars.

In the case of MAF and MMSET, there might be more subgroup exclusive OESMM genes than DESMM genes because some DESMM genes are shared with other subgroups but down-regulated in the other subgroups and over-expressed in MAF or MMSET. Additionally, the OESMM genes shared by more than one subgroup (Figure 4-15) dramatically decrease compared to the DESMM genes (Figure 4-14) except for gene overlaps between MMSET and HD. There are 775/2,749 (28%) DESMM genes (Figure 4-14) which are shared by all subgroups with respect to the normal state (out of all DESMM genes) and 240/1,664 (14%) OESMM genes (Figure 4-15) which are shared by all subgroups compared with PCs. In terms of the transcriptome, this points to the different MM subgroups acquiring distinct profiles compared to the healthy state.

When the SMM enhancers are taken into account (Figure 4-11), it can be seen that MAF has the highest number of putative SMM enhancers (1,265) and OESMM genes (346, Figure 4-15). Despite this fact, this functional relationship between the number of putative SMM enhancers and OESMM is not very clear in general. For example, MMSET has a sixth (218) of the enhancers compared to MAF but there are 344 OE genes in MMSET (nearly as many as in MAF). Although enhancers may be redundant (Cannavò et al., 2016) and this may help to explain this fact, it is important to fine tune the requirements to consider enhancer regions as it was done later in this chapter.

To find which gene categories and biological pathways were enriched in these genes Gene Ontology enrichment was performed. All 1,664 OESMM genes (and subgroups specific OESMM genes) were compared to the 57,992 genes quantified and 965 Biological Process and Molecular Function categories with overrepresentation in at least one of the subgroups or all the OE genes between subgroup MM were identified:

MM_vs_PC_supervised_analysis/GO_subgroup_MM_vs_PC_OE_genes_Wallenius.xlsx

70 categories were overrepresented in all subgroups, some including cancer-related such as cell migration or the chemokine-mediated signaling pathway which could be involved in MM (Marchica et al., 2019). Other terms are present in some or all the subgroups which were found to be enriched, for example, angiogenesis as mentioned before, involved in MM (Ribatti and Vacca, 2018) and is enriched in all but CCND1 subgroups, regulation of aspects of the extracellular matrix such as organization or adhesion. The extracellular matrix remodelling has been shown to be of importance in MM (Glavey et al., 2017). Another example is regulation of cell proliferation in MAF.

4.3.10. DASMM enhancers regulating protein coding DESMM genes

To identify regions of open chromatin in subgroups vs. PC that might be regulating DESMM genes a window around DASMM enhancers of 1Mb was used as specified in the Materials and Methods chapter. There were 6,090 region – gene pairs are DA and DE in at least one subgroup vs. PC. These are referred to as DASMM enhancers regulating protein coding DESMM genes. They can be seen in the table:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_DE_ATAC_DE_RNA_1Mb.xlsx

The interactions corresponding to each subgroup and their overlaps that are active compared to PCs can be seen in Figure 4-17. There are 2,404 interactions where a subgroup may contain either ATAC, RNA or both either more accessible (and OE) or equal in PC with respect to all MM subgroups. The number of subgroup specific interactions between DASMM enhancers and protein coding DESMM genes are qualitatively correlated with the combination of the individual DASMM enhancer regions (Figure 4-8) and DESMM genes (Figure 4-14) as it is expected. For example, MAF is the dominant group with 1,399 interactions generated from 1,798 regions and 298 genes, this is followed by MMSET with 585 interactions and HD with 512, finally, CCND1 has 139 interactions. There are 106 interactions common to all subgroups.

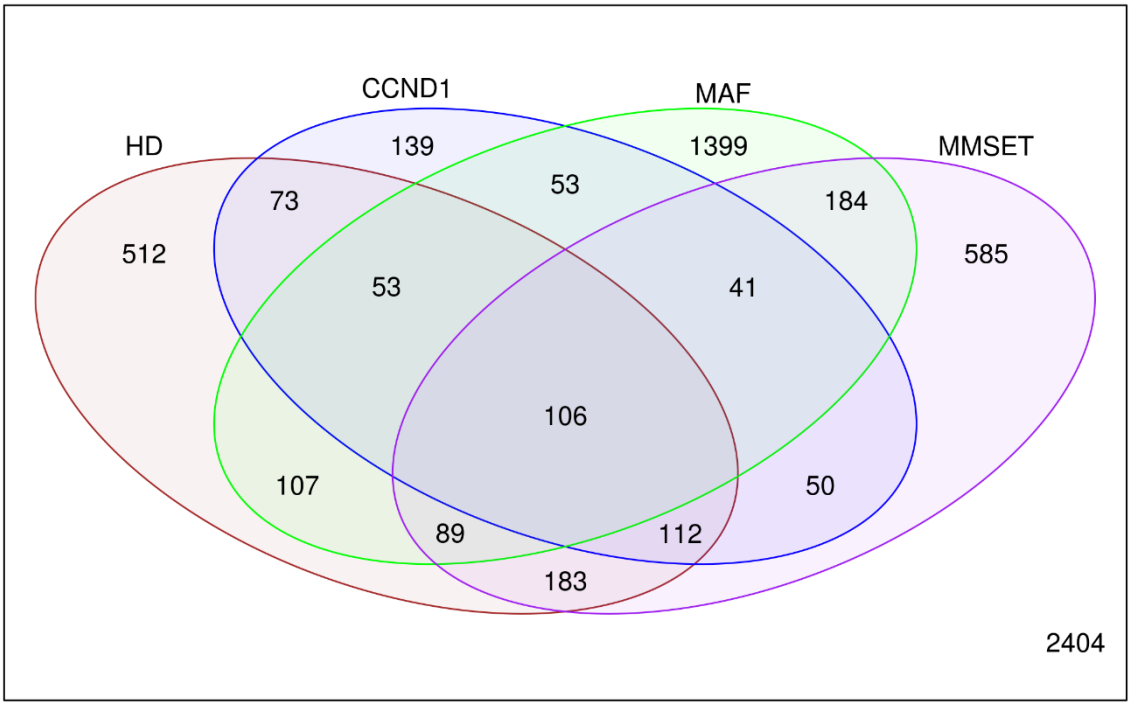


Figure 4-17: DASMM enhancers regulating protein coding DESMM genes. Enhancer – gene interactions overlapping each subgroup.

Protein coding DESMM genes regulated by DASMM enhancers overlapping each MM subgroup, a subgroup is considered overlapped if both the region and gene are over accessible and OE respectively compared to PC: HD (red), CCND1 (blue), MAF (green) and MMSET (purple). Outside the bubbles, bottom right: interactions that have either ATAC, RNA or both either more accessible (and OE) or equal in PC with respect to all MM subgroups.

4.3.11. Chromatin accessibility and gene expression subtyping classification profiles for DASMM enhancers regulating protein coding DESMM genes

Having determined pairs of DASMM enhancers and DESMM genes, I set out to determine whether these pairs could classify cytogenetically annotated and unannotated samples. Using DASMM enhancers regulating protein coding DESMM genes (individually for regions and genes), the samples were subjected to clustering (including samples with cytogenetic information unavailable).

As can be seen in Figure 4-18, all of the subgroup samples were grouped together using chromatin accessibility profiling. Clusters were determined by the average distance between the elements of one cluster and the elements of the other, creating clusters equally dissimilar to other clusters. The cytogenetically unannotated samples A27.20, A27.18 and A24.10 clustered together with HD samples, suggesting a similar chromatin accessibility profile. Additionally, the sorting suggests that the sample A17.5 could in fact be an MMSET cytogenetic sample (Figure 4-18 marked with a red arrow). To check this, the expression of the MMSET gene which is a hallmark of this subgroup was verified and the expression was at a level comparable of the samples containing the t(4;14) where the IgH enhancer is deregulating the MMSET expression (Figure 4-19).

Clustering was also performed on protein coding DESMM genes (Figure 4-20), clusters were determined by the shortest distance between the furthest elements in clusters, and this created groups where all elements contained in them had limited dissimilarity. The gene profiling based on protein coding DESMM genes also suggested that the sample A17.5 corresponds to the MMSET group (Figure 4-20 marked with a red arrow). Moreover, identically as with the chromatin accessibility profiling, the gene expression profiling also clustered all subgroup samples together and the samples with unknown cytogenetic information were categorized as HD samples.

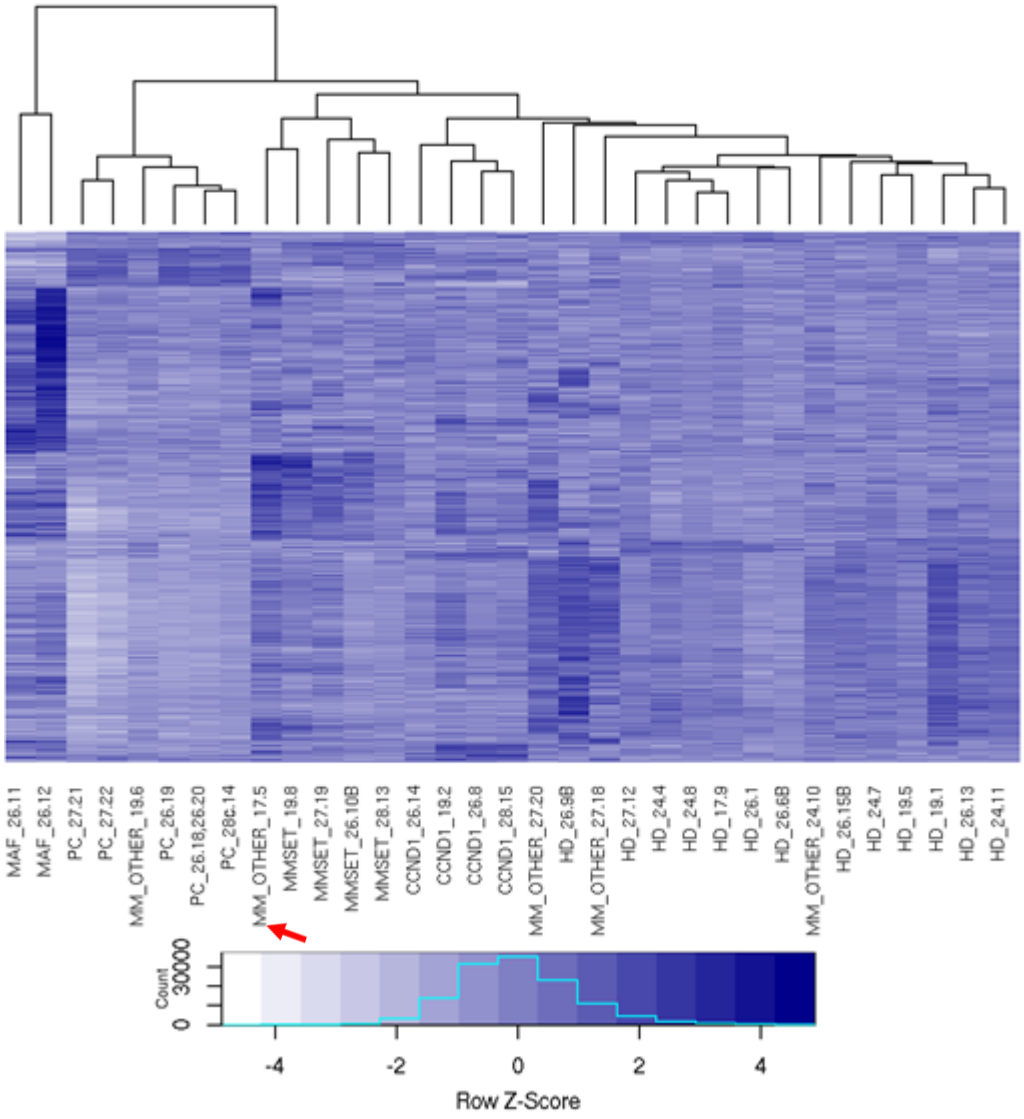


Figure 4-18: MM subgroup chromatin accessibility profiling in terms of regulatory DASMM enhancers.

Each row is a unique DASMM enhancer regulating a protein coding DESMM gene. Columns correspond to samples, names contain cytogenetic groups: Hyperdiploid (HD), translocations MAF, CCND1, MMSET, Plasma Cells (PC) and cytogenetically unannotated MM samples (MM_OTHER). The cells in the heatmap represent the rLog normalized reads in peaks. Samples and enhancers are hierarchically clustered on correlation metric and average linkage. Red arrow indicating non-cytogenetically annotated sample A17.5 clustering with MMSET translocated samples.

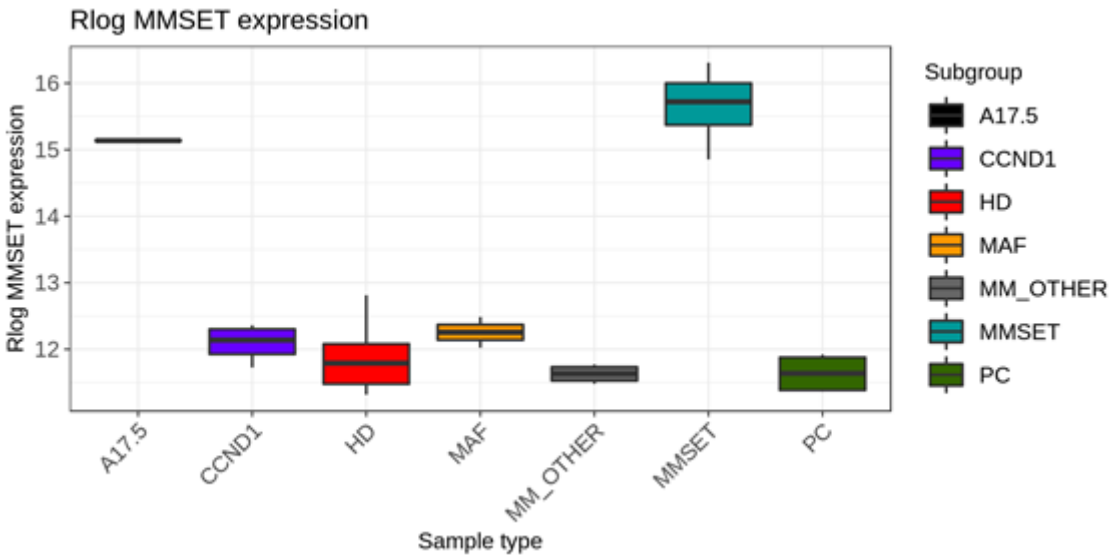


Figure 4-19: MMSET Gene expression for the different categories of primary samples.

Normalized rLog gene expression for all the MM subgroups, PC and the unknown sample A17.5 that clusters with MMSET samples in terms of accessibility in enhancer regions.

Together these results suggest that the DASMM enhancers regulating protein coding DESMM genes are very relevant features capable of classifying samples into their subgroups. This may suggest that the putative enhancers and regulated genes proposed are in fact genuine interactions although this requires further testing. It must be noted that samples corresponding to the same subgroup (such as MMSET) are clustered together based not only on the MMSET gene expression, but on a gene signature of a combination of multiple relevant genes. This was previously explored in the gene ontology for DESMM and OESMM genes. Finally, the chromatin accessible profile will be further explored for TF motif enrichment.

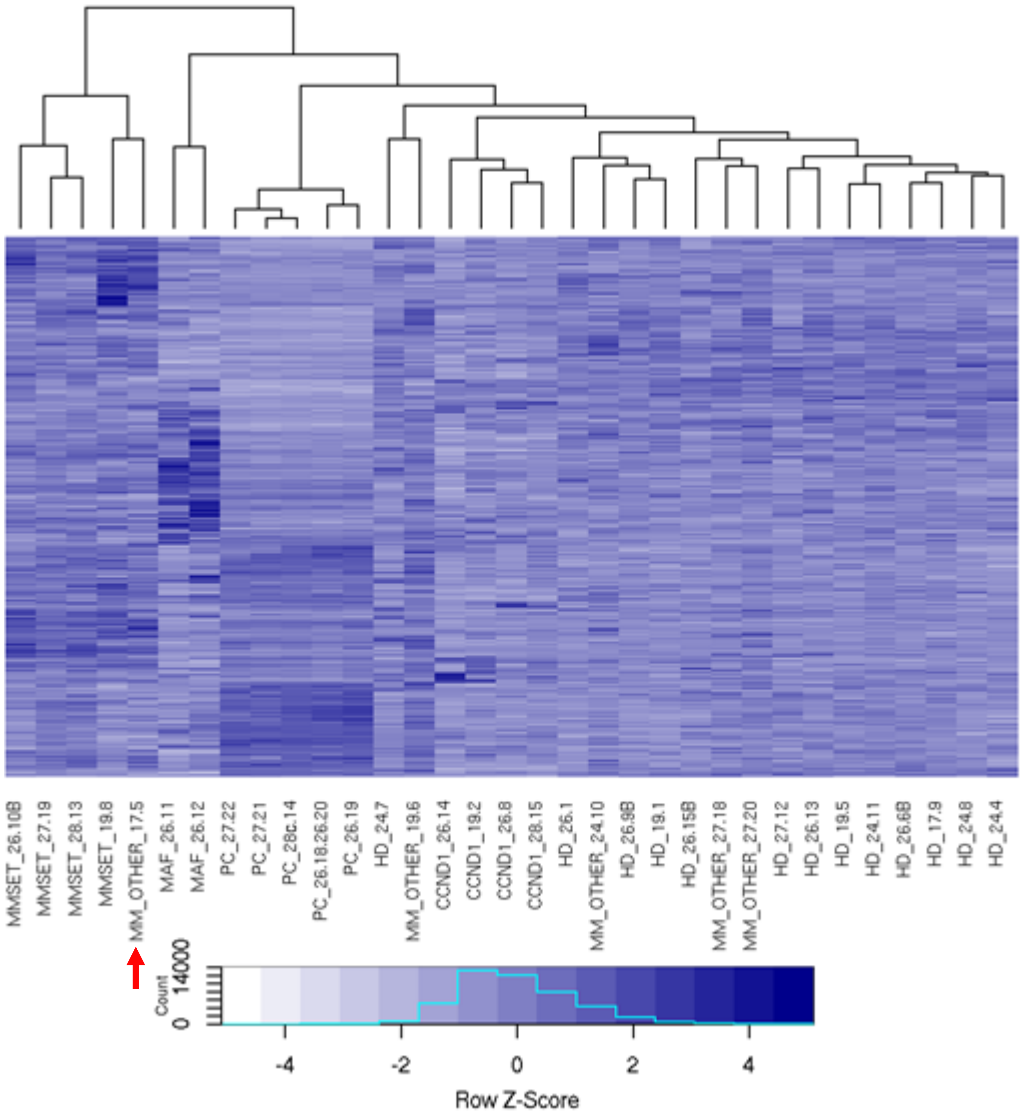


Figure 4-20: MM subgroup gene expression profiling in terms of regulated DESMM genes.

Each row is a unique DESMM gene regulated by a DASMM enhancer. Columns correspond to samples, names contain cytogenetic groups: Hyperdiploid (HD), translocations MAF, CCND1, MMSET, Plasma Cells (PC) and cytogenetically unannotated MM samples (MM_OTHER). The cells in the heatmap represent the rLog normalized reads in genes. Samples and genes are hierarchically clustered on correlation metric and average linkage. Red arrow indicating non-cytogenetically annotated sample A17.5 clustering with MMSET translocated samples.

4.3.12. SMM enhancers near OESMM protein coding genes

Potential subgroup specific enhancers are those that are increased in activity in cancer states and the genes they regulate will presumably be upregulated compared to PC. Thus I sought to relate the subgroup activated enhancers to the corresponding upregulated genes. 1,419 region – gene pairs were obtained as specified in the Materials and Methods chapter, they are

referred to as SMM enhancers regulating protein coding OESMM genes. They can be seen in the table:

MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_OE_ATAC_OE_RNA_1Mb.tsv.gz

Additionally, for each MM subgroup a table was created where the particular subgroup regions and genes are significantly more accessible and expressed than in PC respectively:

MM_vs_PC_supervised_analysis/CCND1_vs_PC_OE_ATAC_OE_RNA_1Mb.tsv.gz

MM_vs_PC_supervised_analysis/HD_vs_PC_OE_ATAC_OE_RNA_1Mb.tsv.gz

MM_vs_PC_supervised_analysis/MAF_vs_PC_OE_ATAC_OE_RNA_1Mb.tsv.gz

MM_vs_PC_supervised_analysis/MMSET_vs_PC_OE_ATAC_OE_RNA_1Mb.tsv.gz

As can also be seen in Figure 4-21, the greatest number of SMM enhancers near OESMM protein coding genes is coming from MAF (704), followed by MMSET (211), HD (169) and CCND1 (66). This correlates well, as it is to be expected, with the relationship between the OESMM genes (Figure 4-15) and SMM enhancers (Figure 4-11) which has MAF with 1,265 active candidate enhancer regions and 346 upregulated genes, MMSET with 218 and 344, HD with 321 region but only 72 genes and CCND1 with 136 and 73 respectively. Surprisingly, there are very few interactions common to all subgroups (13) despite having 240 upregulated genes and 134 active enhancer regions.

Consistent with this observation, when compared with the DASMM enhancers - protein coding DESMM genes interactions (Figure 4-17), there is subgroup specialization in the interactions for all subgroups except HD. For example, about 5% (66/1,419) of activated region - gene pairs are exclusive to CCND1, but only 2% of all differential region - gene pairs are. MAF is always the subgroup deviating most from PC, with MAF exclusive SMM enhancers and protein coding OESMM genes pairs accounting for about half of all the interactions.

Moreover, the interactions common to multiple or all subgroups drastically decline in the SMM enhancers compared with the DASMM enhancers. This result in the SMM enhancers regulating protein coding OESMM genes could be a consequence of eliminating any DA regions which are chromatin accessible in PC (as well as regions more chromatin accessible in PC), removing equally or OE PC genes, while also employing stricter thresholds than for differential gene expression and chromatin accessibility in subgroup comparisons with PC. The combined effect of applying this across multiple subgroup comparisons could decrease the number of overlapping interactions. This, however, increases the confidence in the interactions proposed.

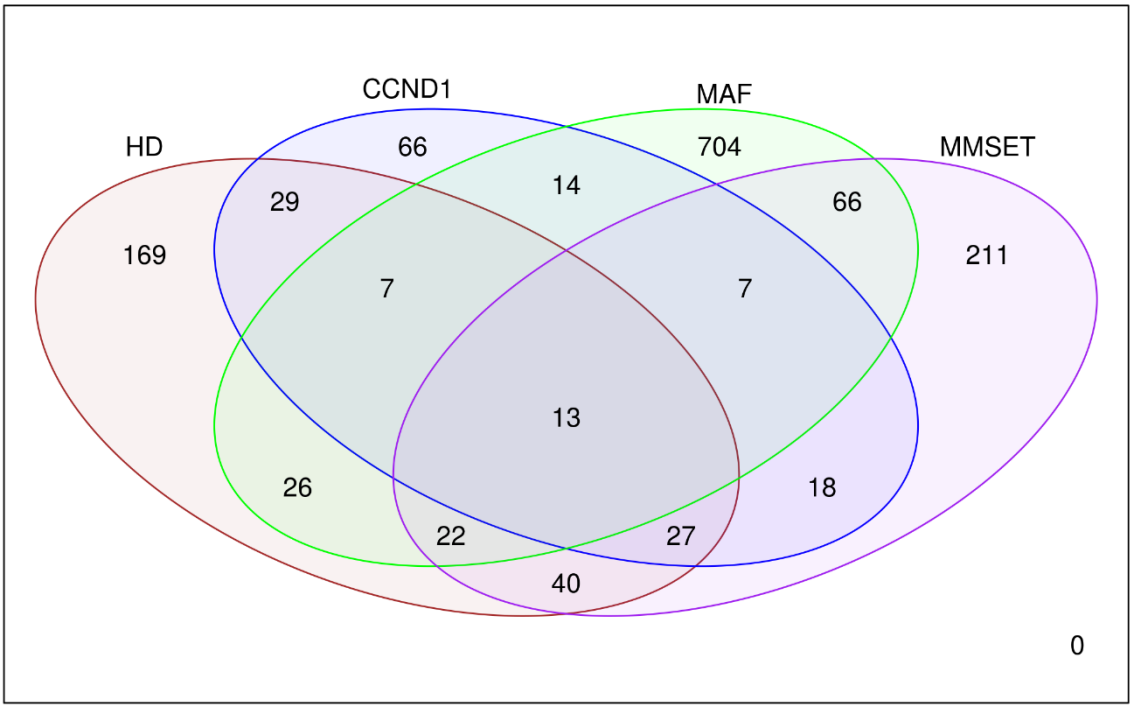


Figure 4-21: SMM enhancers near OESMM protein coding genes interactions. Enhancer – gene interactions overlapping each subgroup.

MM subgroups overlapping the interactions: HD (red), CCND1 (blue), MAF (green) and MMSET (purple).

As previously mentioned, it was observed, that primary MM samples tend to show either CCND1 or CCND2 overexpression. The mechanism of CCND1 overexpression is identified through the IgH enhancer translocation, while CCND2 remains unknown. An example of these more discriminatory interactions can be seen in Figure 4-22 (location 1), for the MAF subgroup with the CCND2 gene and a candidate enhancer (chr12:4,141,400-4,141,861). This chromatin accessible region is also present in the HD sample A26.9B but doesn't produce CCND2 overexpression in this sample, as it will be seen in chapter 5, this sample is an outlier in terms of chromatin accessibility. In addition, location 2 on Figure 4-22 corresponding to region chr12:4,142,636-4,143,378, is also a region predicted to regulate the CCND2 gene (DASMM enhancer regulating protein coding DESMM genes table). As can be seen in the corresponding table (MM_vs_PC_supervised_analysis/subgroup_MM_vs_PC_DE_ATAC_DE_RNA_1Mb.xlsx) this region is accessible throughout PC and MM samples (hence why it is not considered an exclusive SMM enhancer) but significantly more open in the MAF subgroup (nearly 8 fold more accessible than in PC). Since the chromatin accessibility is not an "on" or "off" scale, this region can be considered somewhat open in PC but significantly more accessible in MM and

becoming active and activating the CCND2 gene, resulting in overexpression. The CCND2 gene is expressed all throughout MM and PC samples but has a 128 fold increase in expression in the MAF subgroup compared with PC.

The region chr12:4,142,636-4,143,378 overlaps an enhancer region (peak 6) that was tested for CCND2 regulation (not shown) by (Alexia Katsarou, Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Foundation Trust in London, *pers. comms*.) through CRISPR-dCas9 repression in a MAF translocated cell line (to my knowledge JJN-3) obtaining a significant lower relative CCND2 expression comparable to CCND2 promoter repression (p-value is less than 0.01). Furthermore, Alexia Katsarou performed super-enhancer calling using the H3K27ac histone modification Chromatin Immunoprecipitation sequencing (ChIP-Seq) in MAF-translocated JJN3 cells and identified this region (and surrounding ones) as a putative super enhancer extending the region chr12:4,103,242-4,177,985 (not shown). The accessibility of the putative enhancer chr12:4,142,636-4,143,378 correlates with CCND2 expression (Figure 4-23 A). Region chr12:4,141,400-4,141,861 doesn't seem to be correlated with CCND2 expression (Figure 4-23 B) in general and it is not clear if this region is acting as an enhancer only in MAF or perhaps the MAF samples could just have a wide area of open chromatin including the aforementioned region.

Further supporting the regulatory effect of the selected tested enhancer, the samples in the present study tend to have either high CCND1 expression or high CCND2 expression (Figure 4-24 A and B), consistent with the literature regarding the CCND1 and CCND2 expression dichotomy (Chesi and Bergsagel, 2011; Shah et al., 2018; D. Smith et al., 2016) with CCND3 expression being somewhat lower in some samples with high CCND2 expression (Figure 4-24 B). Moreover, the expression of CCND2 tends to be correlated with the accessibility of region chr12:4,142,636-4,143,378 (implying regulation), with CCND1 abundance being anti-correlated with this candidate enhancer's accessibility, high accessibility for this region is found in CCND2 expression above 12.5 (Figure 4-24 A). Furthermore, there are 3 samples with relatively high CCND1 expression and higher CCND2: the HD samples 19.1 and 26.9B and the MMSET translocated sample 28.13. CCND3 is not found to be DE in any subgroup compared to PC. Additionally, 3D chromatin interactions, as measured by HiC, in the B-cell line GM12878 (which is a less differentiated state compared with PCs) showed some contacts for the 5Kb region containing the candidate regions and the CCND2 promoter (Figure 4-25 blue rectangle intersecting CCND2 promoter).

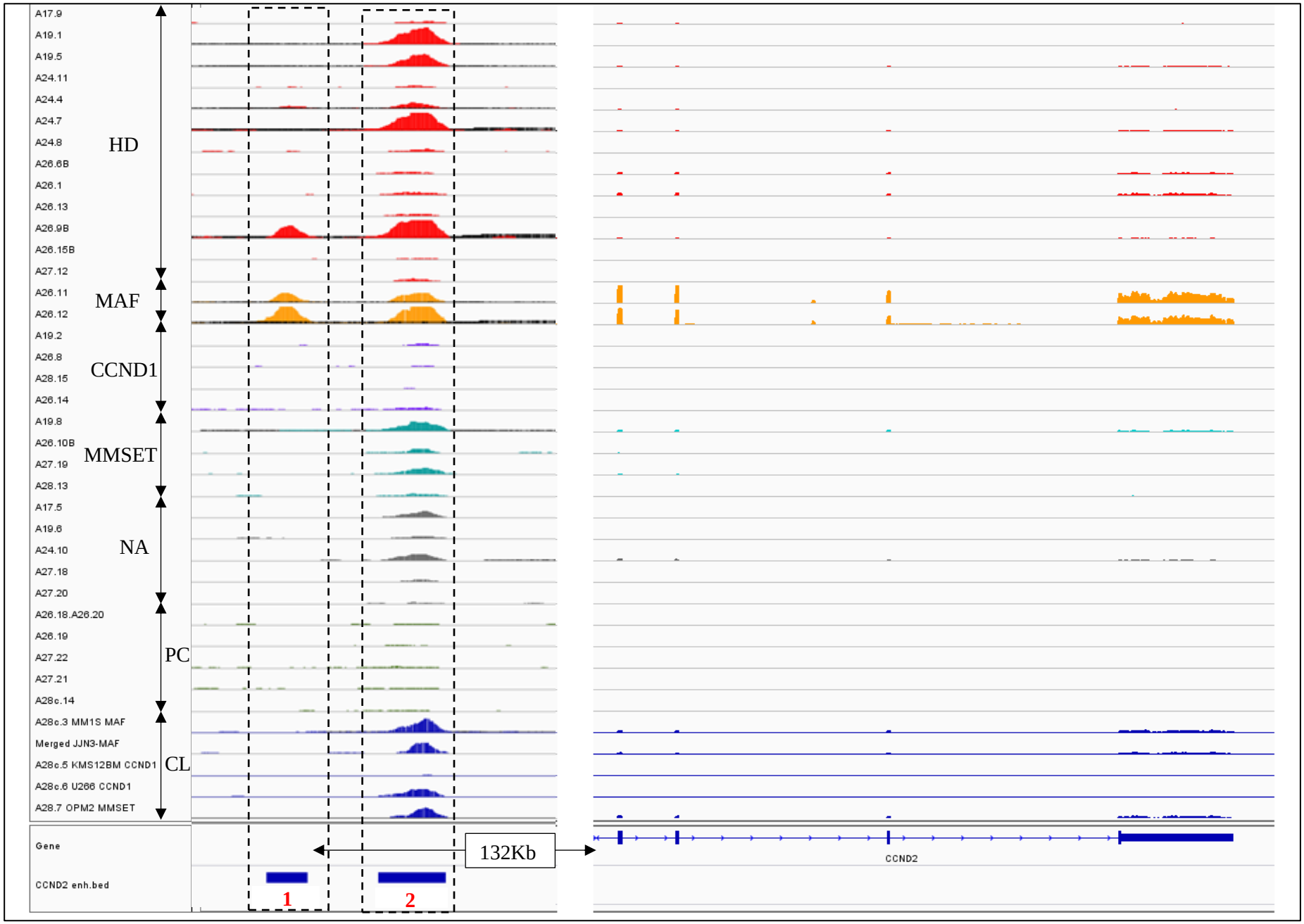


Figure 4-22: MAF CCND2 candidate enhancers.

Left: ATAC-seq tracks showing the different chromatin accessibility profiles of the different samples in different colours depending on the different subgroups: HD in red, MAF in orange, CCND1 in purple, MMSET in cyan, cytogenetically non-annotated samples in grey, PC samples in green and MM CL in dark blue. The baseline calculated chromatin accessibility signal (noise) is overlaid in black for all samples. On the left section, the enhancer at location 1 (chr12:4,141,400-4,141,861) and location 2 (chr12:4,142,636-4,143,378). The scale for each track is 0-3 (fragment pileup normalized for each sample per million reads).

Right: RNA-seq signal corresponding to each of the ATAC-seq tracks (each RNA sample is horizontally aligned with its patient's ATAC sample, ATAC-RNA sample correspondence can be seen in Table 2-1), the colours for the RNA-seq signal are the same as for the ATAC-seq signal. The scale for each track is 0-80 (reads mapping normalized by sample million reads mapped).

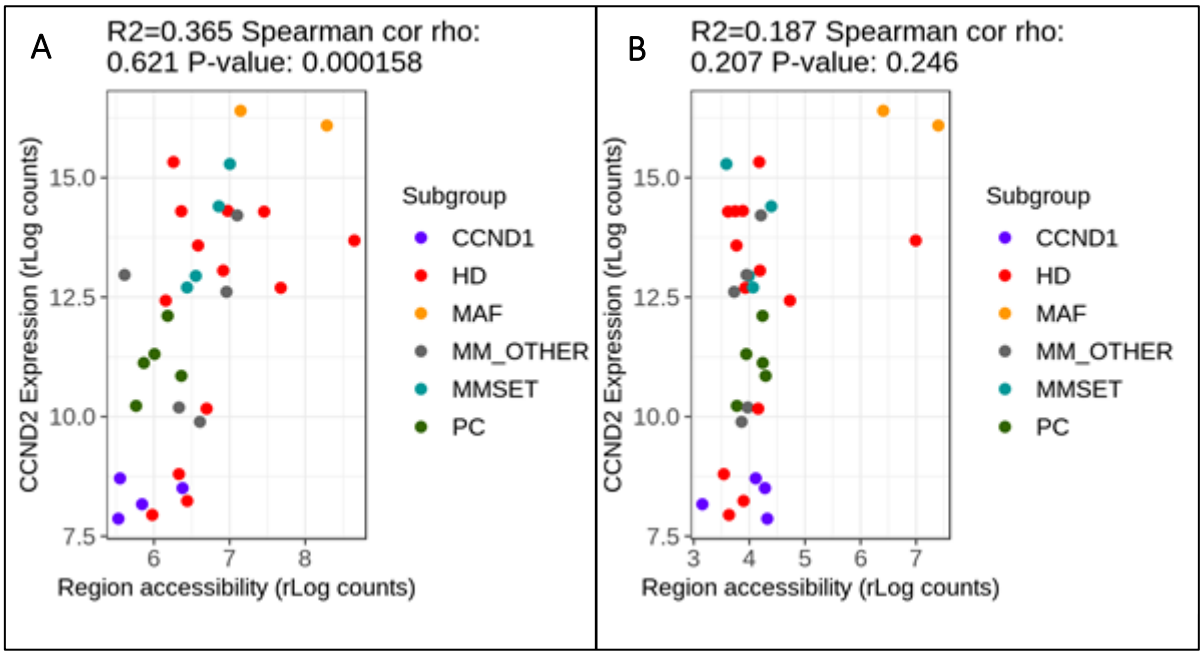


Figure 4-23: Correlation of chromatin accessibility with CCND2 expression for candidate enhancer regions.

CCND2 rLog normalized expression with the rLog normalized candidate regions accessibility. The regions shown are A: chr12:4,142,636-4,143,378 and B: chr12:4,141,400-4,141,861.

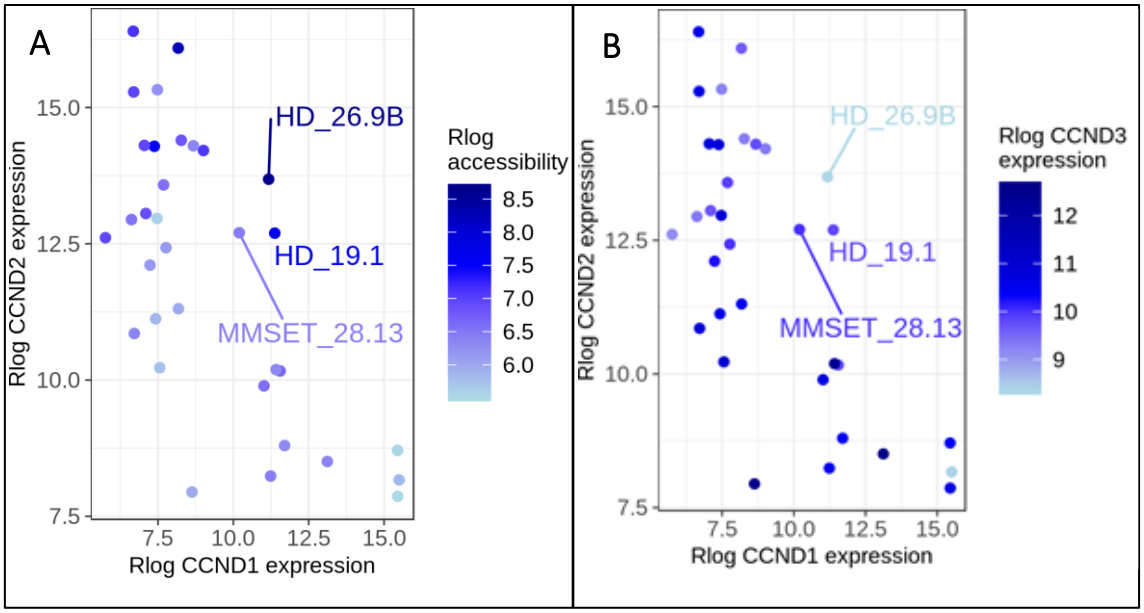


Figure 4-24: CCND1 - CCND2 dichotomy.

A: CCND1 with CCND2 rLog normalized expression and accessibility for the chr12:4,142,636-4,143,378 region.

B: Rlog normalized expression for CCND1, CCND2 and CCND3.

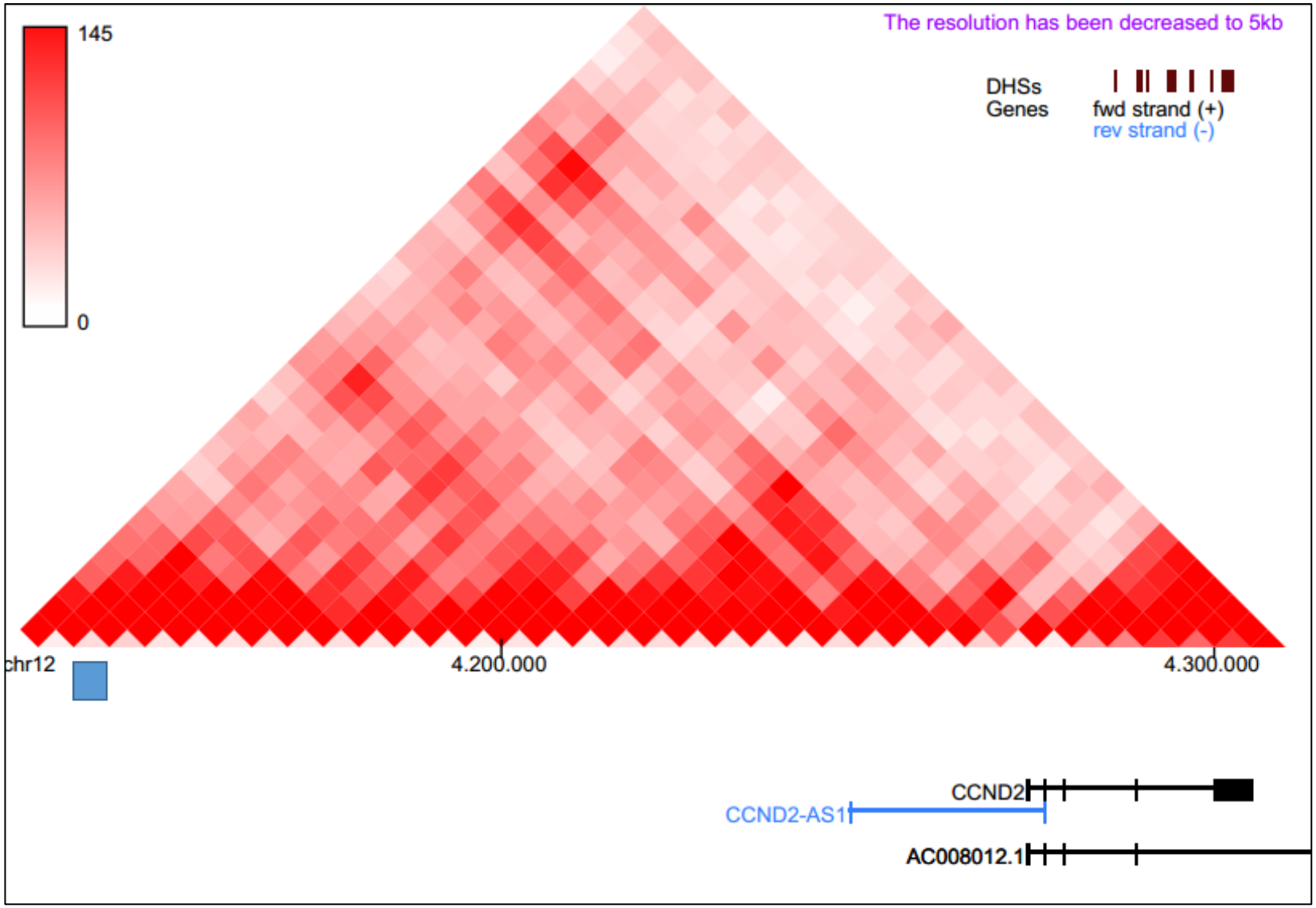


Figure 4-25: B cell line (GM12878) Hi-C for the CCND2 gene and CCND2 candidate enhancer regions.

Hi-C raw data on hg38 assembly at 5Kb resolution (Rao et al., 2014), location of the 5Kb regions of the studied candidate enhancers (chr12:4,142,636-4,143,378 and chr12:4,141,400-4,141,861) in the blue rectangle region. Visualization obtained using the Feng Yue lab at Northwestern University Genome Browser (Yue, n.d.).

4.3.13. TF binding in SMM enhancers regulating protein coding OESMM genes

Now that a set of subgroup specific genes that I hypothesise may be activatated via the triggering of nearby enhancers had been identified, I set out to determine which activator proteins might be binding to the active enhancers. Motif enrichment was performed on the regions for each subgroup with the most confidence of having regulatory potential: the unique SMM enhancers regulating protein coding OESMM genes obtained in the previous section.

As can be seen in the Materials and Methods section 2.6.2, the regions selected for motif enrichment for each subgroup are 143 regions for CCND1, 662 for MAF, 314 for MMSET and 264 for HD. They can be seen in:

MM_vs_PC_supervised_analysis/Subgroups motif enrichment/CCND1_enhancers.bed.gz

MM_vs_PC_supervised_analysis/Subgroups motif enrichment/HD_enhancers.bed.gz

MM_vs_PC_supervised_analysis/Subgroups motif enrichment/MAF_enhancers.bed.gz

MM_vs_PC_supervised_analysis/Subgroups motif enrichment/MMSET_enhancers.bed.gz

Thirty-eight TF motifs were found to be enriched in SMM enhancers (any subgroup) regulating protein coding OESMM genes (Figure 4-26). Most of the enriched TFs have been found to be relevant in MM, for example, BCL6 is found to be OE in patient MM cells in coculture with bone marrow (BM) stromal cell-culture supernatant (Hideshima et al., 2010), ERG expression is discovered to be high in MM samples (Knief et al., 2017) or MAZ OE and activating MYC expression in a subset of MM samples (Zhan et al., 2002). Furthermore, CTCF, ELF2, ETV family, IRF family, POU2F1, POU2F2, PRDM1, RXRA, SP3 and SPI1 are found to be MM CL super enhancer-associated TF genes (Jin et al., 2018). Each MM subgroup has specific TF motif enrichment in its activated enhancers, some (such as CTCF and CTCFL) are common to all subgroups. HD and MMSET have the larger set of the enriched TFs. It must be noted that because of thresholds used, it is possible that despite TFs appearing as non-enriched for certain subgroups, they might be binding to a very small subset of enhancers.

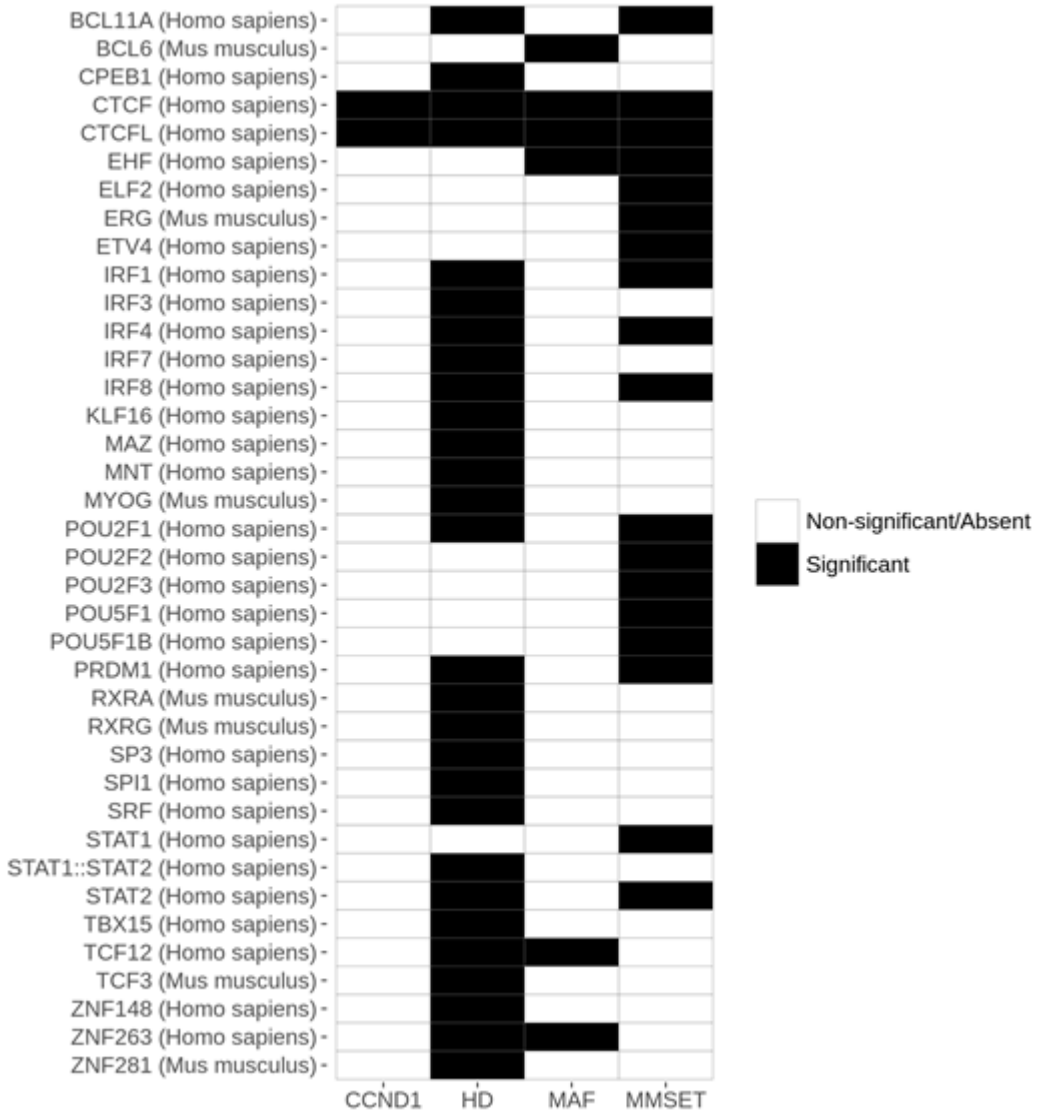
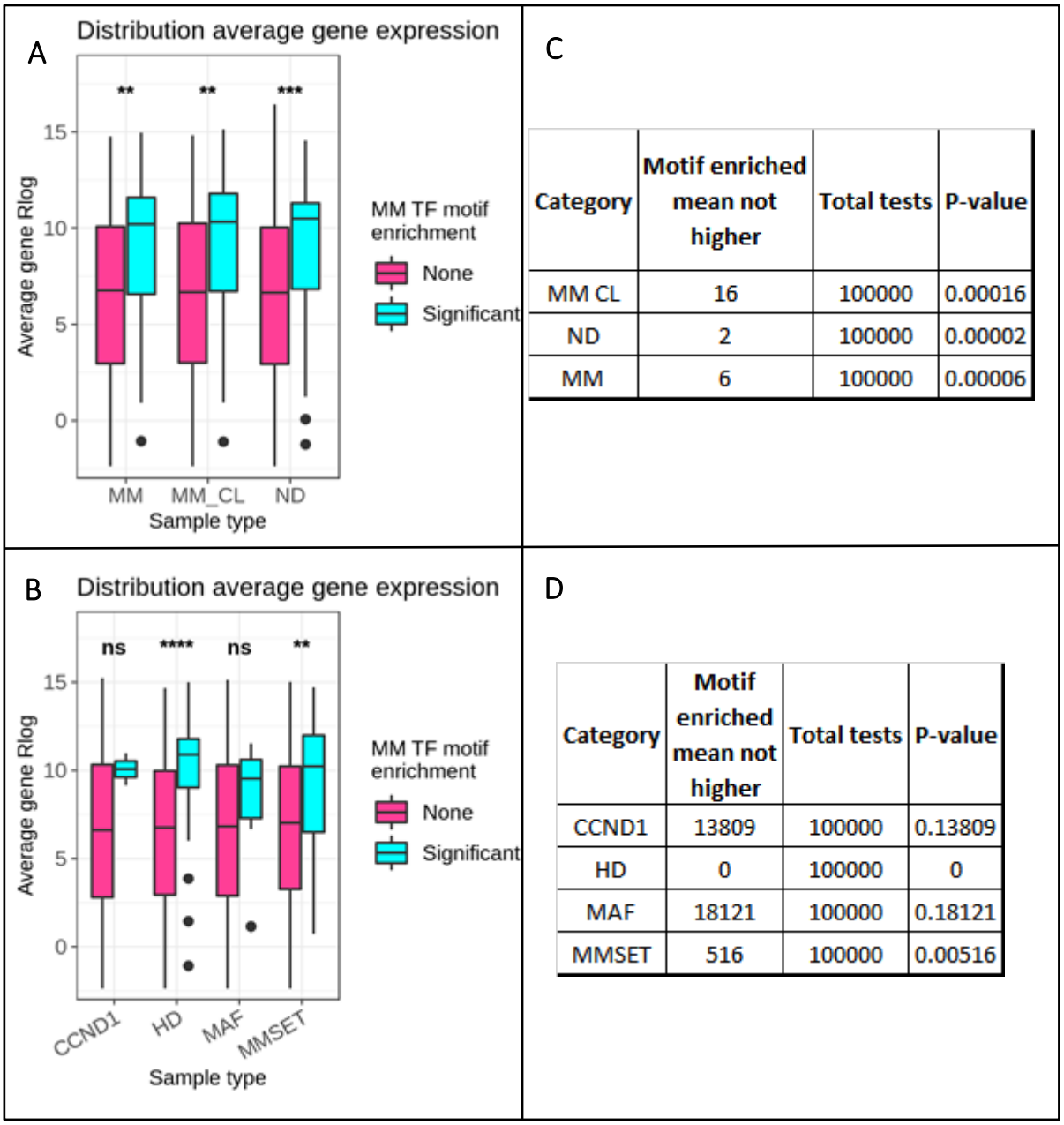


Figure 4-26: TFs motif enriched in DASMM enhancers regulating protein coding DESMM genes.

Black boxes indicating significant enrichment in a particular TF for the regions in the DASMM enhancers – protein coding DESMM that are DA and DE for a subgroup vs. PC.

If these enriched TFs are to regulate the subgroup specific enhancers, it is important to determine whether they are readily available in the cells to activate the putative enhancer regions and promote the expression of myeloma related genes. The method described in the section 2.6.3 of the Materials and Methods was used. The TF genes with significant motif enrichment in any sample group were statistically significantly more expressed than those with non-enriched binding motifs (Figure 4-27 A) when considering all PC samples (labelled “ND”), primary MM samples (labelled “MM”) and MM CLs (labelled “MM_CL”).



C

Category	Motif enriched mean not higher	Total tests	P-value
MM CL	16	100000	0.00016
ND	2	100000	0.00002
MM	6	100000	0.00006

D

Category	Motif enriched mean not higher	Total tests	P-value
CCND1	13809	100000	0.13809
HD	0	100000	0
MAF	18121	100000	0.18121
MMSET	516	100000	0.00516

Figure 4-27: Expression of TF motif enriched in DASMM enhancers regulating protein coding DESMM genes.

A: Average rLog normalized gene expression distribution of any MM subgroup enriched TFs (37, not counting the STAT1::STAT2 dimer in Figure 4-26) compared with the non-enriched TFs in any MM subgroup (388) for different conditions (MM, MM CL, ND: PC). Significance shown for two-sample one tail Kolmogorov-Smirnov test within each group of samples (MM, PC and MM CLs), null hypothesis: distribution of gene means not greater for TFs with significant motif enrichment in any MM subgroup compared with TFs without significant motif enrichment in any MM subgroup.

B: Same as A, but comparing the average rLog normalized gene expression distribution of the particular MM subgroup enriched TFs (not counting the STAT1::STAT2 dimer in Figure 4-26) with the non-enriched TFs for the particular subgroup (from all the 425 TF genes considered in this section) for different conditions MM subgroup samples (CCND1, HD, MAF and MMSET). Significance shown for two-sample one tail Kolmogorov-Smirnov test within each group of samples, null hypothesis: distribution of gene means not greater for TFs with significant motif

enrichment in particular MM subgroup compared with TFs without significant motif enrichment in particular MM subgroup.

C: Results for 100,000 permutation tests, each test creates a random sample of size the number of TFs with a motif significantly enriched in any MM subgroup from all the 425 TF genes considered in this section, and compares it with the TFs with a motif significantly enriched in any MM subgroup (not considering the STAT1::STAT2 dimer). Within each comparison, the mean of the mean expression of the selected TFs was calculated for the samples considered (MM, ND:PC or MM CL) for both groups of samples. The number of tests having the random sample mean TF gene mean expression higher or equal to that of TF motif enriched annotated genes in any MM subgroup was obtained and p-values were calculated.

D: Same as C but using gene expression for each particular MM subgroup specific enriched motif binding TFs compared with that of all considered TFs.

To account for the fact that distribution of TF gene means is unknown, an additional permutation test was performed (Figure 4-27 C), also obtaining statistically significant results that the enriched TF motifs were more highly expressed than the non-enriched motifs. The average subgroup gene expression distribution for the enriched TFs in each subgroup was higher compared to that of the non-enriched TFs in that subgroup (Figure 4-27 B) but only statistically significantly higher for the HD and MMSET subgroups. The mean of the mean gene expression for enriched TFs in each subgroup was also higher than that of all the TFs (Figure 4-27 D) but again, only statistically significant for the HD and MMSET subgroups. When combined, the results show strong evidence that the regions with regulatory potential that are becoming more chromatin accessible in MM and hypothesized to regulate genes, are enriched with relevant TFs in PC, MM in general (HD and MMSET in particular) and MM CL. This is consistent with the fact that these regions are thought to become active by binding of these TFs. Within MM samples, these pathways leading to cancer are significant in the HD and MMSET groups as can be seen in Figure 4-27. These subgroups have the great majority of enriched motifs, perhaps exhibiting this mechanism to deregulate genes more profoundly.

4.4. Discussion

In this study, samples with higher number of single ends contain a higher number of corresponding accessible peaks while the latter is only weakly correlated with the assigned fraction. Therefore, out of the covariates studied (Figure 4-1), in general, the distribution (not the number) of chromatin accessible reads from ATAC-seq is the factor dictating the assigned fraction. A higher assigned fraction implies a higher difference between background and higher

accessibility. This means that MAF, MM CL samples and HD samples (all containing a high assigned fraction) could have more total areas of high accessibility with very large contrasts compared to the surrounding regions while the rest of the subgroups might have a higher proportion of the DNA in a somewhat accessible state to a large degree. For HD samples, the ratio of peaks to input reads is around half that for MAF and MM CL samples. Its high assigned fraction might be explained by a significant proportion of the peak regions overlapping genomic areas that have been duplicated (a feature of HD). This could translate to one peak being assigned reads from multiple duplicated regions. Finally, it must be noted that 2 out of 5 MM CLs contain a MAF translocation, this might explain why MM CLs samples have similar metric ratios as the MAF samples. In this regard, it is possible that MAF overexpression or its downstream repercussions are creating this effect on the chromatin.

A set of 295,238 chromatin accessible regions relevant in the PC – MM subgroup context were produced. The different sequencing depth of each sample was taken into account by subsampling all samples to a common sequencing depth corresponding to the lowest of all samples included in the study (28,231,242 single end reads). These were used to obtain the PC and MM subgroup consensus regions from which the rest of regions of interest were derived. The number of samples in each subgroup was unbalanced (Table 4-1): HD is the highest with 13, while MAF is the lowest with 2, in between MMSET and CCND1 with 4 and PC with 5 (3 donors with different CD19 status). Despite this, the number of subgroup peaks doesn't correlate with the number of samples (Table 4-1), with MAF and PC producing the largest number of peaks per sample. Various factors may be influencing this such as the heterogeneity of the accessibility of the samples corresponding to each subgroup or how close the subgroup peaks are to saturation.

Despite balancing the per sample sequencing depth, the total sequencing depth per subgroup (taking into account the total number of samples per subgroup) was not accounted for. The reason for this was that for chromatin accessible peak calling, a minimum starting material is required for each sample to build up signal and distinguish genuine regions from noise. Additionally, if the lowest number of reads per sample and lowest number of samples per subgroup across the whole set had been used, most of the sequencing material would not be used in this process (particularly in the case of HD), since there are more than 6 times more samples than for the MAF subgroup. Finally, once the consensus regions were used as features, quantification of each full sample signal was used in each consensus region to consider subgroup specific enhancers. This accounts for subgroup specific balancing of reads in each consensus region but is also susceptible to some groups of samples being more

deterministic than others in determining a subgroup effect: for example, since there are only two MAF samples, an area containing a high accessibility in one MAF sample has a high influence in determining a MAF subgroup open chromatin region (also applicable in RNA-seq changes). The strategy employed means that for the starting consensus regions, the degree to which chromatin accessible profiles are studied is different for each subgroup and it has to be taken into account when considering subgroup specific regions throughout this work, nevertheless, it also presumably correlates to the incidence of each subgroup's primary oncogenic driver in patients. Consistent with this, HD has the highest number of exclusive accessible consensus regions respectively with 98,801 (Figure 4-2 A) about one third of all consensus peaks for PC and MM subgroups when including overlapping PC accessible regions (Figure 4-12 pink bar for HD category). This also suggests that is it possible that the accessible regions for other MM subgroups might not have saturated and additional samples may be required. MAF is the second MM subgroup with most exclusive consensus peak regions with 13,473 (Figure 4-2 A) regions which could be a result of the high per sample influence explained above.

There is an enrichment of promoter regions in consensus accessible regions throughout all subgroups consistent with the fact that promoter accessibility tends to be a requisite of gene expression (chapter 3). Furthermore, the general nature of chromatin accessible changes required from the normal to the cancer state differs between each subgroup with CCND1 and HD having an enrichment in accessibility gain in already PC accessible regions while MMSET has a high proportion of *de novo* chromatin accessible regions and MAF lying somewhere in between.

DASMM enhancers and SMM enhancers have MAF as the subgroup deviating most from the PC state in terms of chromatin accessibility, the number DA regions shared across all subgroups vs. PC [1,057/6,897 (15%)] dramatically decreases when considering only activated enhancers shared by all subgroups vs. PC [134/2,801 (5%)], pointing at a MM subgroup specialization in the chromatin profile.

Similarly to accessibility, there is also a MM subgroup specialization acquiring distinct transcriptomic profiles compared to the healthy state when going from DE to OE genes but in contrast with accessibility, the degree to which this occurs is lower in absolute and relative terms. There are 775/2,749 (28%) DESMM genes which are shared by all subgroups with respect to the normal state and 240/1,664 (14%) OESMM genes which are shared by all subgroups compared with PCs. Not only is the relative ratio of DE features shared by all

subgroups double for genes compared to regions but when considering only activated features in all subgroups the relative ratio is triple. This suggests a higher importance of subgroup specific chromatin accessibility changes. MAF had the highest range in gene expression change with respect to PC while MMSET had the lowest variability.

MAF and MMSET had an enrichment for subgroup exclusive DESMM genes and OESMM genes with DESMM genes associated pathways such as vasculogenesis (MAF), chemokine production (all subgroups) found to be involved in MM angiogenesis (Marchica et al., 2019) and cAMP (cyclic adenosine monophosphate) mediated signalling (all subgroups but MMSET subgroup). Interestingly, despite the fact that the cAMP-mediated signalling pathway is also enriched in a study (Jin et al., 2018) in 11/23 MM samples, it is not enriched in any of the 3 MMSET t(4;14) translocated samples, consistently, in the analysis presented in this thesis the ontology category is overrepresented in all MM subgroups except MMSET. As Jin et al. mentioned, targeting of the cAMP has been performed and the study in this thesis has extended the fact that the MAF translocated samples may also have a similar response as all subgroups except MMSET t(4;14). OESMM genes have associated pathways related to aspects of the extracellular matrix such as organization or adhesion, of importance in MM (Glavey et al., 2017). Similar to the accessibility, each of the MM subgroup transcriptome becomes very specialized.

As it was done in Chapter 3, a range of 1Mb was used to relate DASMM enhancers regulating protein coding DESMM genes. These interactions were shown to be key for unsupervised sample classification into subgroups (from a chromatin accessibility and gene expression independent point of view), consistently classifying cytogenetically annotated and unannotated samples (with evidence for correct grouping). They are also strong candidates to be considered genuine interactions for MM patient classification, prognosis and to be exploited as therapeutical targets.

As mentioned in chapter 3, Hoang *et al.* previously linked 6 interactions on the basis of recurrent region mutations affecting candidate genes in MM with one of them and an additional being associated with cytogenetic subgroups. In this study, the interactions discovered by Hoang *et al.:* chr2:165,615,060-165,624,028 with COBLL1, chr9:37,375,172-37,395,282 with PAX5 and chr1:16,944,603-16,958,779 with ATP13A2 were proposed as enriched in 1% - 6% of the MM cases considered. Despite not overlapping, the present study in this thesis found multiple regions upstream of the COBLL1 and PAX5 thought to regulate these genes and being more accessible in most subgroups compared with PC. Additionally, the

regions chr1:16,444,028-16,444,514 and chr1:17,027,461-17,027,702 are found more active in the MAF subgroup and thought to regulate ATP13A2 in the subgroup analysis. The work in this thesis therefore provides novel MM subgroup resolution into interactions.

Finally, SMM enhancers regulating protein coding OESMM genes are also obtained, there are very few interactions common to all subgroups (13) pointing at a clear subgroup specialization. It is observed that primary MM samples tend to have either high CCND1 or CCND2 expression and one studied illustrative example of gene regulation occurring in the MAF subgroup is found in the region chr12:4,142,636-4,143,378 regulating the CCND2 gene which is key in PC and MM biology (Bergsagel et al., 2005; Nahar et al., 2011; Tooze, 2013). Additionally, CCND2 is involved in PC differentiation biology: in pre-B cells, once the pre-B cell receptor is activated, cell cycle exit is induced through high BCL6 expression which targets and represses MYC and CCND2 (Nahar et al., 2011). Furthermore, high levels of CCND2 are required for the transition from Plasmablasts to PCs (Tooze, 2013).

Cyclin D1, D2, or D3 are genes involved in cell cycle progression and expression of at least one appears to be altered in practically all MM tumours, also, profiling this expression has been shown relevant in MM stratification (Bergsagel et al., 2005). In terms of D-cyclin deregulation in MM, there are two models proposed for how they are an oncogenic initiating event contributing to malignancy: on the first one, healthy PCs are assumed to have exited the cell-cycle irreversibly and deregulating D-cyclins establishes an “abnormal quiescent, but not post-mitotic state, which is not observed in normal plasma cells” (Tooze, 2013), the second one regards normal PCs as having this state and deregulation of D-cyclins just lowers the threshold to allow for cell-cycle re-entry during malignancy (Tooze, 2013).

Prior studies have pointed at a general CCND1 – CCND2 dichotomy in MM cases (D. Smith et al., 2016), having high CCND1 expression in CCND1 t(11;14) translocated samples and HD samples with polysomy in chromosome 11 leading to biallelic CCND1 expression while having high CCND2 expression in MAF t(4;14) and MMSET t(14;16) translocated samples and a subset of HD and non-HD (Shah et al., 2018), with very few MM cases reporting a CCND3 overexpression due to the t(6;14) of the IgH enhancer (Chesi and Bergsagel, 2011). This has been confirmed in the present study (Figure 4-24), there are only 3 samples with relatively high CCND1 expression and higher CCND2, the cytogenetic information can be seen in the samples details table (MM_vs_PC_supervised_analysis/ATAC_and_RNA-seq_stats.xlsx): HD samples A19.1 (1q gain) and A26.9B (TP53 deletion, 1q gain) and the MMSET translocated sample A28.13. The high CCND1/2 expression for HD samples has been previously found in a

small proportion of HD samples (Kaiser et al., 2013). Despite the fact that gaining of 1q arm in HD samples is associated with high CCND2 (as reflected in the results) and silenced CCND1 expression (Shah et al., 2018), there is no cytogenetic information regarding gaining 1q for any of the samples (associated with CCND1 expression increase). This may be because no sample in the study has this feature or perhaps because it is not determined in the analysis, leaving the possibility open for the highlighted samples to have that feature. Out of these samples, A28.13 is a sample with the lowest CCND1 expression, on the borderline of a high CCND1 expression, but still having a significantly higher CCND2 expression.

Deregulation of different D-cyclins occurs through different mechanisms: CCND1 and CCND3 are known to be deregulated by the previous mechanisms proposed with the t(6;14) CCND3 affecting translocation occurring in less than 5% of all MM cases (Chesi and Bergsagel, 2011; Kaiser et al., 2013; Sarasquete et al., 2013; D. Smith et al., 2016). Different mechanisms have been suggested causing CCND2 abnormal expression, for example, it was proposed that at least two CCND2 mRNA isoforms are produced and, via alternative polyadenylation, the samples with CCND2 overexpression contain a greater abundance of one of the isoforms containing a shortened 3'UTR length that loses miRNA binding sites, this characteristic impairs miRNA repression of this gene (Misiewicz-Krzeminska et al., 2016). Interestingly, the study also found that this CCND2 mRNA shortening was significant when repressing CCND1 and the longer CCND2 isoform abundance was greater with CCND1 and CCND3 overexpression, postulating the role of CCND1 and CCND3 regulating CCND2 through alternative polyadenylation. Furthermore, it was found that the explained mechanism was unlikely to occur in CCND1 translocated t(11;14) MM cell lines with no detectable CCND2 isoform and in these cases, it was found that methylation of the CCND2 promoter was repressing its expression (Misiewicz-Krzeminska et al., 2016).

Downstream of the novel elucidated CCND2 enhancer regions found in the present work, a previously found different CCND2 super-enhancer provided another hypothesis for CCND2 overexpression (Young et al., 2013). This putative enhancer starting at position 4,247,853 in chromosome 12 (upstream of CCND2) and overlapping the CCND2 TSS was reported in a MM CL (Young et al., 2013), but it could be an extended CCND2 TTS and not an enhancer. The mechanism of CCND2 overexpression suggested in the present work refers to the novel CCND2 super-enhancer region. Suppression of one region within it (chr12:4,142,636-4,143,378) represses CCND2 expression in a MAF translocated cell line (to my knowledge JJN-3) to levels comparable as the ones repressing the CCND2 promoter (Alexia Katsarou, Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Foundation Trust in

London, *pers. comms*.). In primary samples, its accessibility is correlated with CCND2 gene expression and in a close B-cell line there is evidence showing interactions between this region and the CCND2 promoter. Together, these results propose regions within the CCND2 superenhancer (chr12:4,103,242-4,177,985) and particularly chr12:4,142,636-4,143,378 as a very strong candidate in regulating the CCND2 gene overexpression, complementing the already known converging cyclin-deregulation program found in virtually all MM samples. At the time of writing, the chromatin state segmentations (Albrecht et al., 2016) and ChromHMM (Ernst and Kellis, 2017) are not showing active enhancer states for these regions in the 4 primary MM samples available (not shown), but this may be due to not having any MAF translocated samples among them. This evidence points at these putative enhancers (particularly chr12:4,142,636-4,143,378) becoming active in MAF translocated samples and other samples from other subgroups and regulating the CCND2 gene.

TF enrichment on these candidate enhancers unveils a possible regulatory program where TFs which are already found to be relevant in the literature for PC and MM and are highly enough expressed already in both the MM subgroups and PCs might be binding to the enhancers and activating them. The MMSET/NSD2 TF motif was not an available TF motif in the database used, interestingly, the MAF subgroup was not found significantly enriched with the MAF TF at specific candidate enhancers. Perhaps because on the whole, the 662 MAF enhancers were not generally enriched for MAF motifs (while maybe a fraction of them might be). It is also possible that the MAF TF is binding preferentially to promoters in the activated genes of the MAF subgroup instead of enhancers, as it is the case of the CCND2 gene which contains a MAF binding motif in the promoter and MAF can directly transactivate a CCND2 promoter construct in transient transfection assays (Hurt et al., 2004), although it may also bind to a CCND2 enhancer. Together these results help to uncover MM subgroup biology with potential druggable targets.

Previous attempts have endeavoured to stratify cases of MM based on enhancer action, studying the possible oncogenesis generating mechanism. As seen in chapter 3, the ones most related to this work are the previously mentioned study (see chapter 1, 2 and 4) studying gene regulation with *in vitro* differentiated memory B cells used as reference PCs (Jin et al., 2018). This study found 20,000 regions with differential histone acetylation between PC and MM (a proxy for enhancers) but to my knowledge the enhancer list is not public. Also, to my knowledge, there is only one study which has provided MM subgroup-related specific interactions (albeit only 2) which is previously mentioned (Hoang et al., 2018).

5. Chapter 5: Multi Omics Factor Analysis (MOFA)

5.1. Acronyms and Abbreviations used in the chapter

Acronym	Definition
AID	Activation-Induced Cytidine Deaminase
AML	Acute Myeloid Leukaemia
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
ATP	Adenosine triphosphate
BET	Bromodomain and Extraterminal Domain
BM	Bone Marrow
BMP	Bone Morphogenetic Protein
bp	Base pair
ChIP-seq	Chromatin Immunoprecipitation sequencing
CL	Cell Line
CRISPRi	CRISPR interference
DA	Differentially Accessible
DAMM	Differentially Accessible MM
DASMM	Differentially Accessible Subgroup MM
DE	Differentially expressed
DEMM	Differentially Expressed MM
DESMM	Differentially Expressed Subgroup MM
eRNA	Enhancer RNA
FDR	False Discovery Rate
H3K27ac	Acetylation of histone H3 lysine 27
HD	Hyperdiploid
HDACi	Histone Deacetylase inhibitors
IgH	Immunoglobulin Heavy Chain
Kb	Kilobase
LF	Latent Factor
lnc-RNA	Long non-coding RNA
log_2foldchange	Log (base 2) fold change
Mb	Megabase
MGUS	Monoclonal Gammopathy of Uncertain Significance
miRNA	Micro RNA
MM	Multiple Myeloma
MMPC	Multiple Myeloma and Plasma Cell
MOFA	Multi Omics Factor Analysis
mRNA	Messenger RNA
ncRNA	Non-coding RNA
ND	Normal Donor
OE	Overexpressed
OEMM	Over Expressed MM
OESMM	Over Expressed Subgroup MM
PC	Plasma Cell (used interchangeably with ND)

PCA	Principal Component Analysis
PCL	Plasma Cell Leukemia
QC	Quality control
rLog	Regularized Log
RNA-seq	RNA sequencing
SMM	Subgroup MM
snoRNA	Small nucleolar RNA
SNV	Single Nucleotide Variant
TF	Transcription Factor
TSS	Transcription Start Site
UTR	Untranslated Region
VDJ	V-D-J block recombination (PC formation)
WES	Whole-Exome Sequencing
WGS	Whole-Genome Sequencing

5.2. Introduction

As mentioned in the introductory section in chapter 4, multiple MM sub classification systems have been previously used. Some studies have implemented an unsupervised approach to then relate the findings to the cytogenetic MM subtypes known at the time, for example, in a study of PC dyscrasias (diseases): 7 Monoclonal Gammopathy of Undetermined Significance (MGUS), 39 MM and 6 Plasma Cell Leukemia (PCL) patients were profiled (Mattioli et al., 2005) in terms of the RNA expression (microarray). The study found the unsupervised classification of samples tended to create subgroups corresponding to the IgH translocations (despite not creating condition specific groupings). Later on, in a microarray study containing 532, also newly diagnosed, patients treated with 2 different therapeutic protocols , it was found that a profile of 70 genes (with 30% of these belonging to chromosome 1), were associated with different prognostic outcomes (Shaughnessy et al., 2007). This was then reduced to 17 genes, which performed equally well.

Small nucleolar RNAs (snoRNAs) have also been used to sub classify MM. snoRNAs are a type of small non-coding RNAs molecules involved in the modification and processing of ribosomal RNA. More recently it was found that snoRNAs have commonalities with miRNAs in terms of processing pathways, and in fact some snoRNAs can be processed to release miRNA like molecules (Scott and Ono, 2011). For example, in a study involving 55 MM, 8 secondary plasma cell leukaemia samples and 4 PC donors from tonsils (not bone marrow), they were found to be down regulated in disease in general and did not subgroup patients in terms of translocation partner and HD status (Ronchetti et al., 2012). LncRNAs have been used to stratify and determine prognosis in MM patients leading to a four lncRNA signature (Zhou et

al., 2015). Furthermore, using the 500 most variably expressed lncRNA, in a cohort of 30 MM patients (Ronchetti et al., 2018) the same subgroups used in Chapter 4 were recapitulated. Finally, miRNAs have also been used for this purpose. In a study by (Liu et al., 2019) correlation analysis was performed to obtain miRNA-mRNA relationships in MMSET-translocated patients. In another study with 33 MM samples, 5 MGUS and 9 PC, 109 miRNAs were found to be DE in MM compared to normal with most of them being up regulated in MM (Chi et al., 2011). Interestingly, Chi et al. also found DE miRNAs associated with MGUS, with more than half of them overlapping with deregulated miRNAs in MM. Moreover, 26 of the 33 MM samples were cytogenetically annotated and certain miRNAs were aberrantly expressed in the following individual subgroups CCND1 t(11;14), MMSET t(4;14) and also in a group of samples harboring del(13q) (Chi et al., 2011), which, among other events, deletes negative cell cycle regulators and tumor-suppressor genes, an event associated with worse prognosis (Chavan et al., 2017).

More recently, using single cell RNA-seq, disease progression has been studied in bone marrow MM 597 PCs from 15 patients at different MM disease stages: MGUS, smoldering MM, newly diagnosed MM and relapse/refractory MM (Jang et al., 2019). In this study, classification based on 790 most variable expressed genes placed the cells from each patient in 2 or more groups.

Combination of different data types has been performed in MM before, for example to assess cell line resistance changes with regards to accessibility, DNA methylation and gene expression (Dimopoulos et al., 2018).

Also, in a study, combination of whole genome, exome sequencing and RNA-seq was used first to determine translocation breakpoints and clustering of samples according to their gene expression (Barwick et al., 2019). Furthermore, chromatin accessibility and ChIP-seq were used to determine a superenhancer near the immunoglobulin lambda (IgL) locus (Barwick et al., 2019). A combination of ChIP-seq and ATAC-seq was used to show that the E2F-DP1 dimer acting as a TF and its effect was increased promoter chromatin accessibility at E2F-DP1 bound promoters (Fulciniti et al., 2018). Furthermore, Fulciniti et al. also used RNA-seq data to confirm the relationship between E2F1 expression and expression from E2F-DP1 bound promoters, implying regulation. RNA-seq data was also used to assess prognosis in terms of expression of E2F.

Hoang and colleagues combined whole-exome sequencing (WES), whole-genome sequencing (WGS) and matched RNA-seq in MM samples with publicly accessible promoter Capture Hi-C, ATAC-seq and regulatory histone modifications generated on naïve B-cells, in order to determine promoters and associated cis-regulatory elements (potential enhancers) and their

relationship to recurrent mutations affecting expression (Hoang et al., 2018). This study however, found few interactions enriched in 1%-6% of the total samples considered in the differential analysis and only two interactions correlated with MM subgroup incidence (with less than 5% presence in HD and MYC translocated subtype samples). Thereby lacking a comprehensive view of the MM subgroup enhancer – promoter interactions.

Finally, ATAC-seq and gene expression data were combined in the previously mentioned (see chapter 1, 2 and 4) study studying gene regulation with *in vitro* differentiated memory B cells used as reference PCs (Jin et al., 2018).

Depending on the data, unsupervised analyses can be a very valuable tool to elucidate novel biology. To my knowledge, this is the first exhaustive unsupervised coordinated analysis involving chromatin accessibility and gene expression data comparing PCs to their malignant Myeloma state, including cytogenetic subgrouping.

5.2.1. Chapter Aims

The software package MOFA (Argelaguet et al., 2018) was used in combination with chromatin accessibility and gene expression data in an unsupervised way with the aim of classifying MM patients. This resulted in identification of meaningful axes separating samples. This procedure intended to answer questions such as what sample separations, each of the resulting axes was identifying and whether they related to previously known subgroups based on cytogenetic or other biological information. Regions and genes were identified and their contribution to each separation was quantified, for the resulting genes, they were related to previously known genes (whether from the supervised analysis in chapter 4 or the literature). Furthermore, regions and genes were related to identify important interactions guiding each resulting axis.

Further technical questions regarding the resulting MOFA model were also investigated. For example, to what extent adding ATAC-seq to RNA-seq data aid in the PC vs. MM and its subgroups classification and whether the model generated is robust to the number of input features used and within the input features used to starting seeds. Also, I set to determine whether the input features selected in an unsupervised way were meaningful (for example genes with a minimum expression) and if any correlations existed (for example in terms of library sizes).

5.3. Results

5.3.1. MOFA separates samples into their constituent subgroups

The MOFA model (Argelaguet et al., 2018) was run using as inputs the normalized paired ATAC-seq (reads in peaks) and RNA-seq data (Figure 5-1) from the same 33 primary MM and PC samples (Table 2-1) as in chapter 3, including the primary MM samples with no cytogenetic information as specified in the Materials and Methods chapter. Since there is no gender information about the samples and to eliminate that source of biological variability, gender chromosomes and genes lying in them were removed from the ATAC-seq counts in consensus chromatin accessible peaks (for primary PC and MM) and RNA-seq gene counts respectively. Moreover, to obtain an enhancer – regulated genes based classification, TSS were removed from the accessible regions too. From the remaining, the top 5,000 regions and genes based on variability are inputted to MOFA.

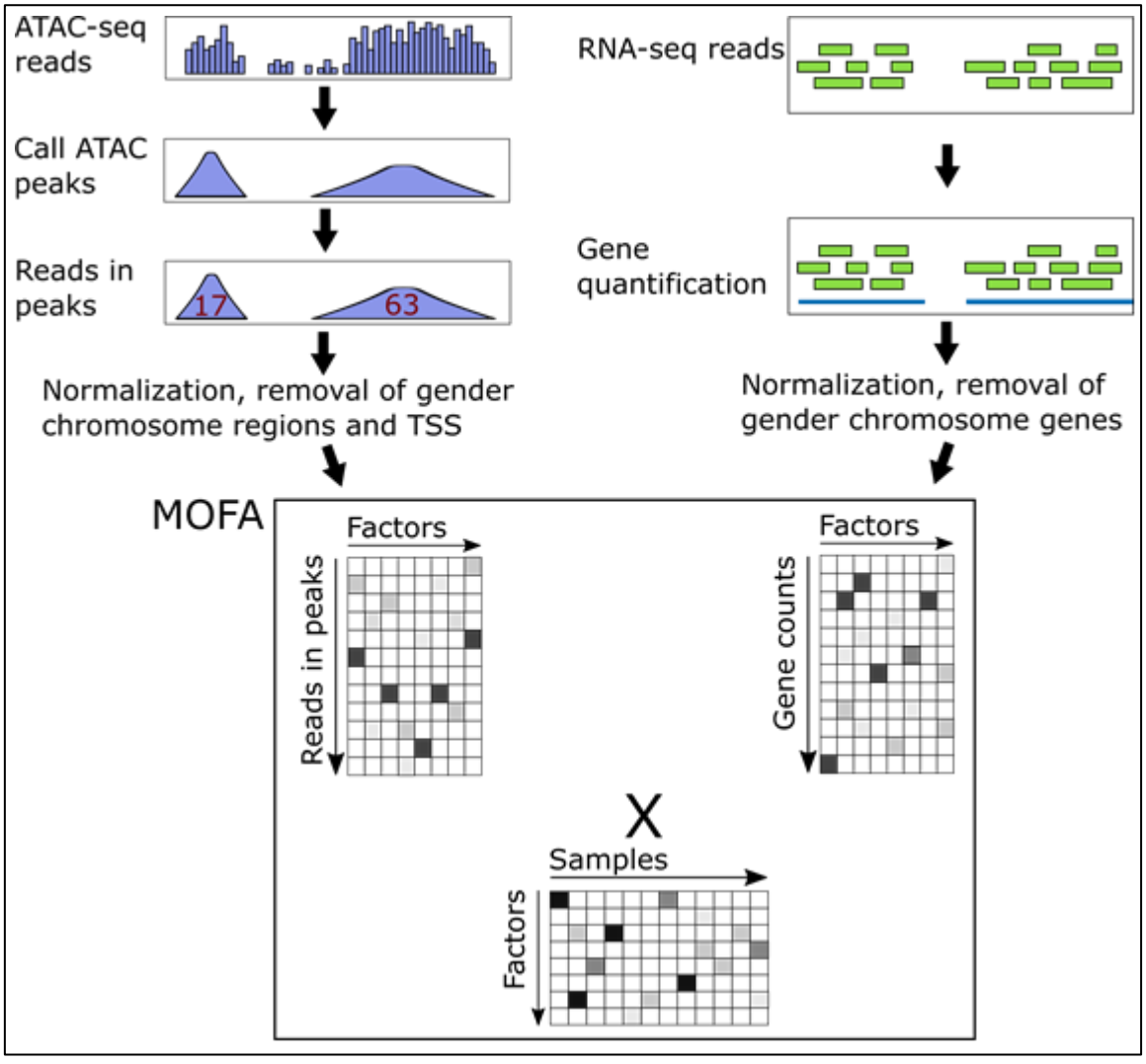


Figure 5-1: MOFA schematic.

ATAC-seq reads in consensus peaks and reads in genes are normalized. Regions corresponding to gender chromosome regions and overlapping annotated and unannotated TSS of more than 1 exon are not considered, genes having TSS in gender chromosomes are also disregarded. The top 5,000 regions and genes are inputted to MOFA based on variability. MOFA creates 17 Latent Factors (LFs) and divides the overall variability into these dimensions.

Briefly, MOFA is similar to PCA, in that it creates a framework to explain the data in an unsupervised way (subgroup classification information is not used *a priori*) based on variability using 17 Latent Factors (LFs) (in this case). Importantly MOFA trains a model with different data types (in this case chromatin accessibility and gene expression) simultaneously and each LF can be thought of as a dimension where each of the ATAC-seq and RNA-seq features contributes in a particular way.

The LF – sample weights resulting from the chosen model can be seen in:

MOFA/samples_LFs_weights_matrix.tsv.gz

The factor weights for RNA-seq and ATAC-seq respectively can be seen in:

MOFA/RNA_weights.tsv.gz

MOFA/ATAC_weights.tsv.gz

To verify that the model used was robust, 5 runs with different seeds were used and the one with the best data fit was selected, all models showed very similar grouping of samples and little correlation between LFs. To validate that the results were independent from sample library sizes, no significant correlations were observed (not shown) between sample LF weights and the ATAC-seq final number of single ends inputted to the peak caller, ATAC-seq reads in peaks or RNA-seq mapped read pairs. Likewise, the influence of the number of input features were accounted for in the model by training a model with the top 10,000 variable ATAC and RNA features producing a very similar model to the one with 5,000 (not shown).

In addition, it was important to determine whether the features used in the model were biologically meaningful (for example, genes that were expressed to some degree in a subset of the samples). Particularly, the mean signal across all samples of the chromatin accessible features included in the model was compared to that of the MMPC enhancers obtained in the supervised analysis in Chapter 3 in the context of the LF1 separation (Figure 5-2 A). As can be seen, the feature selection based solely on variability choses variables with similar minimum means as in the supervised analysis (which uses robust tools), suggesting variability is not high

purely because the signal is low and noisy. An analogous check was performed in terms of gene expression (Figure 5-2 B), lower gene expression averages than the supervised MM vs. PC DEMM genes were found, however, the LF weights for these genes drastically tend towards a 0 weight (meaning they are not meaningful to the model in this separation).

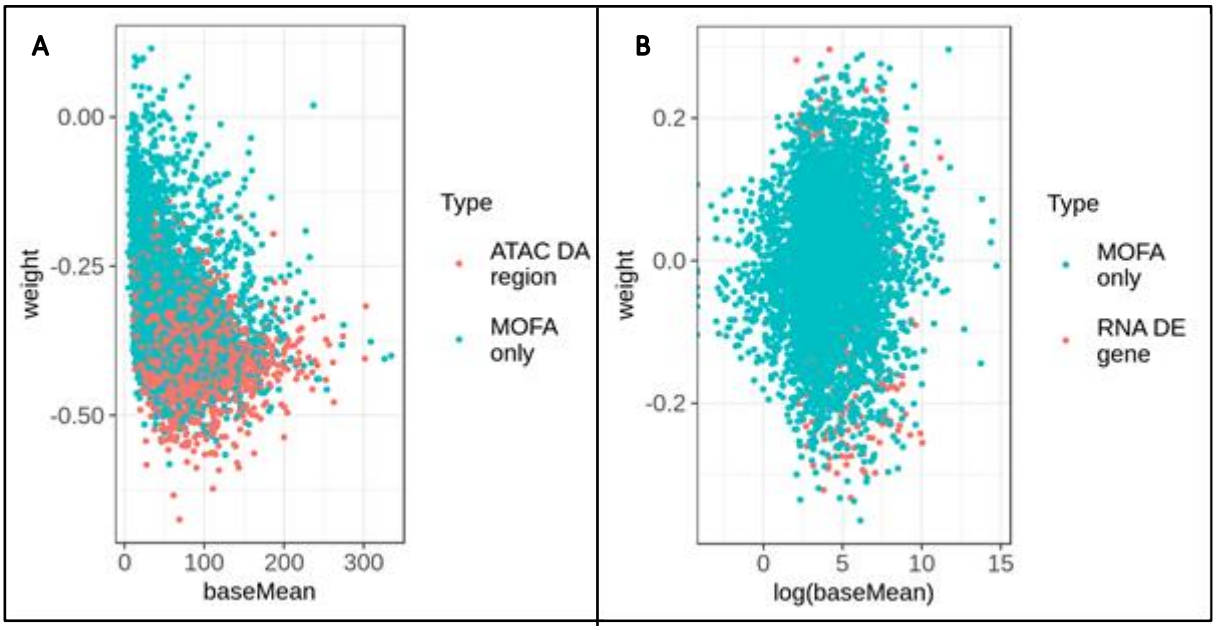


Figure 5-2: Quality control metrics from the MOFA model features.

A: Mean of normalized ATAC-seq counts for all samples in MOFA with MOFA LF1 ATAC-seq weights for MOFA regions intersecting MMPC enhancers from the supervised analysis (labelled "ATAC DA region") and "MOFA only" regions. B: Mean of the normalized gene counts for MOFA gene features corresponding to DEMM genes (labelled "RNA DE gene") or non-DEMM genes.

Despite the fact that the model is unsupervised, the different LFs capture the variance in such a way that the samples are split into the different subgroups (Figure 5-3). This occurs more profoundly when simultaneously using chromatin accessibility and gene expression data (Figure 5-3 left column), LF2 completely separates PC from MM, with LF1 creating a similar separation. Furthermore, LF3 parts the MMSET samples from the rest and LF5 creates an axis dividing the samples into MAF and CCND1 at each of the extremes with the rest of the samples in between. The variance explained plot (Figure 5-4 A) shows how LF3 and LF5 have an even variance explained profile between accessibility and gene expression, and LF4 (as it will be seen later) separates a HD outlier sample exclusively based on its accessibility profile. LF1 and LF2, despite generating a matching healthy – cancer separation exhibit a different nature in

terms of the source of variation captured. LF1 is chromatin accessibility dominant, explaining more than half of all the data variance (while having a significant gene expression influence) and LF2 is practically solely gene expression based. LF1 and LF2 are correlated (Figure 5-4 B). For this reason multiple different model executions were performed. Despite this, these LFs always remained independent. This is to be expected because the ordering of MM samples is different for each LF, even pointing at an anti-correlation between LF1 and LF2 MM sample weights. As Figure 5-4 B shows, the only other significant correlation is between LF14 and LF8, since the variance explained at this point is low and no relationship with any of the covariates studied was seen (CD19 status, PC donor id, subgroup or condition), these LFs were not considered.

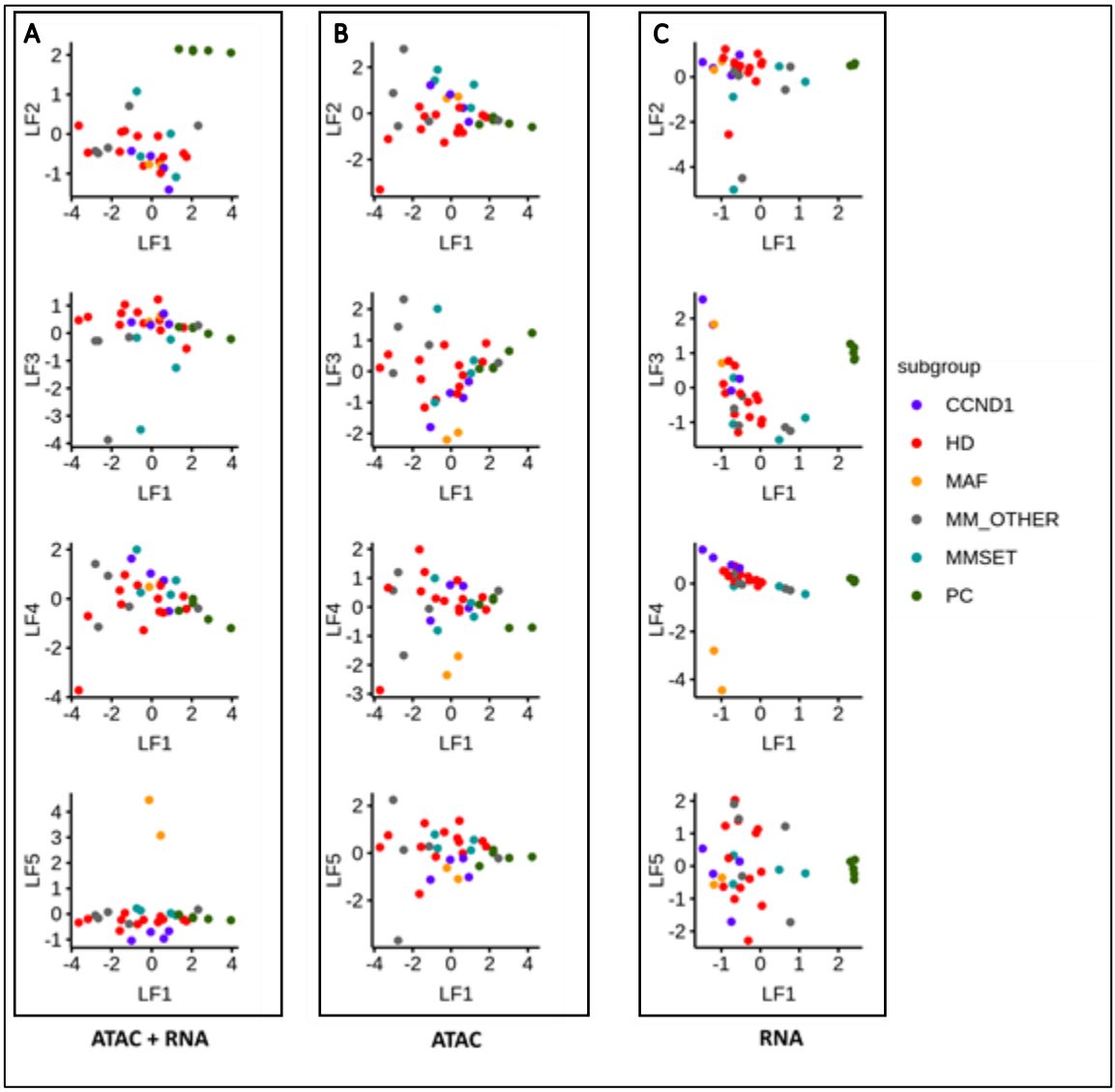


Figure 5-3: Separation of MM subgroups and PC samples by MOFA for each LF with different feature inputs.

Separation of samples by the LFs of the MOFA model sample – LF weights with different inputs. Left column: 5,000 top variable ATAC and RNA features. Middle column: 5,000 top variable ATAC features only. Right column: 5,000 top variable RNA features only.

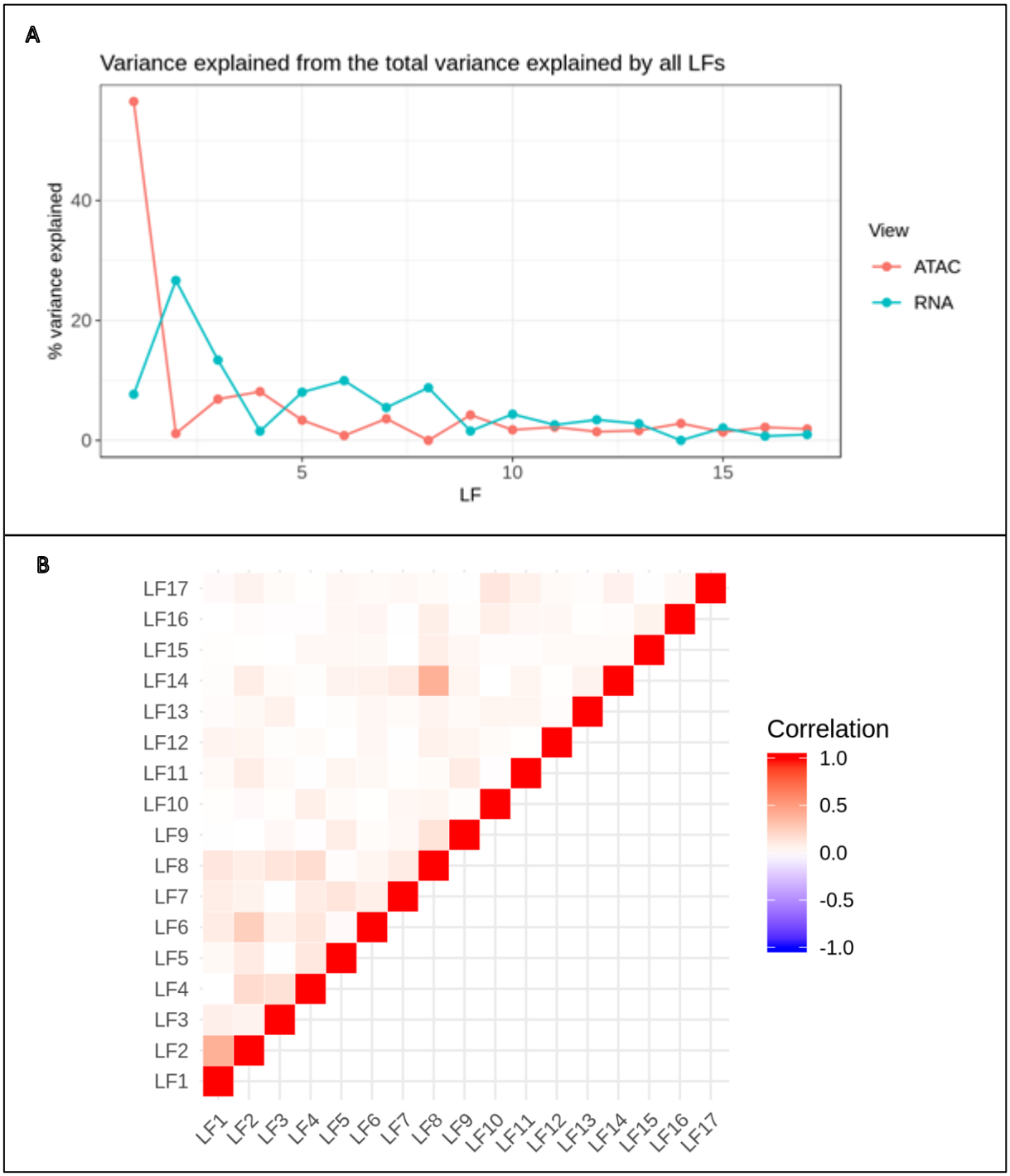


Figure 5-4: Quality control metrics from the MOFA model.

A: Percentage variance explained by each dimension (LF) for each data type: ATAC-seq and RNA-seq with each LF. B: MOFA LF correlation plots (LF – sample).

These results point at a genuine subgroup profile when taking into account chromatin accessibility and gene expression data. It is important to determine which features are creating this effect and relate them to the supervised MM vs. PC (Chapter 3) and MM subgroup vs. PC (Chapter 4) analysis.

As Table 5-1 shows, none of the regions used in the MOFA analysis (i.e. the 5,000 most variable regions) are more accessible in PC than MM by the pan-myeloma analysis in Chapter 3. The gene expression features in MOFA overlapped by the supervised analysis are more balanced in terms of being OE in PC compared with MM or MM subgroups. In some cases, as with the CCND1, HD and MAF subgroups, the number of overlapping genes that are OE in PC is larger than in the specific MM subgroup.

	Unsupervised analysis				
Supervised analysis			**Dominant accessibility/expression**	**MOFA regions overlapped**	**MOFA genes overlapped**
	MM vs. PC DE/DA		**MM**	1,912	237
			PC	0	148
	Subgroup MM vs. PC DE/DA	**CCND1**	**CCND1**	590	277
			PC	13	782
		HD	**HD**	904	398
			PC	0	636
		MAF	**MAF**	668	465
			PC	48	692
		MMSET	**MMSET**	647	645
			PC	15	485

Table 5-1: Genes and regions in MOFA overlapped by supervised analysis DE genes and regions.

Columns: "Dominant accessibility/expression" the count for MM, MM subgroup or PC elements with higher accessibility (column "MOFA regions overlapped") or higher gene expression (column "MOFA genes overlapped") for each pairwise comparison.

"MM vs. PC DE/DA": The supervised analysis MM vs. PC (Chapter 3) MMPC enhancers and DEMM genes.

"Subgroup MM vs. PC DE/DA": The MM subgroup vs. PC analysis (Chapter 4) DASMM enhancers (DA) and DESMM genes.For the overlap in DASMM enhancers, each MOFA region inputted is only overlapped by either one SMMPC enhancer or none. In cases where they are overlapped by multiple DASMM enhancers, the one with the largest overlap is selected.

5.3.2. Paired accessibility and expression data is more optimal classifying samples

After examining the overall results, it is important to identify if using ATAC-seq and RNA-seq data simultaneously with MOFA provides benefits in terms of the profiling of samples compared with using either data source alone.

As Figure 5-3 shows, the MOFA algorithm using accessibility and gene expression data (Figure 5-3 left column) outperforms using either data source by itself (Figure 5-3 middle and right columns) in the task of correctly classifying samples. The model using ATAC-seq data by itself only creates distinguishing axis for MAF samples (LF3) and MM vs. PC samples (LF1). The use of RNA-seq data is more effective than ATAC-seq, creating a MM vs. PC separation for LF1 and a MAF - CCND1 axis in LF4.

To quantify the difference in the ability of the various data types to separate subtypes, silhouette scores, which measure the coherence of clusters in a multi-dimensional space, were calculated for each class in each model (Figure 5-5 A). Chromatin accessibility in conjunction with gene expression data (Figure 5-5 A, pink circles) creates more distinct subgroup clusters than gene expression by itself (Figure 5-5 A, cyan circles), while gene expression alone classifies MM vs. PC better. Mirroring the LF plots (Figure 5-3), the MMSET, HD and CCND1 samples are better classified within the cluster when using a combination of accessibility and expression data. Cytogenetically unannotated samples are more similar to other clusters when using a combination of data sources, this is to be expected, consistent with the fact that this group of samples contains samples belonging to other cytogenetic groups. MAF is regarded as slightly more similar to other clusters when using the combination of data types compared with using only expression data. This however rectifies when considering only LF1-LF5, which are the LFs found to be most relevant to this analysis and explaining the most variability in the data (Figure 5-5 B): LF1 explains 57% total ATAC-seq and 8% RNA-seq variability, LF2 1% and 27%, LF3 7% and 13%, LF4 8% and 2%, LF5 3% and 8% respectively. Interestingly, when regarding only these dimensions, the only subgroup that is classified more optimally with gene expression data alone is HD.

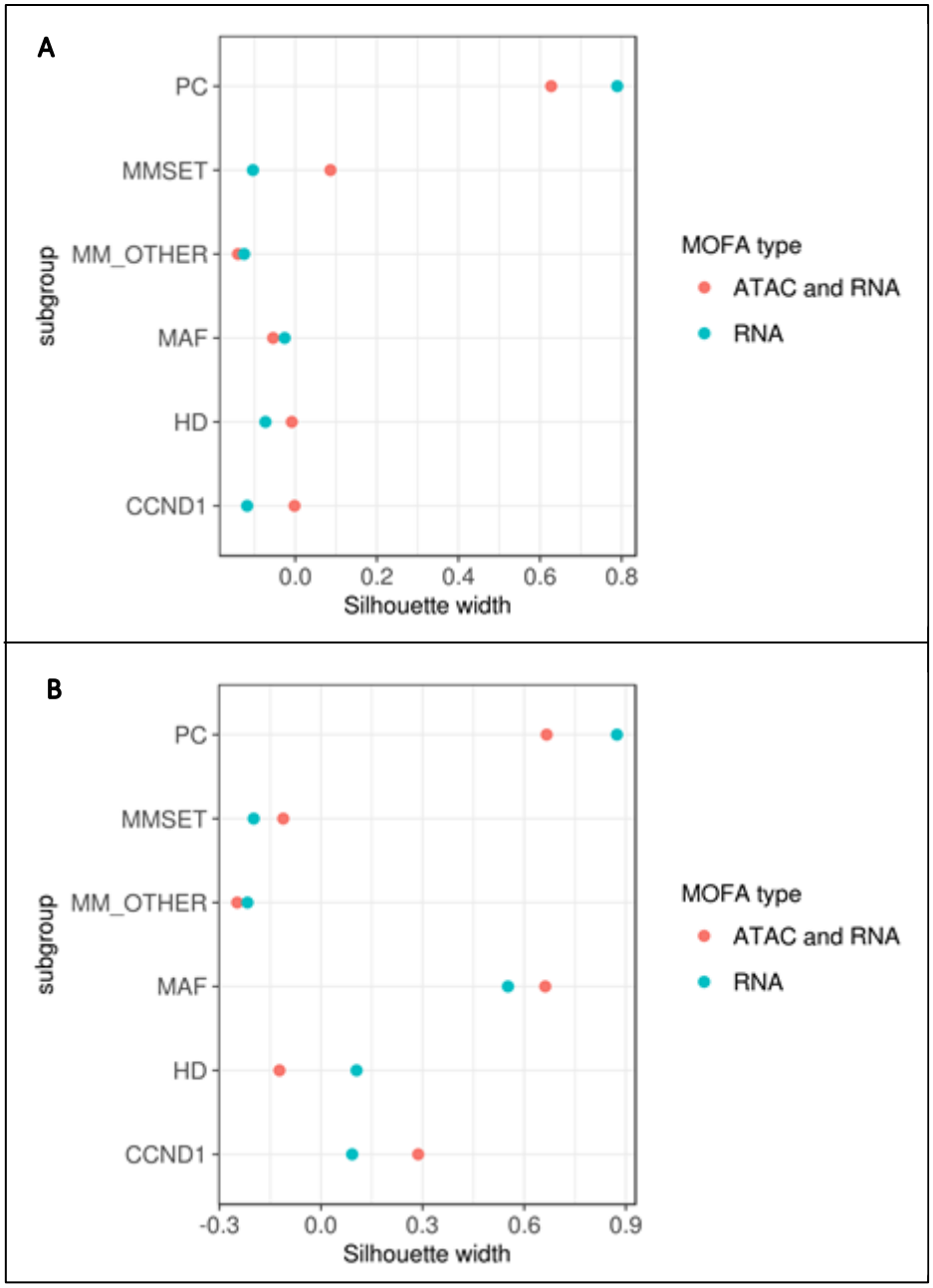


Figure 5-5: Average silhouette width per subgroup.

Average per subgroup silhouette comparison after assigning each sample to its corresponding subgroup and calculating the Euclidean distance between samples, comparison between MOFA models comprising of ATAC and RNA input features and RNA input features only. A: Using all samples LFs 1-17 factors as distance dimensions. B: Using only LF 1-5 factors as distance dimensions. The silhouette score for each sample goes from -1 (very similar to another cluster) to 1 (very similar to own cluster). MM_OTHER (cytogenetically unannotated samples).

5.3.3. LF1 and LF2 create MM vs. PC separation

Now that it has been established that simultaneous integration of accessibility and gene expression data provides an advantage in terms of classification of samples, the LFs creating

meaningful separations can be studied in detail, relating them to the supervised analyses in Chapter 3 and 4 and integrating the coordinated importance for features of both data types provided by MOFA. To cover the different categories of genes and enhancer – gene interactions, throughout this chapter, examples are provided containing genes with or without associations in the literature with MM and PC, in cases where a significant association between gene and MM disease is found (Piñero et al., 2020), an association score is attached. Additionally, it is indicated whether a gene is previously found to be DE or OE in MM (or a MM subgroup) or in PC, equivalently on interactions, whether an enhancer is obtained in the supervised analysis for each LF. Futhermore, for completeness, examples are also selected according to whether the gene or region weights are extreme for a given latent factor.

As was observed in Figure 5-3, LF1 establishes approximately PC vs. MM sample separation with positive LF1 weights corresponding to PC samples (Figure 5-6 A and B). This LF explains nearly half of the accessibility variability (coherently with having around 40% of MOFA features overlapping candidate MMPC enhancers as can be seen in Table 5-1) and a significant fraction of that for the gene expression (Figure 5-4 A).

The table with the MOFA ATAC-seq features containing supervised analysis MM vs. PC information can be found in:

MOFA/MOFA_ATAC_MM_vs_PC_details.gz

The very predominant negative weights in the LF1 ATAC-seq features (Figure 5-7 A and B) reflect the general opening of MM chromatin vs. PC (Figure 5-6 A) as can also be seen by the negative correlation of MM vs. PC openness change (but not general openness, not shown) with increasing LF1 weights (Figure 5-7 A). This correlation is only significant when taking into account all MOFA features (which include regions that are equally or more open in the healthy compared with the cancer state) and not when using only overlapping MM vs. PC enhancer regions from the supervised analysis. Furthermore, all the MOFA features found to overlap differential features in the MM vs. PC analysis in Chapter 3 are more accessible in cancer (Table 5-1) and have an enrichment for negative weights compared with all MOFA features (Figure 5-7 B). Finally, it can be observed how 146/200 regions with the most negative LF1 weight are also DAMM regions more accessible in MM. Pointing at a likely enhancer signature overlapping both analyses.

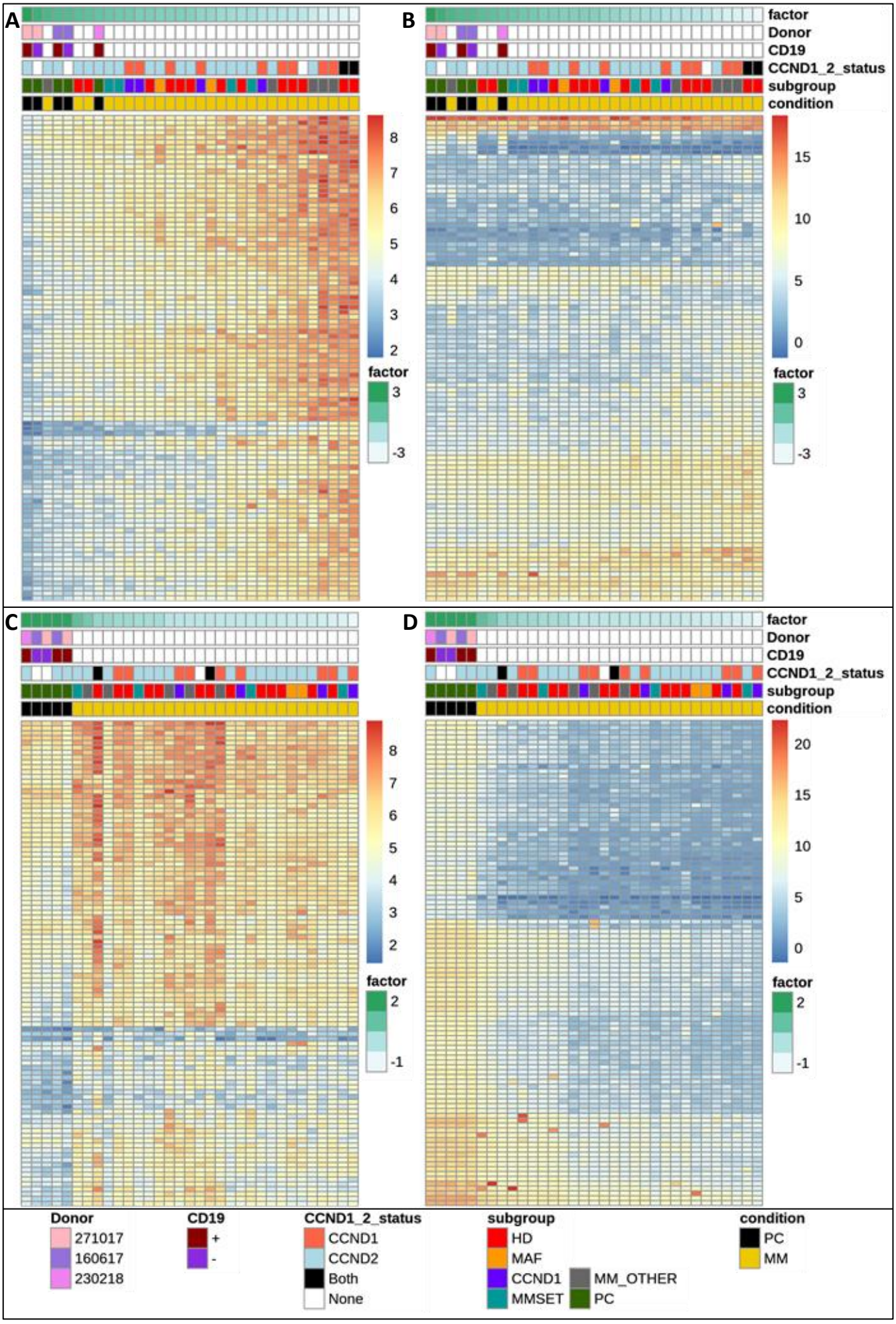


Figure 5-6: Heatmaps for the top 100 ATAC and RNA features (absolute factor loading) for LF1 and LF2.

rLog normalized feature counts heatmaps with the different sample covariates, heatmap scale and factor scale shown to the right of each plot. Rows hierarchically clustered by correlation metric and complete linkage. A: LF1 with ATAC-seq features, B: LF1 with RNA-seq features, C: LF2 with ATAC-seq features, D: LF2 with RNA-seq features.

Donor: PC donor id. CD19 status: CD19 receptor status for the PC donor. CCND1_2_status: sample CCND1 and CCND2 normalized log expression positive if larger or equal to 11. Subgroup: Plasma Cells (PC), Hyperdiploid (HD), MM_OTHER: cytogenetically unannotated samples, primary IgH translocation driving event: MAF, CCND1, and MMSET.

Together these results show that this LF separates a change in chromatin openness between conditions (not correlated with absolute accessibility) and the MM and PC candidate enhancer regions from Chapter 3 are very relevant to this separation. MOFA, however, not only recapitulates a portion of these regions but also incorporates new genomic areas that are more accessible in PCs and are meaningful to this separation.

The table with the MOFA RNA-seq features containing supervised analysis MM vs. PC information can be found in:

MOFA/MOFA_RNA_MM_vs_PC_details_association_score.gz

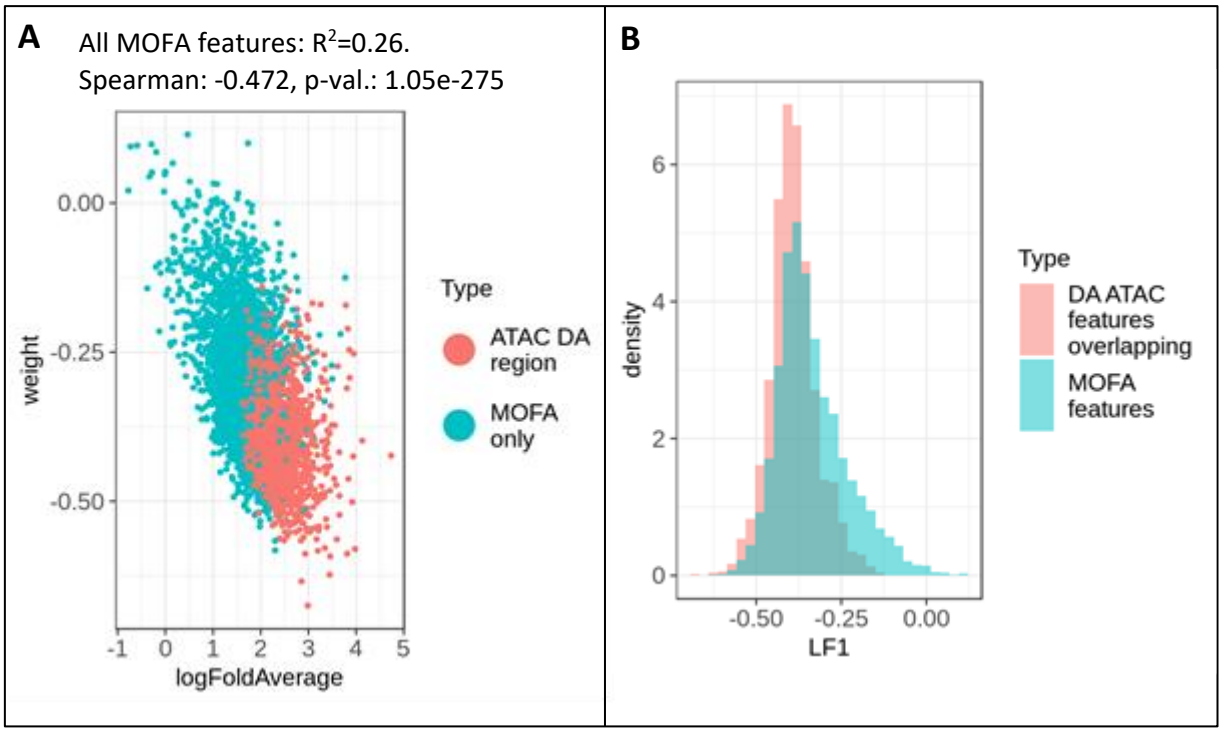


Figure 5-7: Chromatin accessibility MOFA features for LF1.

LF1 weight loadings for MOFA chromatin accessibility.

A: correlation with MM vs. PC log_2FoldChange average for the MM vs. PC analysis in Chapter 3 (MM vs. average PC CD19, MM vs. PC donor average and MM vs. PC accounting for batch), features with distinction for MMPC enhancer features (labelled "ATAC DA region") overlapping and non-overlapping "MOFA only" features (Chapter 3). Correlation metrics taking into account all MOFA features.

B: Weight density distributions, distinction between MMPC enhancer (labelled "DA ATAC features overlapping") features overlapping MOFA features and all MOFA features (labelled "MOFA features").

The gene expression feature weights for LF1 are evenly distributed (Figure 5-6 B and Figure 5-8 A and B) and negatively correlated with MM vs. PC fold expression change only when taking into account all MOFA features (Figure 5-8 A). There is no relationship between the LF weights and overall gene expression (not shown). The features that overlap DEMM (Chapter 3) consist of over and under expressed genes (Table 5-1) with an enrichment on the corresponding extreme weight ends when compared to gene expression for all features using in MOFA (Figure 5-8 B).

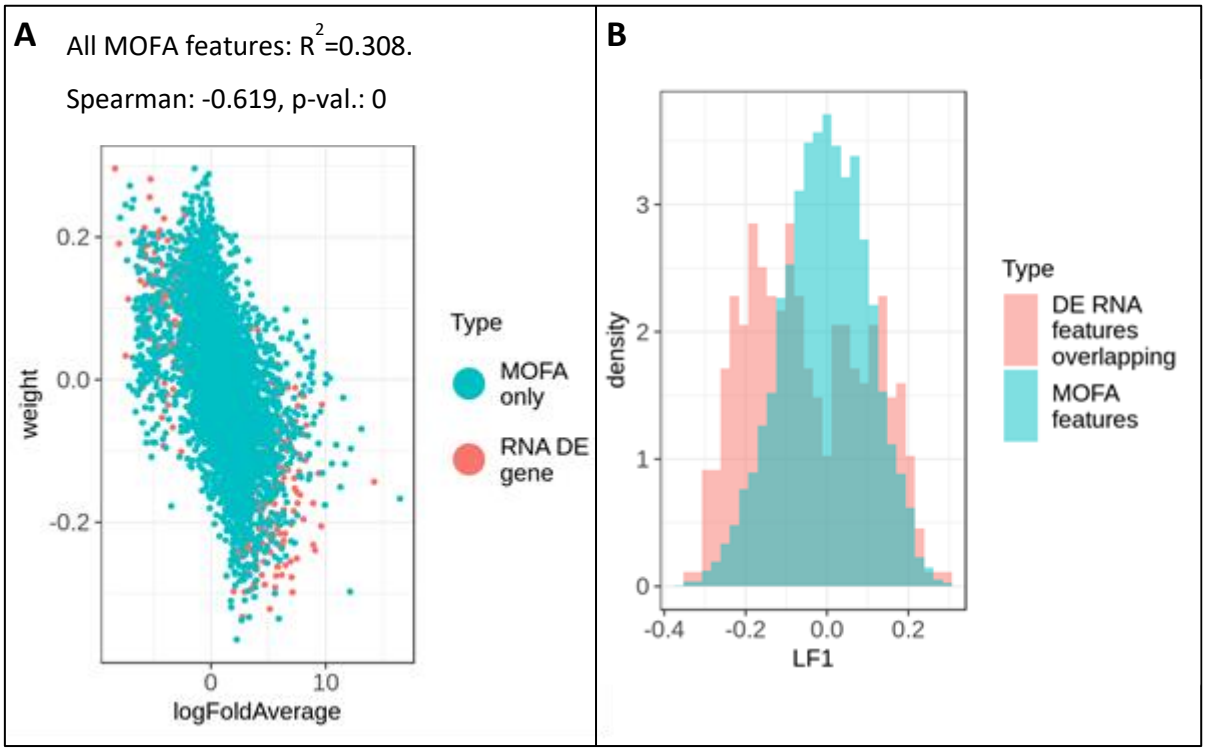


Figure 5-8: Gene expression MOFA features for LF1.

LF1 weight loadings for MOFA gene expression.

A: correlation with MM vs. PC log_2FoldChange average for the MM vs. PC analysis in Chapter 3 (MM vs. average PC CD19, MM vs. PC donor average and MM vs. PC accounting for batch), features with distinction for DEMM genes

(labelled "RNA DE gene") overlapping and non-overlapping MOFA features (labelled "MOFA only"). Correlation metrics taking into account all MOFA features.

B: Weight density distributions, distinction between DEMM genes features overlapping MOFA features (labelled "DE RNA features overlapping") and all MOFA features (labelled "MOFA features").

Analogously to chromatin accessibility features, LF1 splits conditions based on a change in gene expression occurring during Myelomagenesis with already studied and novel genes. This LF axis could be a consequence of the cancer state creating a general chromatin opening (or further opening already PC accessible regions) which is not always associated with general overexpression.

Gene Ontology analysis was carried out on the LF1 and LF2 top 10% genes by extreme negative weight (see Gene Ontology analysis on top MOFA LF1 and LF2 MM activated genes, section 2.9.5). The results can be seen in:

MOFA/LF1_LF2_top_500_neg_weights_GO.xlsx

For LF1, there is enrichment in 88 ontology categories, some such as type I interferon signalling and positive regulation of interleukin-8 production involved in immunity or PC-related like osteoblast differentiation and BMP signaling pathway. Others which may be cancer related such as negative regulation of apoptotic signaling pathway and positive regulation of proteolysis; involving gene regulation: gene silencing by miRNA, posttranscriptional gene silencing by RNA, mRNA splicing via spliceosome, posttranscriptional regulation of gene expression; epigenetic and chromatin remodelling categories: DNA conformation change, DNA packaging, epigenetic regulation of gene expression, chromatin assembly or disassembly, nucleosome organization.

The next step is to take advantage of the integrated accessibility and expression data, for this the MOFA ATAC-seq features that are within 1Mb of the promoters of MOFA genes are computed, resulting in 15,903 interactions per LF. The table containing the 15,903 interactions for each LF, MM vs. PC and subgroup MM vs. PC accessibility and expression log_2FoldChanges can be found in:

MOFA/MOFA_all_LFs_ATAC_1Mb_RNA_promoters_MM_vs_PC_and_subgroup_MM_vs_PC_details.tsv.gz

These interactions are analysed with emphasis on extreme negative LF1 weights (signifying high importance in the MM vs. PC separation produced by MOFA and generally MM more

open chromatin and gene overexpression compared with the healthy state) in Figure 5-9. 173 of these interactions recapitulate some of the 311 pairs of MM enhancers near OEMM protein coding genes found in Chapter 3. One such example with association to MM, with gene MM association score of 0.01 (Piñero et al., 2020), is the region chr5:137,739,042-137,739,301 interacting with CDC25C (marked as "chr5:137,739,042-137,739,301 CDC25C", coloured magenta in Figure 5-9), a gene found to be up regulated in a sub-population of cells within MM cell lines (Nara et al., 2013). Another gene previously found in Chapter 3 to be OE and regulated by candidate MM enhancers chr8:106,524,326-106,525,957 and chr8:107,457,615-107,457,954 is ANGPT1 (gene disease association score of 0.03), which was discovered to be up regulated in MM (Munshi et al., 2004). Three additional putative MM enhancers are found in the MOFA analysis that were disregarded by the supervised analysis for having a MM vs. PC log_2foldChange slightly below the used threshold: chr8:106,745,254-106,745,782, chr8:107,382,945-107,383,638 and chr8:106,597,737-106,598,615 (this last one annotated in Figure 5-9 as "chr8:106,597,737-106,598,615 ANGPT1" in grey), these regions and gene were found to have the corresponding negative weights in the MOFA analysis.

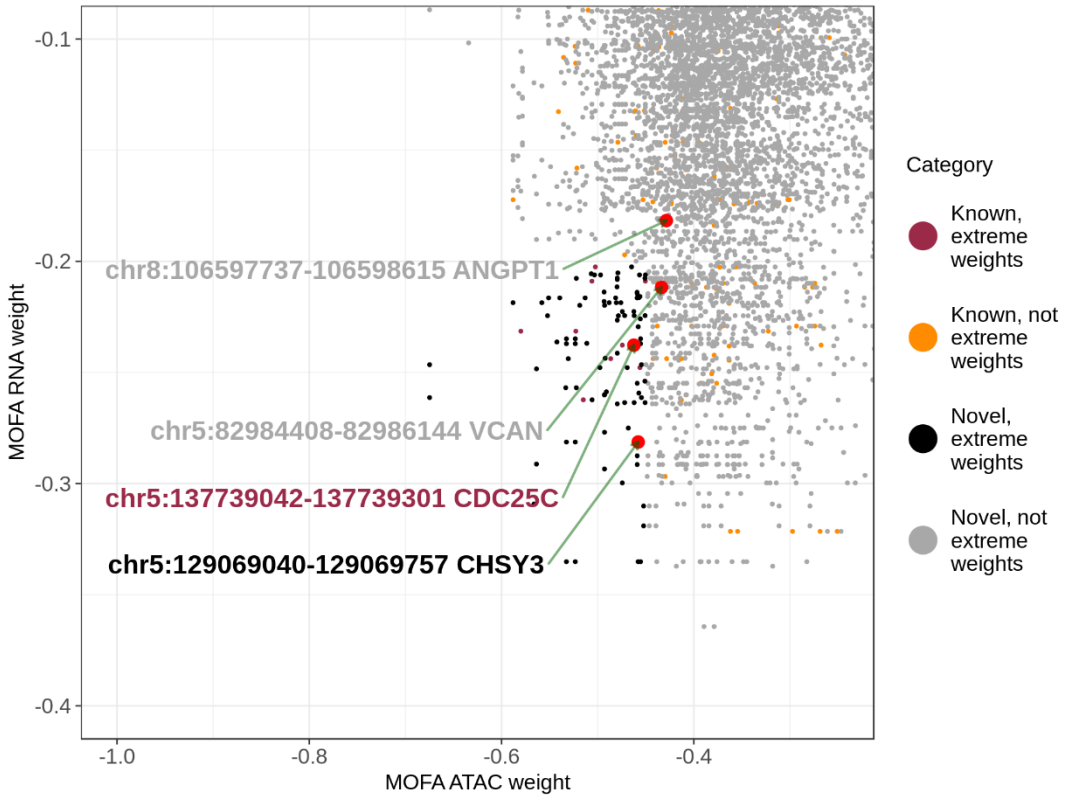


Figure 5-9: MOFA candidate enhancer - gene interactions for LF1 in the context of MM vs. PC analysis.

Each dot represents an interaction between a MOFA accessibility and gene feature within 1Mb. Different categories are created depending on whether or not the interaction overlaps an already found MM enhancer regulating a OEMM protein coding gene ("Known") or not ("Novel"). Also an interaction is classified on whether it contains extreme negative weights for both LF1 ATAC (less or equal to -0.45) and RNA-seq features (less or equal to -0.2). Only a fraction based on the range of weights of all interactions is shown.

The novel candidate MM enhancer chr5:82,984,408-82,986,144 is found, among others, in proximity to the cancer OE gene VCAN (highlighted in Figure 5-9 as "chr5:82,984,408-82,986,144 VCAN" in grey). Its protein product accumulates in MM lesions (Hope et al., 2016) and sustains inflammation favouring tumours while suppressing the immune response preventing T-cell infiltration (Pagenkopf et al., 2017), it has a MM association score of 0.02. A new interaction with extreme weights is CHSY3 (a gene with transferase activity) in the vicinity of chr5:129,069,040-129,069,757 (among other candidate regions present in MOFA). CHSY3 is reported (marked with the label "chr5:129,069,040-129,069,757 CHSY3" coloured black in Figure 5-9) in the literature as OE in the disease compared with PC (Zhou et al., 2009).

Other candidate interactions examples present in the MOFA analysis involving genes not significantly associated with MM (Piñero et al., 2020) and having negative weights found to be expressed or OE in MM are: CASP12 (Chauhan et al., 2010), ATP10B (Broyl et al., 2010; Kassambara et al., 2012), ELOVL4 (Condomines et al., 2009), PARD3B containing SNV in MM (Egan et al., 2012) and SH3BGRL2 (Kassambara et al., 2012). Also, the following genes with negative weights may be of future interest to be studied: SNORA63 (a type of small nucleolar RNAs that enable chemical modifications of other RNAs such as methylation), LINC02029 (a long intergenic non-protein coding RNA) and IFI44, an interferon protein linked to IRF4, a key plasma cell gene (Klein et al., 2006), also linked in lymphomas (Wang et al., 2014). There are other genes selected as MOFA features that despite not having extreme LF1 weight values are of relevance, for example, FGFR3. FGFR3 contains a very strong association score with MM (0.7), it is a translocation partner, in t(4;14) MM samples in conjunction with MMSET (Kalff and Spencer, 2012). Other genes are AZGP1: correlated with survival of non-MMSET myeloma patients (Wu et al., 2016), SFRP2: found to be secreted by myeloma cells, inhibiting the Wnt signaling pathway and preventing bone formation (Oshima et al., 2005) and having a MM association score of 0.02.

Similarly to LF1 and correlated (Figure 5-4 B), LF2 completely separates PC samples from the cancer samples (Figure 5-3) with PC having positive weights (Figure 5-6 C and D), the variance explained by this LF is dominantly for gene expression with little chromatin accessibility (Figure

5-4 A). ATAC-seq features have an even distribution of positive and negative weights, with the weights for those overlapping MMPC enhancers from the supervised analysis being slightly enriched for negative weights in comparison (Figure 5-10 C). This occurs because the factor decreases through Myelomagenesis, therefore increasing accessibility in the cancer state is reflected by negative weights and the recapitulated regions therefore have importance in this separation.

The accessibility LF weights are centred on a MM vs. PC log_2FoldChange of around 2, with a moderate negative correlation between these two variables (Figure 5-10 A). As with LF1, this axis separates samples by chromatin openness between healthy and cancer state.

LF2 gene expression features have both positive and negative weights, with a bias for positive. There is a bimodal distribution of weights: positive corresponding to known (chapter 3) MM under expressed genes and negative ones with OE in MM (Figure 5-10 D). Consistent with this, a strong negative correlation between MM vs. PC gene log_2FoldChange and weights exist (Figure 5-10 B) which occurs both with all genes used in MOFA and only with overlapping DEMM genes from the supervised analysis. This LF delineates PC to MM gene expression change with most of the positive weights being more extreme than the most negative weights.

This points at a stronger importance on the known PC OE genes (maybe tumour suppressor genes), than the MM ones (perhaps proto oncogenes). The reason for this could be due to the fact that there are less of the former than the latter and its importance is inversely proportional to its number. Together with the accessibility, LF2 reflects coordinated changes in general chromatin opening with corresponding gene expression changes through Myelomagenesis.

LF2 Gene Ontology on the top 500 genes by extreme negative weight has enrichment in 9 categories, importantly, there are no ontology categories overlapping extreme-LF1-weighted-genes. Interestingly, most are neuron and synapse related, also present is the voltage-gated cation channel activity.

Interactions of MOFA regions with genes within 1Mb were analysed in the context of LF2 (Figure 5-11), 391 out of the 2,698 total MMPC enhancers regulating DEMM protein coding genes were recapitulated by the analysis.

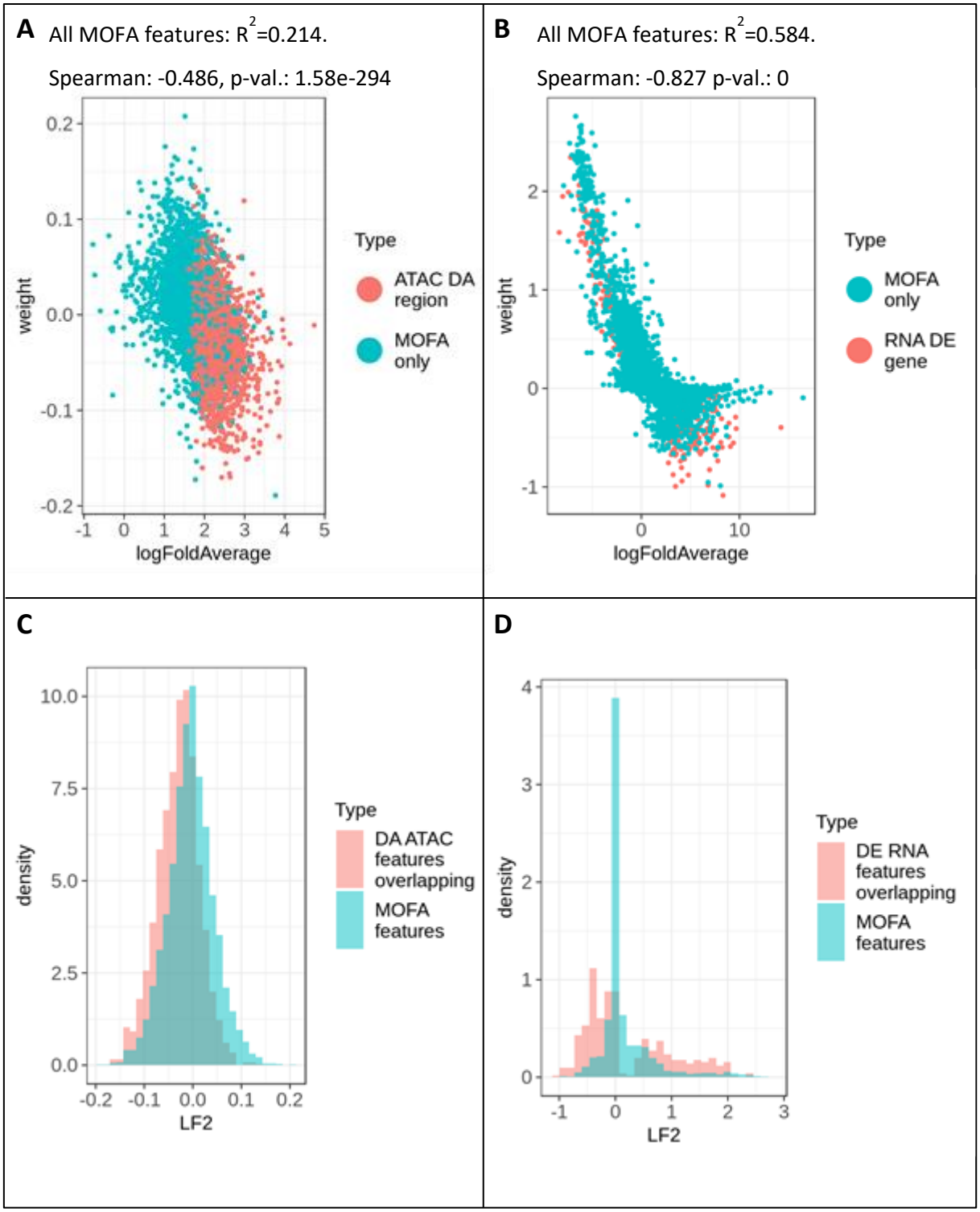


Figure 5-10: Chromatin accessibility and gene expression MOFA features for LF2.

LF2 weight loadings for MOFA input features.

A and B: Correlation with log_2FoldChange average (A: accessibility, B: gene expression) for the MM vs. PC analysis in Chapter 3: MM vs. average PC CD19, MM vs. PC donor average and MM vs. PC accounting for batch. Correlation metrics taking into account all MOFA features. A: features with distinction for MMPC enhancer features (labelled "ATAC DA region") overlapping and non-overlapping (labelled "MOFA only") MOFA features. B: features with

distinction for DEMM genes (labelled "RNA DE gene") overlapping and non-overlapping (labelled "MOFA only") MOFA features.

C: Weight density distributions for accessibility features with distinction between MMPC enhancers (labelled "ATAC DA features overlapping") MOFA features and all MOFA features (labelled "MOFA features").

D: Weight density distributions for gene expression features with distinction between DEMM genes (labelled "DE RNA features overlapping") overlapping MOFA features and all MOFA features (labelled "MOFA features").

An example of one such interaction is the chr15:33,915,253-33,916,443 candidate enhancer with the FMN1 gene (marked in magenta with the label "chr15:33,915,253-33,916,443 FMN1" in Figure 5-11), the expression of which is increased in advanced stages of MM (Lu et al., 2018). Both the genomic region and gene have significant negative weights and is more than 5 fold more accessible and nearly 11-fold OE in MM. The gene is expressed throughout both the cancer and PC condition, while the candidate enhancer is already accessible in PC, hence why it is not classified as a MM enhancer), exclusive to MM. Another potentially functional regulation that recapitulates a MM enhancer regulating a protein coding OEMM gene is chr14:24,246,900-24,247,115 with STXBP6 (marked in orange with the label "chr14:24,246,900-24,247,115 STXBP6" in Figure 5-11). With a MM chromatin openness and overexpression of more than 6 and 9-fold respectively and negative weights signifying the activation in MM. STXBP6, is a gene found to be deregulated when applying Histone deacetylase inhibitors (HDACi) and DNA methyltransferase inhibitors (DNMTi) as therapeutics on a murine MM model. The human ortholog has prognostic value in MM patients (Maes et al., 2015).

A novel MM putative enhancer – promoter communication with extreme negative weights is chr7:121,994,561-121,995,365 correlating FEZF1-AS1 overexpression (marked in black with the label "chr7:121,994,561-121,995,365 FEZF1-AS1" in Figure 5-11). FEZF1-AS1 is a non-protein coding RNA involved in multiple cancerous processes within different cancer types such as colorectal (Bian et al., 2018), pancreatic and lung adenocarcinoma among others and linked with poor prognosis (Shi et al., 2019). It has been reported that its expression is increased in primary MM samples and cell lines, within them FEZF1-AS1 promotes MM cell proliferation through the MIR610/AKT3 axis and FEZF1-AS1 suppression creates arresting of the cell cycle and produces cell death *in vitro* (Li et al., 2018).

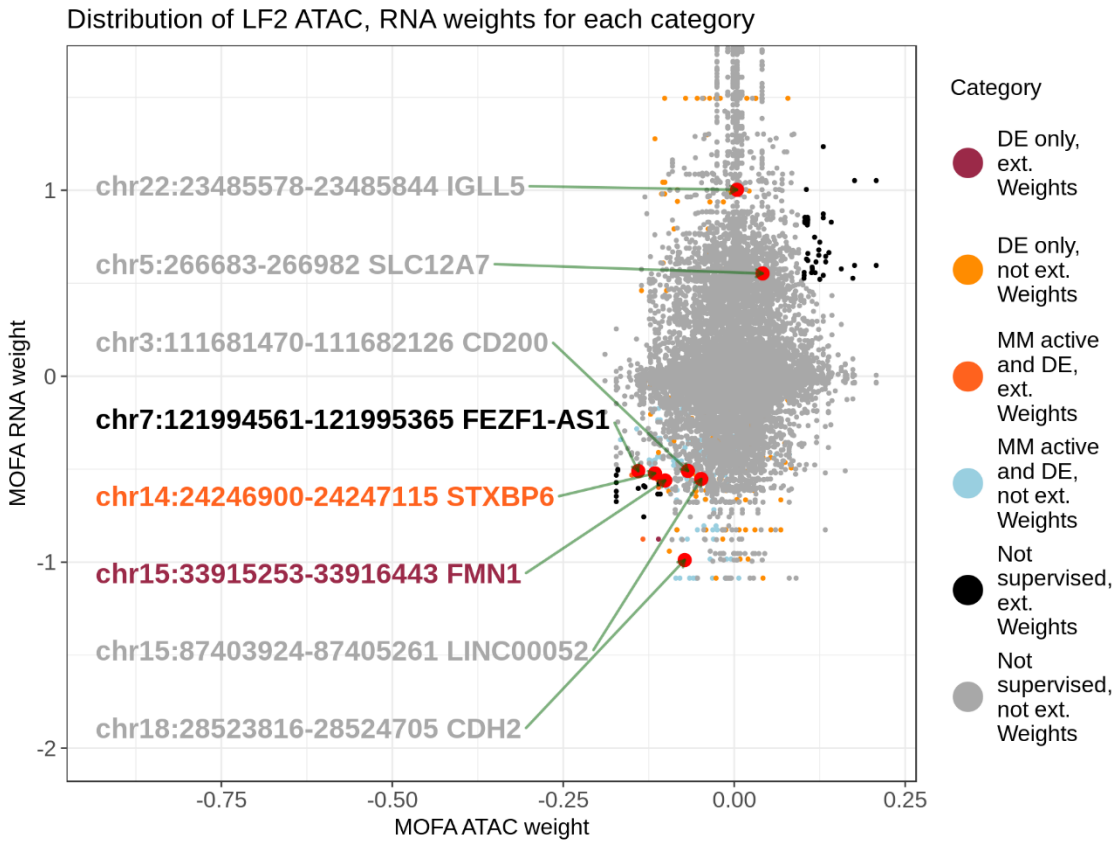


Figure 5-11: MOFA candidate enhancer - gene interactions for LF2 in the context of MM vs. PC analysis.

Each dot represents an interaction between a MOFA accessibility and gene feature within 1Mb. Only the weights within the figure limits are shown. Different categories are created depending on:

Whether or not the interaction overlaps an already found MMPC enhancer regulating a DEMM protein coding gene (label including "DE").

Whether the interaction overlaps a MM enhancer regulating an OEMM protein coding gene (and therefore also DE), label including "MM active and DE".

"Not supervised" label: MOFA interaction not found in any supervised analysis.

"Extreme weights": defined as ATAC weight equal or more negative than -0.10 or equal or larger than 0.10, MOFA RNA weight equal or more negative than -0.5 or equal or larger than 0.5.

Other regulatory connections are activated in MM vs. PC with negative (albeit not extreme weights) either for the gene or accessible feature, but contain relevant genes to MM: chr3:111,681,470-111,682,126 with CD200 (marked in grey with the label "chr3:111,681,470-111,682,126 CD200" in Figure 5-11). CD200 is a gene expressed in MM and not in PC with prognostic value (Douds et al., 2014). Another interaction is chr18:28,523,816-28,524,705 with

the CDH2 cadherin 2 gene (marked in grey with the label “chr18:28,523,816-28,524,705 CDH2” in Figure 5-11), having a 266-fold over expression in MM and MM association score of 0.02. There is evidence that targeting of N-cadherin in MM is an effective treatment option (Mrozik et al., 2015). Similarly, chr15: 87,403,924-87,405,261 and LINC00052 (also known as lnc-AGBL1-4), a Long Non-Protein Coding RNA with a 102-fold expression in MM, implicated in breast (Salameh et al., 2017) and gastric cancer (Shan et al., 2017). Through *in silico* analysis of interactions, in primary MM, LINC00052 is thought to interact with MIR185, a tumor suppressor in other cancers (Lei et al., 2018; Tang et al., 2014). This enhancer – promoter interaction is marked in grey in Figure 5-11 with the label “87,403,924-87,405,261 LINC00052”.

Other genes are found containing LF2 weights signifying MM activation, with interactions that may be of interest and relevant overexpression in MM with no MM – gene association score (and therefore may be of novel importance to MM). Examples of these include DCAF4L2, involved in colorectal cancer invasion and metastasis (Wang et al., 2016); ADGRG6, proposed to contribute to pathological angiogenesis in urothelial bladder carcinomas (Wu et al., 2019); VTRNA1-1, whose expression is proposed to contribute to cancer cell line resistance to chemotherapy (J. Chen et al., 2018); CYP2J2, a gene shown to promote apoptosis and curb cell proliferation in lung cancer when repressed by MIRLET7B/let-7b (Chen et al., 2012); CNTN1, a neural cell adhesion protein that promotes prostate cancerogenesis (Yan et al., 2016), also in breast cancer (N. Chen et al., 2018) and gastric cancer (Chen et al., 2015) and classified as a possible potential T-cell antigen for gliomas (Dettling et al., 2018). Examples of genes with no prior association to MM include: LOC100508631, MTMR11, LMAN1L, HIST2H4B, RXFP4, HIST2H2AA4, HIST2H2BC, KIAA1217, RIMS2, FMNL3 or GPRC5A. A protein coding OEMM also found in the MOFA analysis is GALNT13.

Also, genes with previously found MM association scores (0.02 and 0.06) are found: RELN and SYT1 respectively. RELN, the Reelin protein-coding gene which protects MM cells from doxorubicin caused cell death by MM cell linkage to fibronectin (Lin et al., 2017a), SYT1 is a protein coding OEMM.

Some interactions involve MM suppressed genes such as IGLL5 (with a 9-fold lower expression than in PC), which may be regulated by chr22:23,485,578-23,485,844, a region which is more accessible in PC and has positive weights (corresponding to PC activation with respect to MM). The interaction is marked in grey with the label “chr22:23,485,578-23,485,844 IGLL5” in Figure 5-11. IGLL5 is found to be a tumor suppressor in large B-cell lymphoma (Cornish et al., 2019).

Also, chr5:266,683-266,982 regulating SLC12A7 with diminished accessibility and expression in MM (marked in grey with the label “chr5:266,683-266,982 SLC12A7” in Figure 5-11).

5.3.4. LF3 distinguishes the MMSET subgroup

MMSET samples have the most negative LF3 weights (Figure 5-3). The LF weights for the chromatin accessibility and gene expression features are distributed around 0 with some bias for negative weights (not shown), which represent MMSET activation. 4,855 out of 5,000 of the features used in MOFA overlap a subgroup consensus peak (MOFA features derive from MM and PC accessible peaks). Nevertheless, the distribution of LF3 weights for these subgroup features is representative of all (not shown). As Table 5-1 shows, there are 662 MMSET – PC DA features present in the analysis (647 more accessible in MMSET samples). Predictably, there is an enrichment of negative LF3 weights in the DA genomic regions (Figure 5-12 A), since the great majority are MMSET activated. As expected, there is a moderate negative correlation between LF3 weights and MMSET vs. PC $\log_2$FoldChanges for both MOFA regions corresponding to subgroup MM and PC peaks and DA areas between the MMSET subgroup and the healthy state (Figure 5-12 B). This correlation doesn’t translate in terms of weights with absolute accessibility (not shown), signifying that LF3 explains a change in accessibility between MMSET and the remaining samples.

There are 1,130 DE MMSET – PC genes (645 of them more expressed in MMSET, Table 5-1) which makes an even number of under and OE genes. The LF3 weights for the MMSET DESMM genes are more extreme (positive and negative) than for all the MOFA gene features (Figure 5-13 A) and this is consistent with the division this LF produces (MMSET activation). Concordant with this, as it is the case with the accessibility features there is a moderate negative correlation between MMSET vs. PC $\log_2$FoldChange in expression and LF3 weights, but in this case the anti-correlation exists only when taking into account the DE genes (Figure 5-13 B, Spearman’s correlation -0.497) and not all the MOFA genes (-0.389, not shown). The absolute gene count means are not correlated with the LF weights, suggesting, as with accessibility, this axis reflects MMSET activation vs. the rest of samples.

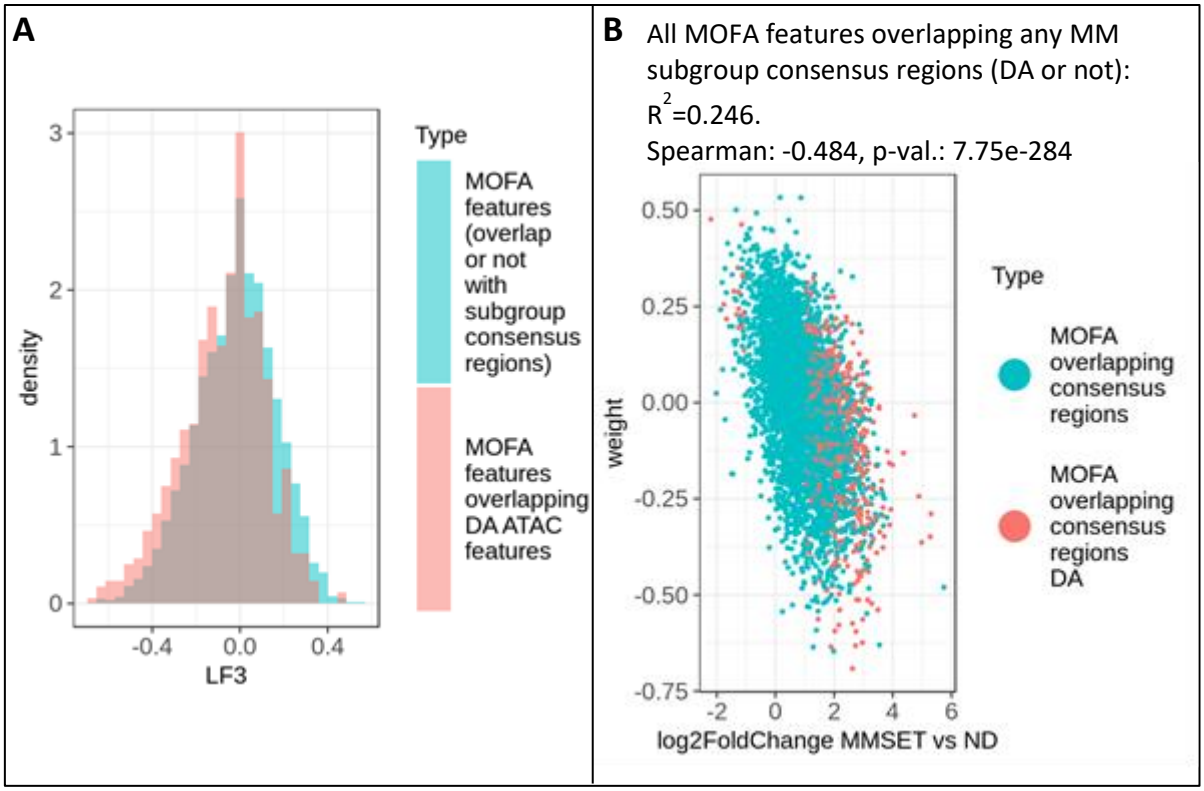


Figure 5-12: MOFA LF3 accessibility features overlapping DASMM regions.

LF3 weight loadings for MOFA input reads in peaks features.

A: Weight density distributions for accessibility features, with colour distinguishing between all MOFA features and MOFA features that are also MMSET – PC DA (label "MOFA features overlapping DA ATAC features").

B: Correlation of LF3 weights with MMSET vs. PC log_2FoldChange (Chapter 4). Features that overlap MM subgroup vs. PC consensus regions (Chapter 4) which are also MMSET – PC DA are distinguished by colour. Correlation metrics taking into account all MOFA regions overlapping MM subgroup vs. PC consensus regions.

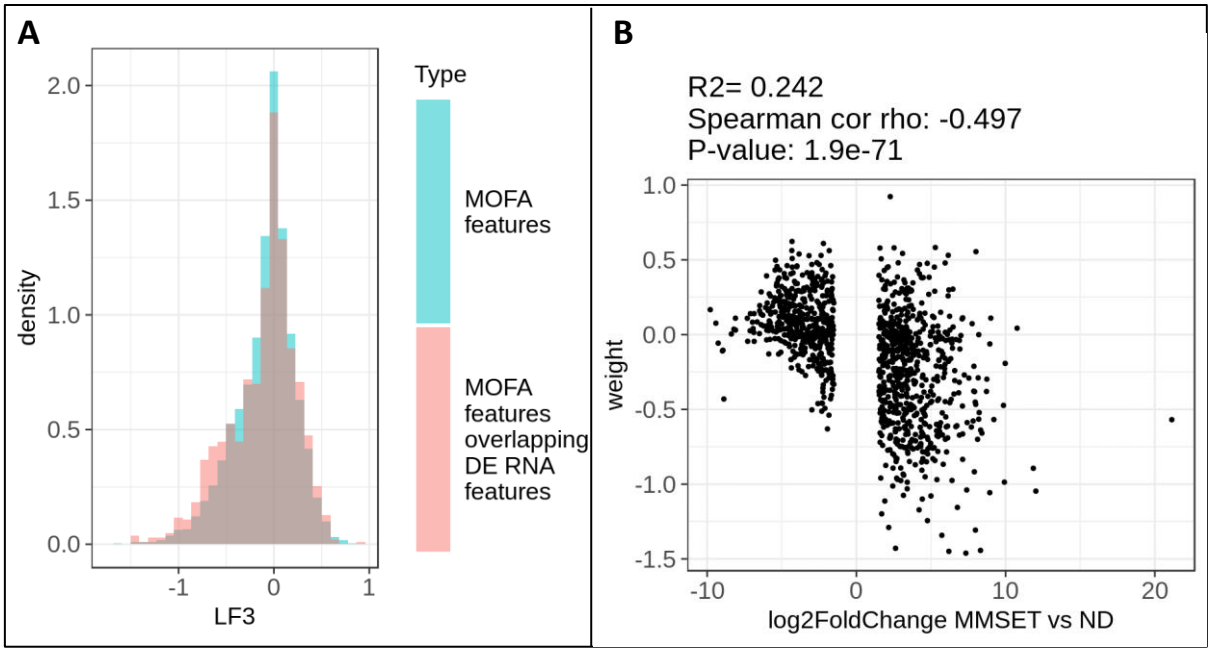


Figure 5-13: MOFA LF3 gene expression features overlapping MM subgroup DE genes.

LF3 weight loadings for MOFA gene counts. Any MOFA features with no MMSET vs. PC log_2FoldChange or mean count available due to Deseq2 filtering are disregarded.

A: Weight density distributions for features with distinction between all MOFA features (labelled "MOFA features") and MOFA features that are MMSET – PC DE (label "MOFA features overlapping DE RNA features").

B: Correlation of LF3 weights with MMSET vs. PC log_2FoldChange (Chapter 4) for MOFA features overlapping MMSET – PC DE genes.

5.3.4.1. Highly weighted genes in LF3

Consistent with this MMSET separation, the loading value on the NSD2/MMSET gene (Figure 5-14 B in grey labelling) is -0.756, ranking in the top 3% loadings by absolute value, which shows its high relative importance in the model. Figure 5-14 A, exhibits how the negative values of the LF are associated with high MMSET expressing samples. In this category, the cytogenetically unannotated sample A17.5 shows the most extreme negative weight. As was shown in Chapter 4, this sample has an MMSET-like expression and clusters together with the MMSET subgroup in terms of accessibility and gene expression.

This and other genes thought to be relevant for MMSET, with the corresponding LF3 weight loading ranking, are shown in Figure 5-14 B. Genes in MOFA overlapping a list of 71 genes differentially expressed in MMSET patients vs. Non-MMSET patients (Wu et al., 2016) are marked in Figure 5-14 B with red circles. Within them, two genes with positive (CCND1 and NOL4), near zero (EDNRB and CDH2) and negative weights (CDC42BPA and MAP1B) are

labelled in red (Figure 5-14 B). CCND1 (LF3 weight of 0.36) is an IgH translocation target having a modest MMSET vs. PC log_2FoldChange of 1.48, its expression seems to be discriminatory between MMSET and non-MMSET MM patients (Wu et al., 2016). CCND1 is very significantly overexpressed in CCND1 translocated MM patients compared with the rest (including the MMSET subgroup), thereby having a positive weight (non-MMSET activation). NOL4 is the only other gene identified by Wu and colleagues with positive LF3 weight (0.23), it was found to be cancer-testis antigen with expression in MM making it a possible candidate as a biomarker and therapeutical target for MM patients (Ghafouri-Fard et al., 2015). Although not significant, NOL4 is more than 13 times more expressed in the MMSET subgroup compared with PC. However, it tends to have similar overexpression in other MM subgroups compared with PC, while being highly OE in HD (and therefore around half of all the samples), hence why it may appear to be deactivated in MMSET compared with the rest of samples and have a positive LF3 weight.

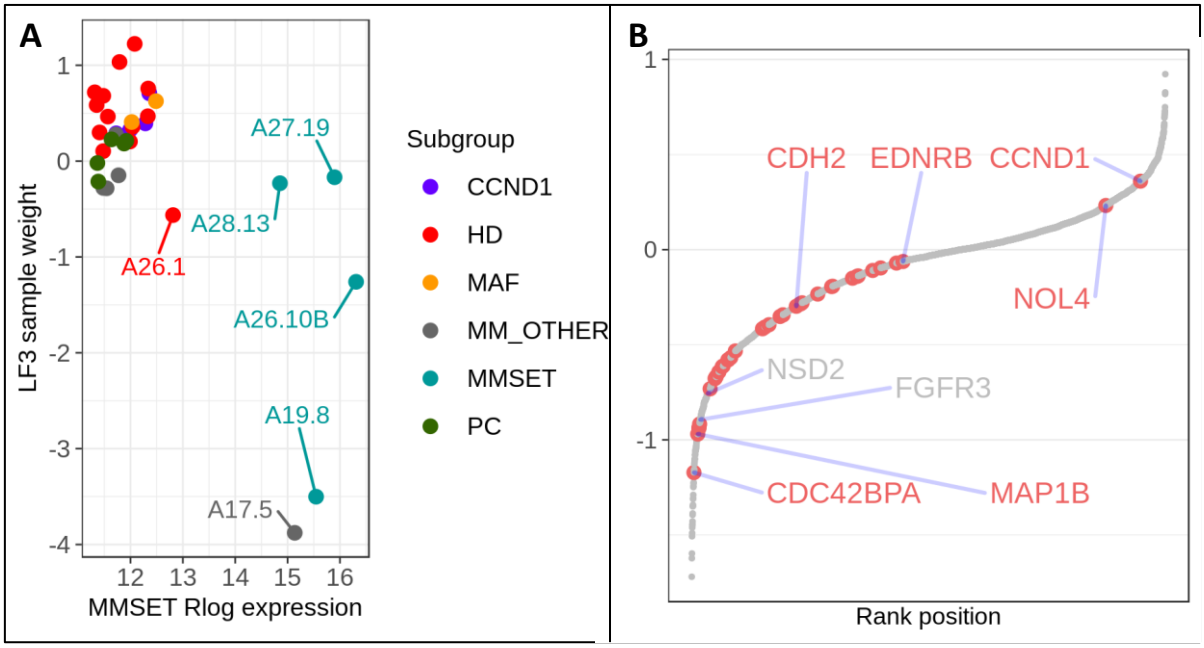


Figure 5-14: LF3 with MMSET subgroup genes.

A: MMSET Rlog normalized expression (as inputted to MOFA) with LF3 sample weight for the different subgroups. MM_OTHER: samples with no cytogenetic information available.

B: MOFA gene expression weight loadings for LF3 ranked with relevant MMSET features marked. Genes marked with red circles overlap with a list of 71 genes differentially expressed in MMSET patients vs. Non-MMSET patients (Wu et al., 2016), from these, the analysed genes are labelled in red. Genes involved in IgH translocation in MMSET are abelled in grey.

EDNRB and CDH2 are genes involved in activating a phosphatidylinositol-calcium second messenger system and generating calcium-dependent cell adhesion molecule and glycoprotein respectively. EDNRB and CDH2 have a LF3 weight of -0.06 and -0.30, despite having a nearly 500-fold and 428-fold increase in expression from MMSET to PC respectively; the weight is indicating low importance for this separation. The reason why these genes might not be contributing significantly to the MMSET vs. rest separation might be because they are similarly overexpressed in the majority of samples from other subgroups. This is the case in both, with overexpression in the HD (281 and 286-fold) and CCND1 (13 and 25-fold) subgroups with respect to PC for CDH2 and EDNRB expression respectively.

MAP1B is a protein implicated in normal cells in microtubule assembly and also as part of neurogenesis, the gene has an extreme LF3 weight of -0.97 and correspondingly it is 42-fold overexpressed in MMSET vs. PC, second to the CCND1 subgroup with only 3-fold overexpression. It is thought to interact with *REIIBP* (one of the two protein isoforms of the MMSET gene) in H929 cell lines harbouring the MMSET translocation (Mirabella et al., 2014). MAP1B has a 42-fold higher expression in MMSET samples compared with PC in this study. It was found as possible target of multiple microRNAs (miRNAs) in MMSET translocated vs. non-translocated MM samples through regulation models and hypothesized to interact with FGFR3 in conjunction with MYRIP and CDC42BPA (Liu et al., 2019). CDC42BPA (also labelled in red in Figure 5-14 B) has a negative weight of -1.17 indicating MMSET-only activation, which is the case: CDC42BPA is significantly overexpressed with respect to PC in MMSET while being underexpressed in the rest of the subgroups.

Tumours containing the IgH - NDS2/MMSET t(4;14) also have high FGFR3 expression in around 75% of the cases (Stewart et al., 2004). Consequently, this gene has an extreme negative weight, ranking in the top 2% by absolute weight value (Figure 5-14 B in grey labelling). It can be seen that the previously determined genes separating MMSET from non-MMSET MM samples (Wu et al., 2016) overlap extreme and non-extreme LF3 weights (Figure 5-14 B with red circles) with genes such as EDNRB and CDH2 having close to zero weights.

Gene Ontology analysis was carried out on the top 10% genes by absolute LF3 loading using as background distribution the top 5k genes by total variation used as input in the MOFA analysis at FDR 0.1 obtaining an enrichment for only one category: chromatin binding category.

5.3.4.2. Gene – region pairs in LF3

Since LF3 accounts for a significant fraction of the accessibility and gene expression variability (Figure 5-4 A), enhancer – promoter interactions assigned an important weight are likely to have meaningful biological implications distinguishing this subgroup from the rest. Additionally, they could be the consequence of biologically relevant changes, for example, upregulation of a gene such as MYC deregulates multiple target genes (some of them might be tumorigenic). As Figure 5-15 shows, MMSET activated interactions compared to normal (from the supervised analysis in chapter 4) tend to lie in the negative LF3 weights, consistent with the fact that this reflects overall MMSET activated interactions in the MOFA model. Moreover, there are examples of interactions showing deactivation in the MMSET subgroup.

For example, the candidate enhancer chr3:46,386,101-46,386,648, which is less accessible in MMSET, is close to CCR5 (marked with "chr3:46,386,101-46,386,648 CCR5" in magenta in Figure 5-15), a gene with a nearly 25-fold decrease in MMSET compared with healthy state. This relationship has positive extreme weights (larger than 0.25 and 0.45 for ATAC and RNA respectively), signifying deactivation in MMSET. It was also found to be significant in the MMSET – PC differential supervised analysis. CCR5 is a member of the chemokine receptor family. When activated, such receptors trigger cell responses such as chemotaxis (migration of cells). MM cells encourage osteoclast formation and in turn osteoclasts express the MM risk factor CCL3 which signal CCR1 and CCR5 receptors and promote MM growth (Abe et al., 2004; Vallet et al., 2007; Yaccoby, 2010). Only CCR1 (which is more than 8-fold expressed in MMSET compared with PC) inhibition reduces formation of mature osteoclasts in the MM context (Dairaghi et al., 2012), but in the L363 MM cell line, CCR5 is highly expressed and a significant migration of cells occurs towards the CCR5 ligand CCL5 (Udi et al., 2013). Perhaps in MMSET, this oncogenic pathway is less favoured than in other MM subtypes.

Examples of genes upregulated in MMSET with a nearby region that gains accessibility that were also found in the supervised analysis (chapter 4) include CDC42BPA and ROBO1. chr1:227,191,697-227,192,560 (among other candidate regions) interacts with the CDC42BPA gene (marked with "chr1:227,191,697-227,192,560 CDC42BPA" in blue in Figure 5-15) and is active in MMSET (signified by extreme negative LF3 weights) as can be seen in (Figure 5-14 B red circles). As explained in the previous section, CDC42BPA is found altered when between MMSET and non-MMSET subgroup MM patients in a previous study (Wu et al., 2016), thought to affect FGFR3 in a majority of MMSET translocated samples (Liu et al., 2019). Together with CLEC11A, it is thought to be a regulator in MM primary samples and shown to be key in two t(4;14) MM CLs: KMS-26 and NCI-H929, it is also critical for these cells survival (Laganà et al.,

2018). Consistent with this finding, CLEC11A also has an extreme negative LF3 weight (MMSET active) and ranks in the top 3% by absolute value, despite not having a MOFA accessibility feature (either highly or lowly ranked) nearby.

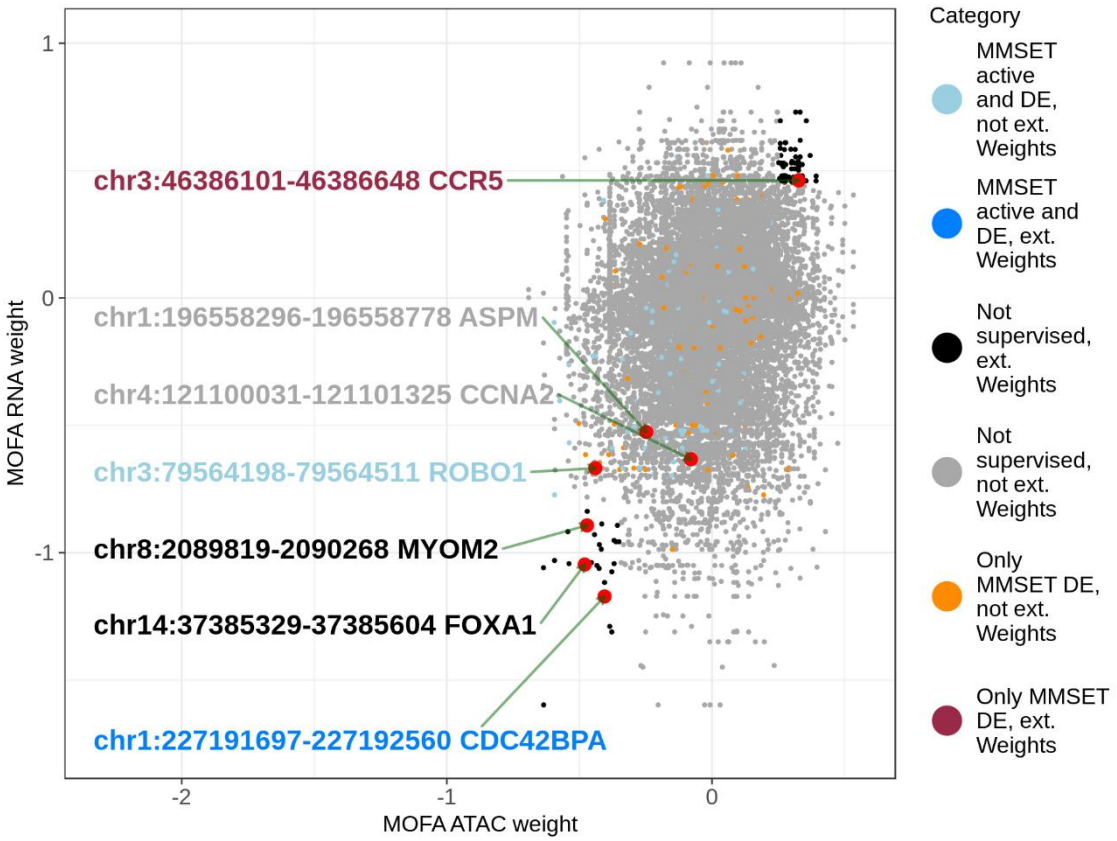


Figure 5-15: MOFA candidate enhancer - gene interactions for LF3 in the context of MM vs. PC analysis.

Each dot represents an interaction between a MOFA accessibility and gene feature within 1Mb.

Different categories are created depending on:

"ext. weights": LF3 extreme weights which are positive or negative simultaneously for both ATAC and RNA: MOFA ATAC weight equal or more negative than -0.35 or equal or larger than 0.25, MOFA RNA weight equal or more negative than -0.8 or equal or larger than 0.45).

"Only MMSET DE": MOFA interaction that is also a MMSET DASMM enhancer regulating a MMSET DESMM gene.

"MMSET active and DE": MOFA interaction that is also a MMSET SMM enhancer regulating an MMSET OESMM gene.

"Not supervised": MOFA interaction not found in any supervised analysis.

Another example of a gene previously found relevant in the MMSET subgroup (Wu et al., 2016) is the ROBO1 gene with the genomic area chr3:79,564,198-79,564,511 (marked with "chr3:79,564,198-79,564,511 ROBO1" in light blue in Figure 5-15), having a 16-fold and 8-fold enrichment in expression and accessibility in MMSET vs. PC respectively. The interaction has region and gene negative weights (despite not being extreme) agreeing with its activation state in MMSET samples (subgroup supervised analysis). ROBO1 has been previously shown to be significantly correlated with MMSET subtype myeloma and also significantly associated with the survival of non-MMSET myeloma patients (Wu et al., 2016).

Genes are found to be MMSET active in MOFA, but which were not found by the supervised analysis include MYOM2, FOXA1, ASPM, CCNA2 and BMP4. MYOM2 and chr8:2,089,819-2,090,268 (marked with "chr8:2,089,819-2,090,268 MYOM2" in black in Figure 5-15) have a 5 and 3-fold enrichment in gene expression and accessibility respectively compared to PC. MYOM2 is a MM "spike" gene, defined as having a "strong overexpression in MM cells of a fraction of patients" (Kassambara et al., 2012). It was shown to be DE between PCs and MGUS PCs and decreased abundance of its protein (M-protein 2) was found have statistically significant association with survival after oral Melphalan and Prednisone therapy in MM patients (Palmer et al., 1988). The region chr14:37,385,329-37,385,604 and the gene FOXA1 (marked with "chr14:37,385,329-37,385,604 FOXA1" in black in Figure 5-15), with more than 4,000 and 53-fold higher accessibility and expression compared to normal and being only expressed in MMSET, is also classified in this category. FOXA1 is complex in the cancer landscape, it is recurrently mutated in prostate cancer (Barbieri et al., 2012) but high expression of this gene tends to be associated with good prognosis in breast cancer (summarized in Hu et al., 2014). Another novel interaction is chr1:196,558,296-196,558,778 with ASPM (marked with "chr1:196,558,296-196,558,778 ASPM" in grey in Figure 5-15) with an 8-fold higher activation (accessibility and expression) in MMSET, having negative weights (despite not being extreme). ASPM produces proteins required in the cell cycle and mitosis, it is found to be down regulated in primary bone marrow (BM) MM cells (Cohen et al., 2014). High expression of this gene predicts shorter time to MM progression (Sarasquete et al., 2013) and its expression is higher in extra medullary (tumour cells found outside of the BM) relapsed MM tumours compared to MM BM PCs (Sevcikova et al., 2015). It is also more highly expressed in some cell populations of cells of various MM CLs. These cells have cancer initiating, stem-like, features such as ability to differentiate, repopulation (proliferation of surviving cells during treatment), capability to form clones and self-renewal: division of a stem cell keeping its characteristics, and are selectively sensitive to aurora kinase and proteasome

inhibitors (Nara et al., 2013). The CCNA2 gene encodes a protein of the Cyclin family and it is involved in the cell cycle, an interaction is proposed in the MOFA analysis with chr4:121,100,031-121,101,325 (marked with "chr4:121,100,031-121,101,325 CCNA2" in grey in Figure 5-15). This interaction is more active in MMSET and also has negative weights. CCNA2 has previously been found to be down-regulated when MM CLs are treated with Oxophenamide and Pterostilbenedown inducing S-phase cell-cycle arrest (Zhang et al., 2018) and when preclinical models of MM are treated with Lenalidomide/Dexamethasone (LDA) and pan-BCL2 inhibitor (Paulus et al., 2014). It is also thought to be regulated by the miRNA miR-150 in MM (Bong et al., 2017). BMP4 is another gene associated with multiple general MM candidate regulatory enhancers which may be of relevance and found to have enrichment of B-cell 3C contacts between the regions and the promoter in Chapter 3, section 3.3.8. It has a 64-fold MMSET enrichment and it is a bone morphogenetic protein involved in bone development. In some MM cultures, BMP4 hinders proliferation and triggers apoptosis (Fukuda et al., 2006). In human samples, however, it is significantly OE in MM BM samples compared to BM PC and this overexpression is accompanied by higher expression of its receptor ACVR1 and lower expression of NOG (a BMP antagonist). Furthermore, during Bortezomib treatment, BMPs seem to provide resistance (Grčević et al., 2010). Additionally, SMAD1 is also OE in MMSET (more than 10-fold enrichment compared to normal), and is associated with MM in the literature. In a treatment of MM which exposes the cells to BMP6, suppression of MM occurs by phosphorylation of SMAD1/5 and induction of mesenchymal stromal cells to differentiate into osteocytes (Grab et al., 2019), a similar process also occurs in AML favouring the disease (Battula et al., 2017). Activin A and B are also found to collaborate in SMAD1/5 phosphorylation (mediated by ACVR1 receptor signalling) triggering MM growth inhibition (Olsen et al., 2018), SMAD1/2/8 phosphorylation can also supress MYC increasing apoptosis in human MM (Jiang et al., 2016).

5.3.5. LF4 isolates the outlier sample A26.9B

As it can be seen in Figure 5-3, LF4 separates the outlier sample A26.9B, a HD sample which has general extreme accessibility (not shown). As can be seen in Figure 5-4 A, LF4 is mainly explaining accessibility variability. It is unknown whether A26.9B is a biological or technical outlier, despite this, from the LFs explaining relevant variability, it only appears isolated from the rest of the HD samples in LF4, ensuring that this separation is capturing the outlier characteristics.

5.3.6. LF5 creates an axis splitting CCND1 and MAF translocated samples

As can be seen in Figure 5-3, LF5 places MAF translocated samples on the positive values of the axis and CCND1 on the negative side with the remaining samples in-between. The distribution of weights for accessibility features is dominant for positive weights (MAF activated enhancers). Furthermore, this distribution is very similar to that of the weight for the 97% of MOFA accessibility features overlapping subgroup MM and PC consensus peaks (not shown).

The MOFA features corresponding to MAF vs. PC regions of DA (supervised analysis) are enriched in positive weights with a smaller number of more negative ones (Figure 5-16 A). The weights are also significantly correlated with MAF vs. PC accessibility (Figure 5-16 C), even when taking into account only MAF – healthy DA regions only (not shown). The consequence being, that the LF5 positive weights resemble the 668 regions more accessible in MAF (Table 5-1) compared to the 48 regions with more open chromatin in PC. There is no connection between the overall accessibility of the features and the weights. CCND1 DA regions are enriched with somewhat more extreme weights on both ends (Figure 5-16 B); 590 of these regions have higher accessibility in CCND1 samples and 13 in the healthy state (Table 5-1). There is very little correlation between LF5 weights and CCND1 fold change in accessibility compared to normal (Figure 5-16 D) or LF5 weights and overall accessibility (not shown). Together, these results show that, at least in terms of MM subgroup compared with normal, MAF samples seem to be the driving force in this separation and CCND1 samples lying on the other extreme is a consequence of it. This results in CCND1 activation (vs. PC) in both extremes of the LF5 weights, supported by the clear dominance in the presence of CCND1 over accessible regions compared to PC enriched on both extremes in terms of LF5 weights. It is possible that the difference captured by this axis is between MAF and the remaining samples (healthy and other subgroups, particularly CCND1).

MOFA gene features have an even distribution of LF5 weights on both ends. As Table 5-1 shows, there are 1,157 MAF vs. PC DE genes, 465 out of these are MAF OE, while for CCND1 there are 277/1,059. Similarly to the accessibility features, a large proportion of the the MAF and CCND1 DESMM gene features overlapping have negligible LF5 weights. There is a gain of extreme positive weights for MAF DESMM genes (Figure 5-17 A) and little gain on negative and positive weights for CCND1 vs. PC DESMM genes (Figure 5-17 B). This suggests that for this LF the majority of genes are not relevant and the overexpression in MAF is of higher relative importance. Positive weights on CCND1 DESMM genes might point to these genes having an active state in MAF, (as seen most of them being CCND1 vs. PC repressed genes).

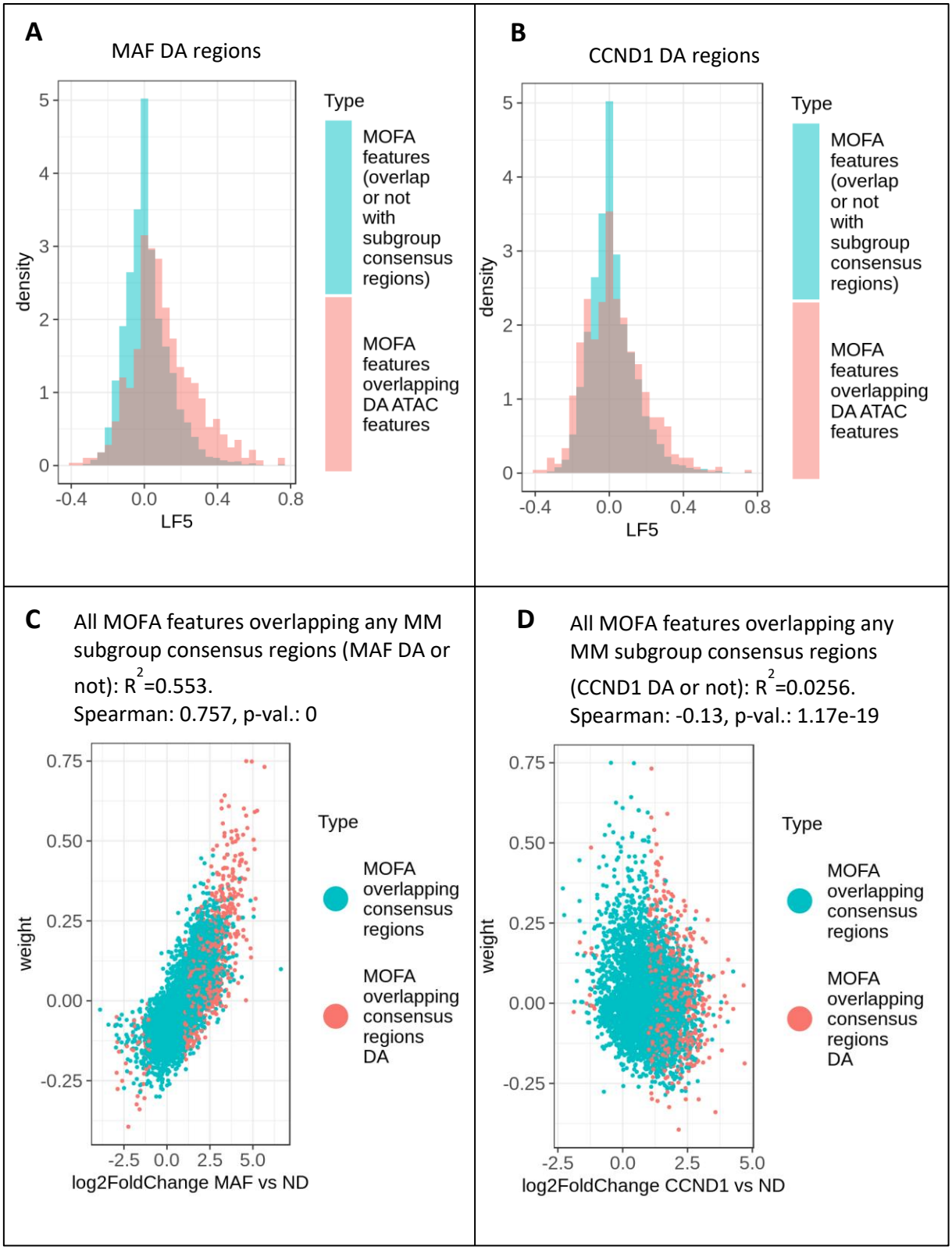


Figure 5-16: LF5 accessibility features with CCND1 and MAF DA regions.

LF5 weight loadings for MOFA input reads in peaks features.

Weight density distributions for accessibility features with distinction between all MOFA features and MOFA features that are A: MAF DASMM enhancers and B: CCND1 DASMM enhancers. The label in both cases for these categories are "MOFA features overlapping DA ATAC features".

Correlation of LF5 weights with C: MAF vs. PC log_2FoldChange (Chapter 4) and D: CCND1 vs. PC log_2FoldChange. Features with distinction for any MOFA region overlapping any MM subgroup vs. PC consensus regions and additionally overlapping each distinct subgroup DASMM enhancers (label "MOFA overlapping consensus regions DA"). Correlation metrics taking into account all MOFA regions overlapping any MM subgroup vs. PC consensus regions.

As with the accessibility features, the correlation between LF5 weights and MM subgroup log2Foldchange is significant for MAF (Figure 5-17 C), particularly when including only MAF vs. PC DE genes: Spearman's rank correlation coefficient of 0.695 (not shown). CCND1 doesn't show a similar trend, with no correlation (Figure 5-17 D). Together these results show that there is a dominant coupling between higher LF5 weight ranking and MAF activation with the difference in expression between MAF and healthy being very representative of this separation. CCND1 samples lie on the other extreme, with different samples having different degrees of similarity to PC and non-MAF subgroups in this dimension.

Similarly to LF3, gene ontology analysis was performed on the top 10% genes by absolute LF5 loading. All genes used as input to MOFA were used as background distribution and a FDR 0.1 threshold used. No pathways were over represented, but there was an under-representation of integrin binding and signalling receptor binding with respect to background.

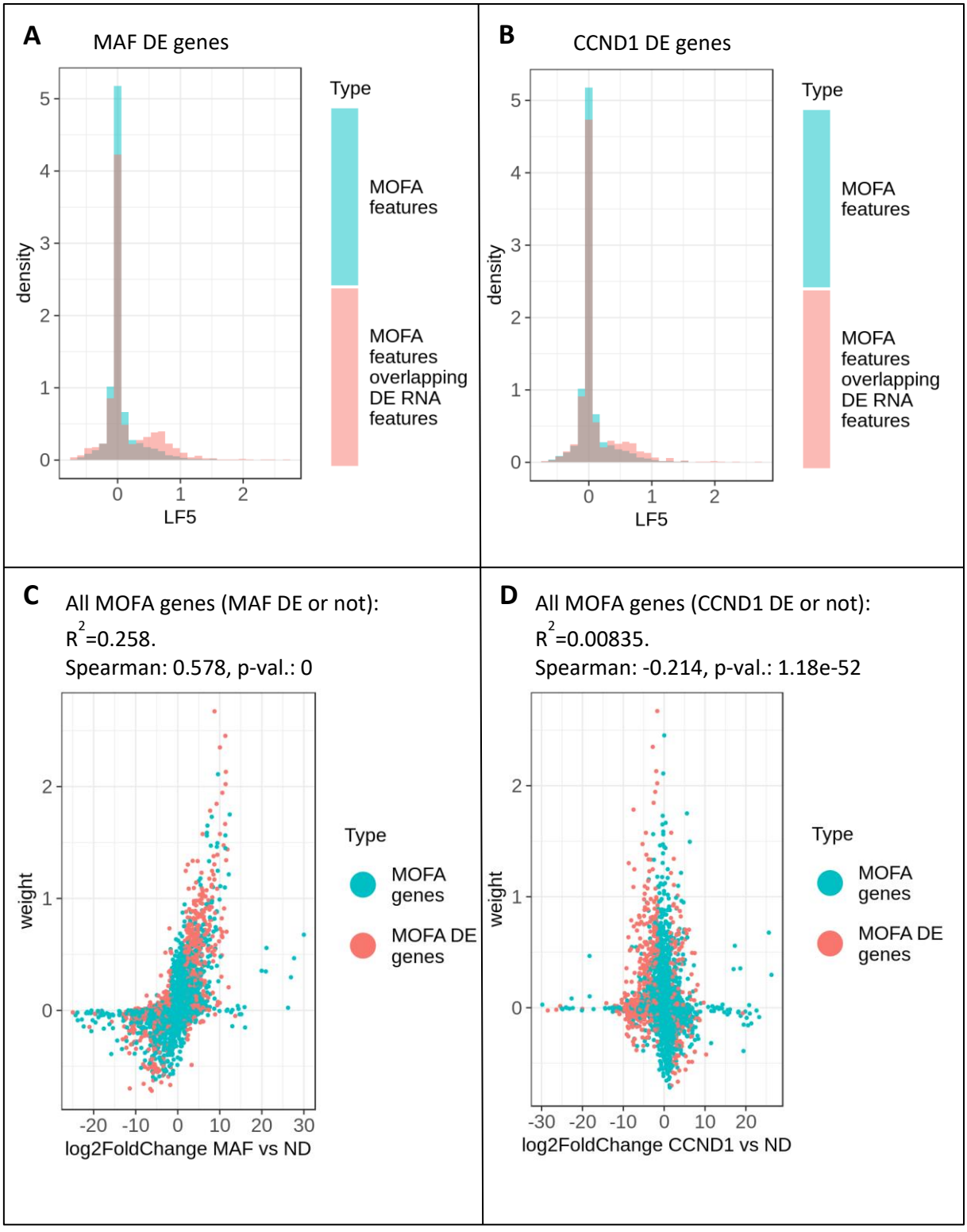


Figure 5-17: LF5 expression features with CCND1 and MAF DESMM genes.

LF5 weight loadings for MOFA gene counts. Any MOFA features with no subgroup vs. PC log2FoldChange or mean count available due to Deseq2 filtering are disregarded. Weight density distributions for features with distinction between all MOFA features (label "MOFA features") and A: MOFA features that are also MAF DESMM genes, B: MOFA features that are also CCND1 DESMM genes. The label in both cases for these features is "MOFA features overlapping DE RNA features".

Correlation of LF5 weights with C: MAF vs. PC log2FoldChange (Chapter 4) and D: CCND1 vs. PC log2FoldChange. Features coloured by whether they are also DESMM genes in each condition (label "MOFA DE genes") or not ("MOFA genes"). Correlation metrics taking into account all MOFA gene features.

5.3.6.1. Highly absolute weighted genes significant in MAF

Figure 5-18 studies genes included in the MOFA analysis and the associated LF5 weights in the context of previously found top over (Figure 5-18 in red) and under-expressed genes (Figure 5-18 in blue) separating clusters of 320 bone marrow MM samples into clinical subgroups (Broyl et al., 2010) representative of: MAF (Figure 5-18 A) and CCND1 (Figure 5-18 B). Among the labelled genes, two examples containing positive, near zero and negative loadings are analysed for each of the two subgroups (in the case of MAF, the only gene with negative loading is DKK1).

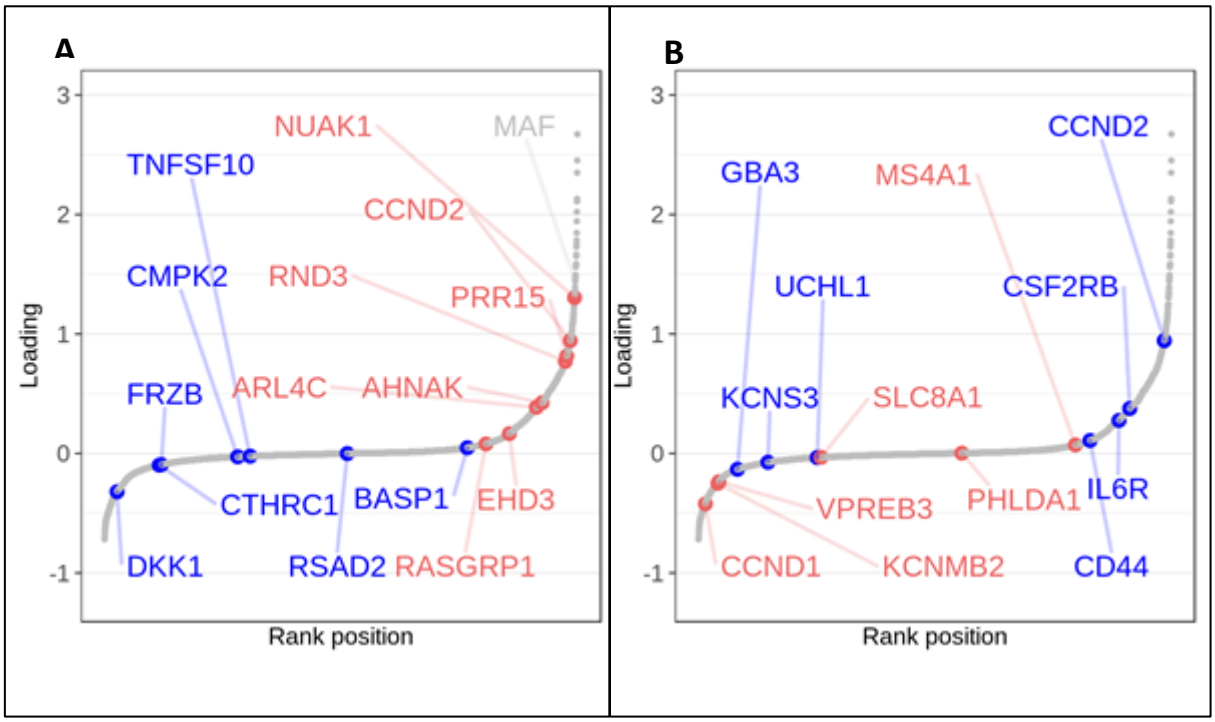


Figure 5-18: LF5 gene weight loading ranks with MAF and CCND1 subgroup relevant genes.

MOFA LF5 gene loadings with top 10 over (red) and under-expressed genes (blue) separating from the rest of clusters A) MAF subgroup cluster (labelled "MF") B) CCND1 subgroup cluster (labelled "CD-2" in 320 bone marrow samples (Broyl et al., 2010) (P < .001, false discovery rate < 5%) showing the greatest fold change per cluster in comparison with remaining clusters. In grey, the MAF gene, relevant to the MM MAF subgroup.

As reference, the Transcription Factor MAF gene is shown for the MAF subgroup (Figure 5-18 A in grey label), showing that it is very important in terms of weight with a loading of 1.44 and

ranking in the top 0.5% by LF5 absolute loadings. MAF is OE compared to PC in HD (an average of 18-fold), MMSET (more than 76-fold) and in MAF translocated samples 3,666-fold. As explained in Chapter 4, another deregulated gene, CCND2 (Figure 5-18 A in red) is a member of the Cyclin family and it is involved in cell cycle processes and PC differentiation (Nahar et al., 2011; Tooze, 2013), cyclins activate cyclin-dependent kinases and these in turn phosphorylate their substrates (Chiles, 2004). CCND2 is highly expressed in a significant number of samples in the present study (Figure 5-19). It ranks in the top 2% of genes by absolute LF5 loading with 0.94 and on average it is OE in relation to the healthy state in HD, MMSET and MAF 13, 18 and 135-fold respectively, with 17-fold under expression for the CCND1 subgroup making it a hallmark MAF subgroup gene in the samples examined. As mentioned in section 4.3.12, deregulation of Cyclins is a common feature in Myeloma (Bergsagel et al., 2005; D. Smith et al., 2016) and has been previously studied as a therapeutical target (Tiedemann et al., 2008). Another gene containing a very high positive weight (1.3) is NUAK1/ARK5 (Figure 5-18 A in red) with very high overexpression compared to PC particularly in the MAF subgroup (417-fold), with MMSET being the only other subgroup with 10-fold overexpression. NUAK1 is a serine/threonine-protein kinase of significant therapeutic interest: simultaneous therapeutical inhibition of CDK4 and NUAK1 in MM CLs and primary samples induces cell-cycle arrest and apoptosis *in vitro* (Perumal et al., 2016).

Genes with close to zero LF5 weights are of low importance to this separation. Examples of previously found genes separating a cluster containing MAF translocated samples from the rest (Broyl et al., 2010) are EHD3 (Figure 5-18 A in red) and BASP1 (Figure 5-18 A in blue), found to be over and underexpressed respectively (Broyl et al., 2010). EHD3 is an ATP and membrane-binding protein, in the present study, EHD3 has a LF5 weight of 0.17. Being expressed throughout all samples, including PC and having no significant expression alteration on any MM subgroup it is assigned a low LF5 weight. BASP1 a membrane bound protein with LF5 weight of 0.05, has a non-significant underexpression in all subgroups compared with PC.

The DKK1 gene (Figure 5-18 A in blue) is found to inhibit the Wnt signalling pathway through repressing LRP5/6 – Wnt interactions. Recently proposed as a biomarker for MM (Feng et al., 2019). In MM murine models, repressing DKK1 – LRP5/6 interaction has shown promising results (Park et al., 2017) and DKK1 has also been shown to be key to inducing MM bone lesions, preventing osteoblastogenesis (Qiang et al., 2008), at least in part by downregulating the miRNA miR-152 which targets DKK1' 3'UTR to repress it (Xu et al., 2015). This gene is highly expressed with respect to normal in HD, CCND1 and MMSET (36, 20 and 11-fold expression) while having a normal-like expression in the MAF subgroup. It has a negative LF weight of -0.32

and signifying that it has a low expression in the MAF subgroup compared with most of the remaining samples.

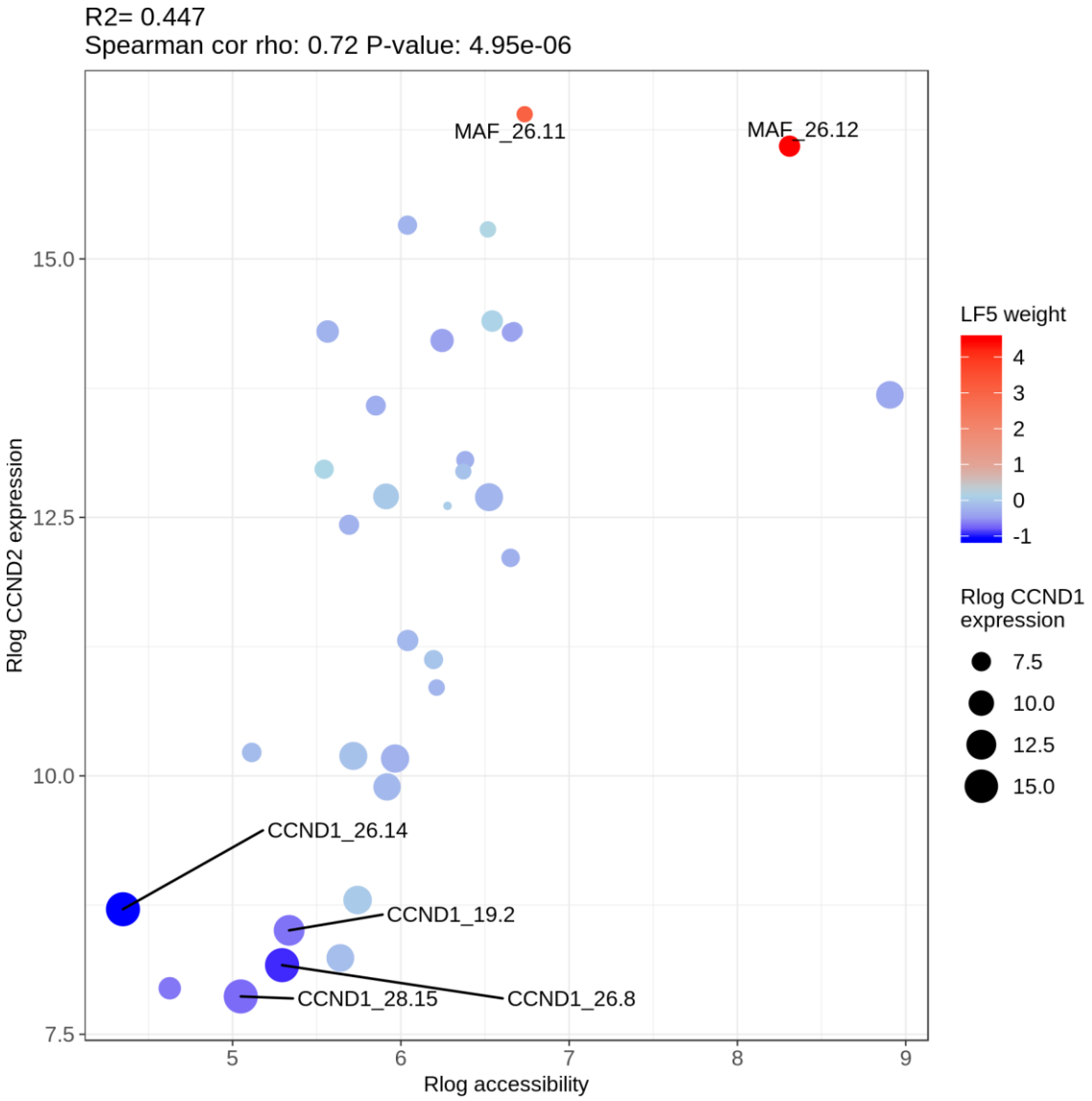


Figure 5-19: CCND1 and CCND2 dichotomy and LF5.

Correlation between normalized rLog CCND2 expression and accessibility for the chr12:4,148,394-4,148,872 region with LF weights shown by colour and CCND1 expression by point size.

As can be seen, in the extremes of the LF5 weights, there is a high correlation between gene expression array data by Broyl and colleagues and RNA-seq from the study in this thesis. In these cases, genes with high positive LF5 weights correspond to significantly higher MAF subgroup expression than the majority of the remaining samples (Figure 5-18 A genes labelled

in red), while very negative weights are assigned to under expressed genes in the MAF subgroup (Figure 5-18 A genes labelled in blue). Furthermore, as it is expected, genes assigned near zero weights (assumed not meaningful in separating MAF samples from the rest) have a constant expression throughout all samples in my study. As it happens with EHD3 and BASP1, this occurs despite the fact that they may be considered significantly altered by Broyl et. al, a disparity that can be caused by the different method of measuring RNA availability.

5.3.6.2. Highly absolute weighted genes significant in CCND1

From the standpoint of the CCND1 subgroup and the top altered genes separating the CCND1 subgroup cluster from the rest in a previous study (Broyl et al., 2010), the deregulated CCND1 gene (Figure 5-18 B genes labelled in red) is OE with respect to PC in HD (16 times average change) and CCND1 subgroup (1,234-fold). Consistent with this, it is assigned by MOFA -0.42 LF5 weight (Figure 5-18 B labelled in red). Another example is the gene KCNMB2 (Figure 5-18 B labelled in red), a gene influencing the calcium sensitivity of Big Potassium channels, has a negative weight of -0.25 and it is not previously found associated in MM. Consequently with the LF5 sample separation, it is overexpressed in all subgroups but MAF with respect to PC: 25-fold in HD, 63-fold in CCND1 and 12-fold in MMSET.

Examples of genes with negligible weight loading for LF5 (-0.03) are UCHL1 (Figure 5-18 B labelled in blue) and SLC8A1 (Figure 5-18 B labelled in red). UCHL1 is a gene with high expression associated to poor prognosis in MM (Hussain et al., 2015), compared with PC expression, it has a 18-fold overexpression in HD and 57-fold in MMSET, while being more than 7 times less expressed in CCND1 and having similar expression compared to PC in MAF. UCHL1 encodes an enzyme that hydrolyses ubiquitin. SLC8A1 is a gene associated to cardiac conduction and mineral absorption pathways; it is expressed throughout all PC and MM samples with no significant differences in MM subgroups with respect to PC. It is possible that since these genes are not separating MAF samples from CCND1 and PC samples by a large factor, it is assigned a close to zero weight.

Genes with positive LF5 weights include CSF2RB (Figure 5-18 B labelled in blue) and IL6R (Figure 5-18 B labelled in blue) with weights of 0.38 and 0.28 respectively. CSF2RB mediates the exchange of calcium ions in cellular processes and it doesn't have any previous associations in literature with MM. The gene has a 25-fold and 3-fold CCND1 and MMSET vs. PC under expression respectively, with the other MM subgroups maintaining PC-like expression. IL6R (interleukin 6 receptor complex) is a cytokine involved in multiple processes such as cell growth, differentiation or immune response. IL6R is a gene very expressed throughout all samples with no significant variability between PC and any of the MM

subgroups. IL6R is located in the 1q21 arm, which appears amplified in MM cases and serves as a prognostic marker (Kim et al., 2011). Overexpression of this gene induces hyper activation of the STAT3 oncogenic pathway (Teoh et al., 2019) and bromodomain and extraterminal domain (BET) inhibitors are shown to decrease IL6R activity (Stubbs et al., 2019).

Genes such as SLC8A1 and IL6R do not have correlating CCND1 subgroup overexpression when comparing data from Broyl's study with my thesis. As it was previously mentioned in the previous section, this can be due to the different methods in RNA availability quantification. Furthermore, there isn't a clear relationship between LF5 weights and over (Figure 5-18 B genes labelled in red) and under-expressed genes (Figure 5-18 B genes labelled in blue) separating the CCND1 subgroup cluster from the rest in Broyl and colleagues study.

5.3.6.3. Highly weighted gene-region interactions

The already established interactions between MOFA accessibility and gene features within 1Mb are studied from a MAF and CCND1 vs. PC analysis (Chapter 4) point of view (Figure 5-20). The MAF vs. PC differential interactions have a greater dispersion in terms of accessibility and expression weights, with CCND1 having an enrichment of interactions containing DE genes with weights around zero. As can be seen in both, the corresponding MM subgroup activated interactions are almost exclusively in the positive and negative weights for MAF and CCND1 respectively, coherently with the MOFA axis separation.

In terms of the MAF subgroup (Figure 5-20 A), a prominent interaction example is CCND2, is a critical gene that as mentioned before, its overexpression is complementary to CCND1's in MM for the practical totality of MM samples. In the MOFA model, it is linked to multiple enhancers overlapping the novel super enhancer region derived from the H3K27ac histone modification enrichment in MAF-translocated JJN3 cells (Alexia Katsarou, Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Foundation Trust in London, *pers. comms*). Furthermore, the interaction shown in Figure 5-20 A (marked with "chr12:4,148,394-4,148,872 CCND2" in black) and Figure 5-19 involving the chr12:4,148,394-4,148,872 region is particularly interesting because it overlaps a tested region for regulation of the CCND2 gene. This is done through CRISPR-dCas9 repression in a MAF translocated cell line by Alexia Katsarou obtaining a significant result (p-value less than 0.05), as can be seen in section 4.3.12. This is consistent with the strong correlation coefficient between accessibility and expression for the primary samples in this analysis (Figure 5-19).

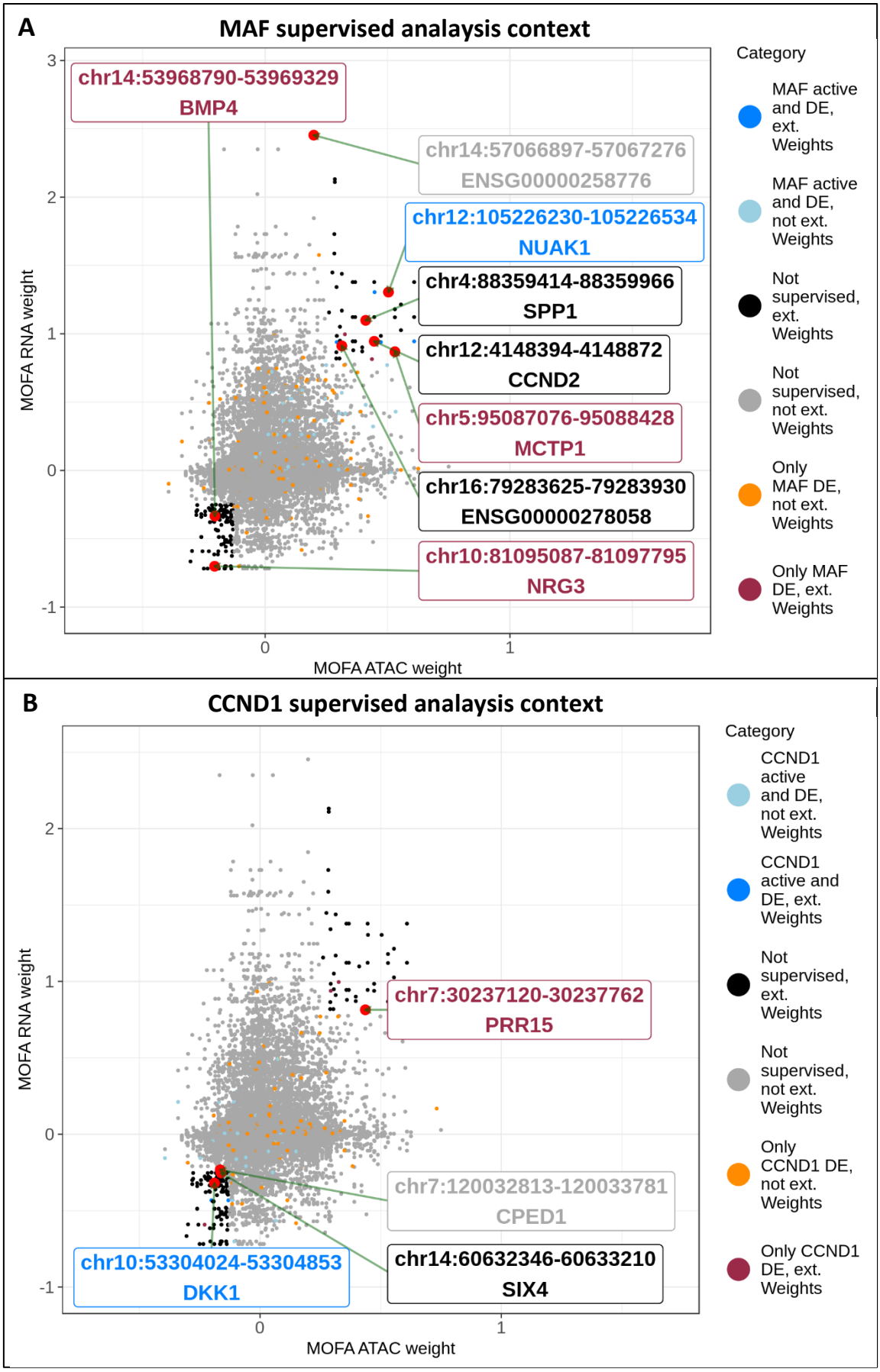


Figure 5-20: MOFA candidate enhancer - gene interactions for LF5 in the context of MAF and CCND1 vs. PC analysis.

Each dot represents an interaction between a MOFA accessibility and gene feature within 1Mb. Different categories are created depending on:

"Extreme weights": LF5 weights being positive and negative simultaneously for both ATAC and RNA: ATAC weight equal or more negative than -0.13 or equal or larger than 0.25, MOFA RNA weight equal or more negative than -0.24 or equal or larger than 0.8).

"Only subgroup DE": MOFA interaction that is also a MM subgroup DASMM enhancer regulating a protein coding DESMM gene.

"Subgroup active and DE": MOFA interaction that is also a SMM enhancer (compared with PC) regulating an OESMM gene for a MM subgroup.

"Not supervised": MOFA interaction not found in the supervised analysis (Chapter 4). MOFA interaction not found in any supervised analysis. A: MAF supervised analysis context. B: CCND1 supervised analysis context.

Furthermore, this region is accessible throughout all samples but more open compared to normal in the MAF samples (4-fold) and repressed in CCND1 samples (3 times). Consequently, the interaction is assigned significant positive weights signifying MAF subgroup and CCND2 gene activation and CCND1 subgroup and gene repression, denoted by the LF5 axis.

There are interactions showing activation both in MOFA and in the supervised analysis with respect to PC. One such example is the enhancer chr12:105,226,230 – 105,226,534 with the gene NUAK1/ARK5 (marked "chr12:105,226,230 – 105,226,534 NUAK1" in blue in Figure 5-20 A). As mentioned before in Highly absolute weighted genes significant in MAF, section 5.3.6.1, NUAK1 is a serine/threonine-protein kinase which has been proven to be an oncogenic target. The interaction is 33 and 417-fold more accessible and expressed respectively in MAF and, consistently, is assigned extreme positive weights in LF5. Another interaction found in the MAF vs. PC analysis is chr5:95,087,076-95,088,428 with MCTP1 (marked "chr5:95,087,076-95,088,428 MCTP1" in magenta in Figure 5-20 A). MCTP1 is a gene in MM that is a MM "spike" gene for 4 out of 206 samples in a study of newly-diagnosed patients (Kassambara et al., 2012), it is not reflected in the study, but perhaps the samples were MAF translocated samples. The region is only significantly more accessible and the gene significantly OE in MAF vs. PC: 14 and 25 times respectively. It is present in the MAF DE analysis but it is not a MAF active interaction with respect to normal because it overlaps an accessible peak region in the healthy state, this candidate regulatory area is accessible in PC but significantly more active in MAF, suggesting that it could be active only in the later subgroup.

A novel active candidate enhancer - promoter pair in MAF found in MOFA that was not found in the previous analysis in chapter 4 is chr4:88,359,414 – 88,359,966 with SPP1 (marked "chr4:88,359,414 – 88,359,966 SPP1" in black in Figure 5-20 A). This gene is expressed in bone marrow tissues and it is involved in diverse functions from bone remodelling to regulating hematopoietic stem cell production. It was found to be regulatory in a variety of cell processes that become deregulated in different haematological cancers such as leukaemia, lymphoma and myeloma (Bastos et al., 2017). This interaction is active in all MM subgroups except MMSET, but is particularly so in MAF translocated samples with an accessibility and expression change over healthy state of 26 and 2,369-fold.

Another example is the lncRNA ENSG00000278058/RP11-70D24.2, associated with the candidate enhancer chr16:79,283,625 – 79,283,930. This lncRNA was recently found to be simultaneously down regulated (and correlated) with MAF expression in CCND1 translocated samples from a study of 30 MM patients (Ronchetti et al., 2018). It is worth noting that the downregulation was calculated by comparing the expression of the 8 CCND1-translocation containing patients with the rest of the samples, out of which, 4 and 7 were MAF and MMSET translocated respectively. This is consistent with the present study: RP11-70D24.2 expression is significantly up regulated in the HD, MMSET and MAF groups (55, 50 and 2,096-fold change over PC) with CCND1 having just a 2-fold expression compared with PC. Furthermore, as mentioned earlier, compared with the normal state, the MAF TF has an 18-fold expression in the HD group, 76-fold in MMSET, 3,666-fold in MAF while having a similar expression to normal cells in CCND1 translocated samples. The associated candidate enhancer region is 3-fold more accessible in HD with respect to PC, nearly twice in MMSET and nearly 7 times more in MAF and 1.5-fold less accessible in CCND1. The LF5 weights for both the DNA feature and gene are extremely positive. Together this interaction is representative of MAF activation and CCND1 deactivation represented by high weights in the LF5 axis and not only does it recapitulate the hypothesis that RP11-70D24.2 might be interacting with MAF (Ronchetti et al., 2018), it extends to the possibility that the means of producing this lncRNA might be through the chr16:79,283,625 – 79,283,930 possible enhancer area.

ENSG00000258776/AL161757.4 is found to be one of the top five long non-coding RNAs (lncRNAs) significantly deregulated in MAF translocated vs. non MAF translocated samples (Ronchetti et al., 2018). In the study presented in this thesis, ENSG00000258776 was found to be DE in the MAF subgroup (and no other subgroup) compared to healthy PC, having a 2570 fold overexpression in MAF. This makes ENSG00000258776 a key lncRNA for separating MAF samples and a possibly important diagnostic/prognostic marker, as reflected by its significant

high LF5 loading (2.45, ranking second). The candidate enhancer region chr14:57,066,897-57,067,276 is associated to it (marked with the label "chr14:57,066,897-57,067,276 ENSG00000258776" in grey in Figure 5-20 A), it is accessible throughout PC and MM but more significantly accessible in the MAF subgroup (5-fold).

NRG3 is a gene under expressed in MAF and located in the negative weights in LF5. It belongs to the neuregulin family, which encodes proteins that activate transmembrane tyrosine kinase receptors ERBB3 and ERBB4, activating intracellular signalling and cellular responses. This is a key process in the majority MM patients (Mahtouk et al., 2006). NRG3 was previously found to be OE in MM vs. bone marrow PCs (Mahtouk et al., 2010). In the analysis presented in this thesis, it is down-regulated only in the MAF subgroup compared to healthy state (by a factor of 92), with CCND1 having a 3-fold overexpression with respect to PC. Perhaps, in MAF translocated samples, this mechanism is replaced by an alternative one resembling the small ratio of MM patients found in the mentioned studies for which the process involving NRG3 is not critical. NRG3 is associated with the nearby region chr10:81,095,087-81,097,795 (marked with the label "chr10:81,095,087-81,097,795 NRG3" in magenta in Figure 5-20 A) which has half the accessibility in MAF that it has in PC. Consistent with this, the both the gene and the region have extreme negative weights and it is also found to be significant in the MAF vs. PC differential analysis.

Finally, another repressed gene – enhancer interaction in MAF is BMP4 (bone morphogenetic protein 4) and chr14:53,968,790-53,969,329 respectively (marked with the label "chr14:53,968,790-53,969,329 BMP4" in magenta in Figure 5-20 A). This interaction is activated in all the MM subgroups except MAF (2 times less DNA openness and 36-fold less expression). BMP4 is previously found in the LF3 axis to be OE in MMSET, it is also OE in the CCND1 and HD subgroups compared to PC. Since bone morphogenetic proteins are known to protect MM cells against Bortezomib treatments (Grčević et al., 2010), it is possible that this pathway is deactivated in the MAF subgroup.

From the CCND1 subgroup supervised analysis point of view there are also biologically relevant interactions with corresponding LF5 weights (Figure 5-20 B). One such example having negative weights and mainly signifying MAF subgroup repression (and secondarily CCND1 sample activation) is chr14:60,632,346 – 60,633,210 with SIX4 (with label "chr14:60,632,346 – 60,633,210 SIX4" coloured black in Figure 5-20 B). The region is active throughout all samples in the study but at least 2-fold more active than in the normal state in the HD, CCND1 and MMSET groups and 1.5 less accessible in MAF. The corresponding gene expression is

consistent with the candidate enhancer accessibility being only repressed in MAF compared to normal 4-fold and OE in the rest of the MM subgroups.

Also, the gene CPED1 is relevant to MM biology, being OE in MM cells when applying *in vitro* combined treatment containing filanesib, pomalidomide and dexamethasone which induces cell death in the disease state (Hernández-García et al., 2017). It is also OE on average in Smoldering Myeloma (a MM pre-malignant condition) patients progressing to MM than in slower or non-progressing MM patients (Storti et al., 2019). In the present study, CPED1 is OE with respect to normal in the HD, CCND1 and MMSET groups, while having nearly half the expression in MAF compared to normal. The associated region to this gene: chr7:120,032,813 – 120,033,781 (marked with "chr7:120,032,813 – 120,033,781 CPED1" in grey in Figure 5-20 B), has a similar normal-like accessibility in MAF while having six or more times more openness in the rest of the MM subgroups and being assigned corresponding negative weights. Assuming CPED1 or other genes' expression with meaningful weights assigned to this separation are associated with the mentioned therapeutical benefit it may be interesting to study if different MM subgroups might benefit from different drug combinations.

The DKK1 gene (referred to in Highly absolute weighted genes significant in MAF, section 5.3.6.1) is associated with multiple candidate enhancers, with a prominent one being chr10:53,304,024 – 53,304,853 (with label "chr10:53,304,024 – 53,304,853 DKK1" coloured blue in Figure 5-20 B), which is more accessible in all MM subgroups except MAF. Coherently with this, it is elucidated in the supervised analysis as a CCND1 vs. PC interaction and contains extreme negative weights (showing MAF repression).

The knowledge of the PRR15 gene is still at its infancy. It is known that it is "expressed almost exclusively in post mitotic cells both during fetal development and in adult tissues" (Meunier et al., 2011). In the present study, it is found to be expressed throughout the normal and disease condition with an overexpression compared to normal in HD, MAF and MMSET of 5, 15 and 4-fold respectively and a 66 fold repression in CCND1. This is consistent with the fact that through RNA microarray gene quantification it was found to be one of the top 10 genes OE in IgH translocated samples with MAF t(14;16) and MAFB t(14;20) vs. the rest (Broyl et al., 2010). PRR15 is associated in MOFA to the chr7:30,237,120-30,237,762 region (label "chr7:30,237,120-30,237,762 PRR15" magenta on Figure 5-20 B). The region is open throughout the PC and MM state but particularly accessible in the HD and MAF subgroups (4 and 20-fold with respect to normal respectively), perhaps, these are the only subgroups surpassing the activation threshold in terms of accessibility. The promoter – enhancer

relationship is found to be differential in PC vs. MAF supervised analysis too, despite this, it is not considered a MAF activated enhancer regulating a MAF OE gene in the supervised analysis (chapter 4) because the region is considered to be accessible in PC too. This interaction is assigned a high LF5 weight, consistently signifying a high activation in MAF and repression in CCND1.

Together these results point at LF5 marking a distinction between MAF and the rest of the samples (whether CCND1 translocated or not) with positive LF5 weights signifying MAF activation.

5.4. Discussion

As can be seen in this unsupervised analysis given the primary samples data used, while gene expression only discriminates the cancer from the normal state, the most effective way to segregate PC from the different MM subgroups is to combine accessibility and gene expression data. This points at a genuinely distinct subgroup profiles in terms of enhancer - promoter interactions. Interestingly, the only exception is the HD group, which is separated from the rest of the subgroups more profoundly with RNA-seq data than by adding chromatin accessibility to it (when classifying based only on the first 5 LFs). A possible reason for this may be that HD is more heterogeneous in terms of chromatin accessibility with different enhancer activity changes that end up in the same gene expression changes. Perhaps, the Myelomagenesis-associated changes occurring in the HD state cause a large number of changes at the chromatin level, but most are unimportant.

Quality control and experimentation with different input features and samples combination was performed when running the MOFA model, in particular, sample RS_4.25 was the only sample in its batch and was assigned the average PC gene expression for each gene. Possible solutions to this were considered, for example, removing batch effects including MM CLs samples (which contained samples in the same batch as RS_4.25), but this would assume the subgroup effect to be the same for a cell line and primary sample based solely on IgH translocation partner or considering all cell lines belonging to the same independent subgroup (having the same effect independently of their cancer driving translocation). Since none of these assumptions seemed rational in the current subgroup analysis context, I evaluated whether not accounting for batch effects or removing sample A28c.14/RS_4.25 yielded similar results as the current model used, a hypothesis that was confirmed. Thus, due to lack of time to redo all the analyses presented here, this sample was included in the analysis.

The 5,000 most variable regions and genes were selected as MOFA input features, while the number of starting MM and PC consensus accessible regions after removing TSS and gender chromosomes is 273,216 and there are 55,110 filtered quantified genes, selecting around 9% of all available genes but only less than 2% of the regions. As a reference to check the selection criteria, the feature means for accessibility (Figure 5-2 A) and gene expression (Figure 5-2 B) were studied in the context of LF1 which produces an approximate MM vs. PC separation (Figure 5-3) explaining more than half of the accessibility and a significant proportion of the expression sample variability (Figure 5-4 A). In both cases, there is a significant proportion of features with means close to zero and features that do not overlap with DEMM genes and MMPC enhancers (this is particularly true for the RNA-seq data). These features are likely to be of low biological relevance and should likely be removed. Nevertheless, there are 33 primary samples divided into different subgroups of very variable number of samples, having each group influencing the mean of a feature in an unbalanced way. For example, it is possible for a region or gene to be highly accessible or expressed respectively in one subgroup (for example in both of MAF samples), while still having negligible values in the remaining samples and therefore having a low overall mean while still showing a biologically relevant MAF subgroup effect. Taken together, the number of features inputted to MOFA (for gene expression particularly) could be reduced without affecting the biological conclusions if the aim is to classify samples. Conversely, due to having 5 times more accessible features than quantified genes, it is possible that some accessibility features with biological meaning can be lacking in the model.

In a similar vein, there are no MOFA analysis regions that are more accessible in PC for the MMPC enhancers (Table 5-1). This is likely due to two main reasons: the first is that there are only 350 out of 9,527 MMPC enhancers that are more accessible in PC (and only 5,000 regions are used in the analysis). Secondly, the study contains 33 samples, out of which only 5 belong to the PC group. The contribution of the variability in the dataset therefore is biased to come from within the MM samples and not between MM and PC, causing a negative selection of variable regions more accessible in PC. In the case of the subgroups, the number of overlapped regions does not entirely correlate with the number of samples. This can be probably due the fact that the regions used in MOFA are coming from the PC and MM consensus peaks, which despite using a balanced per sample sequencing depth strategy to call peaks, the overall sequencing depth per condition is proportional to the samples per condition. Therefore making it likely that in general PC and MM consensus peaks are favoured by the MM condition. For these reasons, it is also plausible that the samples containing higher accessibility

in the feature regions are coming from one or multiple MM subgroups (within the MM samples), this effect can be seen in Table 5-1: the majority of regions are more accessible in MM subgroups than PC. If available it would have been desirable to have more normal donor samples. Short of this, including the same number of consensus peaks being generated from PC and from MM to balance out the dominant accessibility features produced by the 28 MM samples could have been another solution, albeit not a completely unsupervised one in terms of feature selection as it is presently by total sample variance.

In terms of gene expression features there is more balance between the number of genes OE in PC compared with MM or MM subgroups. A possible reason why this is occurring could be the scale of the data, while the chromatin accessibility mean ranges up to 200 normalized read counts (with some outliers having 300), the average gene expression scale ranges is 5-fold higher extending to 1,000 normalized gene counts, with outliers going as far as 30,000 (Figure 5-2). This makes it more likely for gene expression data to have the potential for more variance and therefore more probable that the fewer PC samples could create more overall variability. There are also a higher proportion of PC OE genes (compared with MM) from the total, 258 out of 806 making it more likely that they would be represented in the MOFA analysis, which can also explain this effect.

Genes and enhancer – gene interactions examples are presented considering previous associations in the PC and MM literature, overlap with supervised analysis (Chapter 4 and 5) findings and extreme LF weights for each separation. For LF1 and LF2, a MM disease – gene association score (Piñero et al., 2020) is also included where available. For reference, LF3 considers in the analysis a list of genes previously found to separate MMSET from non-MMSET MM samples (Wu et al., 2016) and LF5 takes into account key genes differentiating individual MAF and CCND1 subgroups from the rest (Broyl et al., 2010). Examples (whether complying or not) with the above criteria are provided for completeness.

The first two LFs establish a change in chromatin openness and overexpression associated with the Myelomagenesis transition. LF1 and LF2 explain a dominant accessibility and expression variation ratio respectively. LF1 reflects a general chromatin opening (or becoming even more accessible) going from the normal to the cancerous state which is not always accompanied by overexpression. As it was found in Chapter 3, chromatin accessibility at promoters is necessary but not sufficient in general in gene expression and similarly, enhancers may need activation (for example through TF binding or acetylation). Gene repression also occurs, this could be an outcome of additional mechanisms such as genetic mutations (Mantovani et al., 2019),

epigenetic changes (Baxter et al., 2014) or changes in chromosome conformation partially mediated by *CTCF* (Braccioli and de Wit, 2019) to name a few. Moreover, the top 500 LF1 negative-weighted genes (MM activated) have enrichment for multiple ontology categories including the bone morphogenetic protein (BMP) signalling pathway, previously found to be enabled by interferon genes in MM (Takaoka et al., 2008). Furthermore, enrichment for epigenetic and chromatin remodelling categories including conformation changes or DNA and nucleosome packaging, which may affect accessibility, are also discovered. This is in line with why, as opposed to LF2, this LF explains a very significant amount of chromatin accessibility variability (reflecting increases in the cancer state). Additionally, post-transcriptional gene repression terms are also found which may help to explain why some genes are more expressed in PC, despite there not being significantly PC over-accessible candidate enhancers. Also, cancer and MM-specific categories such as negative regulation of apoptotic signaling pathway and positive regulation of proteolysis respectively, the latter has been targeted in MM through proteasome inhibitors (Csizmadia et al., 2016). 173 interactions recapitulate some of the 311 pairs of MM enhancers near OEMM protein coding genes previously found in Chapter 3 such as chr8:106,524,326-106,525,957 and chr8:107,457,615-107,457,954 with ANGPT1, a MM up regulated gene also in another study (Munshi et al., 2004) or novel functional relationships such as chr5:129,069,040-129,069,757 with CHSY3 also found to be MM OE (Zhou et al., 2009).

The LF2 axis has a distribution of chromatin features centered on a 4-fold chromatin accessibility enrichment in MM compared to PC, reflecting an axis going from regions being approximately equally accessible in MM and PC to extreme opening of chromatin in cancer. Over and under expression in MM is found in gene expression features, having a very significant anti-correlation (Spearman's correlation coefficient -0.827) between feature loadings and MM vs. PC fold changes. Despite there being no regions that are less accessible in MM than in PC, it is possible that, as in the case of LF1, additional mechanisms could be repressing genes in MM. A representative example in LF2 is chr15:33,915,253-33,916,443 with the FMN1 gene being 5 times more accessible and 11-fold OE in the cancer condition, this region is not classified in chapter 3 as a SMM enhancer because it is considered accessible in PC (albeit less than in the MM state). The top 500 genes with extreme negative weight have enrichment for categories mainly involving neuron and synapses processes, perhaps pointing at this gene-driven axis altering these pathways. This could occur for example through a known gene involved in neuron migration: Reelin (RELN), which becomes very active in MM through hypomethylation of its promoter (Lin et al., 2017a).

LF3 separates MMSET samples from the rest (including the cytogenetically unannotated A17.5 sample which has a high NSD2/MMSET and was implied to be a MMSET subgroup sample in chapter 4). There is a moderate anti-correlation in terms of weights and MMSET vs. PC log_2FoldChanges for both MOFA regions and genes. Furthermore, for genes, the anti-correlation only occurs when taking into account MMSET differential features compared with PC. The weights of genes previously found to separate MMSET from non-MMSET MM samples (Wu et al., 2016) are distributed on the LF3 weight extremes although as it would be expected, although this is not always the case. This can be due to important differences existing between the MOFA analysis and the previous study. Measuring of the RNA availability in the previous study was done through gene expression microarray while in this thesis it was done through RNA-seq. Additionally, the analysis by Wu and colleagues classified MM samples into MMSET and non-MMSET, while the I analysis I presents further includes PC samples in the non-MMSET group.

Also, as can be seen in Table 5-1, there are more total DE genes than regions 1,130 compared to 662, and more balanced in terms of being more activated in MMSET (645 more expressed genes and 647 more accessible regions) compared to activated in PC (485 genes and only 15 regions). A reason for this unbalance in PC more accessible regions might be that perhaps this axis has a higher representation of regions more activated in other MM subgroups compared to MMSET (and not so many PC more activated than MMSET regions) in comparison to genes. LF3 axis delineates novel activation and repression of MMSET subgroup specific promoter – enhancer interactions, including but not limited to MMSET vs. PC interactions previously determined in chapter 4.

Gene Ontology analysis for the top absolute LF3 weights revealed an enrichment in chromatin binding. NSD2/MMSET has been previously associated with transcriptional elongation: MMSET mono and dimethylates the histone H3K36 (H3K36me1/2) (Kuo et al., 2011; Martinez-Garcia et al., 2011), proposed to be required by SETD2 for its role in transcriptional elongation (García-Carpizo et al., 2016). Perhaps overexpression of NSD2 hyper-methylates H3K36 thereby enabling genome-wide SETD2 binding. Furthermore, in MM, these changes in methylation are also associated with recruitment of the histone methyltransferase EZH2 (Popovic et al., 2014) hence why this axis contains chromatin binding enrichment.

LF5 is a MAF subgroup guided axis, separating MAF samples from all the rest (including PC) and having CCND1 samples on the other extreme as a consequence of this effect. The MAF subgroup vs. rest separation top over and under expressed genes (Broyl et al., 2010) are

consistent with LF5 weights in the great majority of cases studied occupying positive (Figure 5-18 A genes labelled in red) and negative LF5 weights (Figure 5-18 A genes labelled in blue) respectively. This is not the case for the CCND1 subgroup genes.

The dichotomy presented in CCND1 and CCND2 expression is reflected in the LF5 separation, having CCND1 assigned a negative and CCND2 a positive LF weight. From higher to lower MM subgroup vs. PC overexpression, the key gene MAF is OE in the MAF, MMSET, HD and CCND1 subgroups. Importantly, this relationship is also maintained in terms of subgroup sample LF5 weights: with MAF having the most positive, MMSET following with close to zero weights, HD samples having negative weights and CCND1 being on the negative extreme. Furthermore, CCND2 expression (linked to MAF) seems to have a similar pattern of expression compared to LF5 weights as MAF (Figure 5-19). Moreover, the HD subgroup samples are heterogeneous: with high MAF and CCND2 expression correspondingly associated to high LF5 weight and high CCND1 linked with negative LF5 weights. This points at the high importance of MAF linked to CCND2 expression simultaneously with the CCND1 – CCND2 dichotomy in the LF5 axis. Additionally, it is possible that MMSET and MAF subgroups converge in over-expression of CCND2 through common routes such as the lncRNA ENSG00000278058/RP11-70D24.2 gene with the candidate enhancer chr16:79,283,625 – 79,283,930 (active in MMSET and MAF). This interaction is marked as "chr16:79,283,625 – 79,283,930 ENSG00000278058" in black in Figure 5-20 A. Also, through different routes such as chr14:53,968,790-53,969,329 with BMP4, which is active in all subgroups but MAF. Novel MAF specific active or repressed interactions are found such as chr5:95,087,076-95,088,428 with the new MM gene MCTP1, only reported in MM as a spike gene in few samples (Kassambara et al., 2012) but significantly more active than in the normal state only in the MAF subgroup. Conversely CPED1 and its associated chr7:120,032,813 – 120,033,781 enhancer are repressed only in the MAF subgroup (with respect to normal), thereby having possible prognostic and targeted therapeutical opportunities for these patients.

Gene Ontology analysis for the highest absolute LF5 weights indicates an under-representation of integrin binding and signalling receptor binding in these genes. Since LF5 assigns weight loadings close to zero to the MMSET samples, perhaps features that are distinct in this subgroup compared with the rest of samples (including CCND1 and MAF subgroups genes) are also assigned close to zero weights. For example, this could underrepresent in the Gene Ontology analysis genes involved in some pathways (including integrin signalling) necessitating of the histone methyl-transferase ability of MMSET that makes the chromatin more accessible in general for other elements to bind (Martinez-Garcia et al., 2011). In this vein, integrins are

involved in signalling pathways between the extra-cellular matrix and cells contained within, relating to functions such as cell adhesion (Hynes R O., 2002), migration, proliferation, differentiation, survival and invasion (reviewed in Brown and Marshall, 2019). In MM, it has been found that MM cell adhesion is promoted by a Reelin/β1 integrin pathway (Lin et al., 2017b, 2016), integrin-related cancerous properties such as bone marrow invasion, cell adhesion and migration have also been confirmed, making Integrin-β7 (ITGB7) a therapeutic target (Neri et al., 2011). Under representation of signalling receptor binding could also be due to a possible MMSET regulated signalling pathway involving CDC42BPA/CDC42, determined by pathway analysis of MMSET affected genes (Xie and Chng, 2014) established by knockdown of MMSET and MMSET re-expression in MMSET knocked down MM cell lines (Martinez-Garcia et al., 2011). As it was seen, LF3 distinguishes the MMSET subgroup, section 5.3.4, CDC42BPA/CDC42 is a gene particularly OE only in the MMSET translocated samples (18-fold higher expression than in PC) and under expressed with respect to normal in the MAF and CCND1 subgroups (13 and 4 fold respectively).

In the past, different studies have approached the deregulation of enhancers in MM from different angles. For example, structural variations such as focal copy number alterations were analysed and it was established that translocations involving the IgL locus were present in a percentage of patients of all cytogenetic subtypes and with different transcriptome profiles and associated with worse prognostic outcome (Barwick et al., 2019). Furthermore, the study suggested that IgL translocations should be used as an independent marker for prognosis and that the IgL locus contained a very active superenhancer in the MM1.S cell line marked by extensive H3K27ac and IKZF1 binding.

A particular case of the E2F-DP1 dimer acting as a TF and BET co-activators was studied finding E2F-DP1 binding MM1.S and U266 MM cell lines promoters but not enhancers and BET binding to both (Fulciniti et al., 2018). Also, it was determined that general increases in DNA methylation and decreases in chromatin accessibility and gene expression were changes occurred in MM cell lines resistance to therapy (Dimopoulos et al., 2018).

On the basis of recurrent mutations in candidate enhancers being associated with target gene expression changes, 6 MM interactions (recalled in chapter 3) such as chr9:37,375,172-37,395,282 with PAX5, chr2:165,615,060-165,624,028 with COBLL1 were proposed as enriched in 1% - 6% of the MM cases considered and two MM subgroup associated interactions were suggested (Hoang et al., 2018). The present study reports one interaction overlapping with the one described by Hoang and colleagues with the PAX5 gene: chr9:37,374,916-37,375,220,

correspondingly, the loading for LF1 on the interaction is negative signifying MM activation. Interestingly, the MOFA analysis adds MM subgroup distinction: the interaction is assigned extreme negative weights in LF5 signifying deactivation (with respect to normal) in the MAF subgroup and conversely activation in CCND1 translocated samples, which is consistent with the fold changes in the corresponding subgroups with respect to the normal state. The PAX5 gene has been found to be involved in PC differentiation (Manier et al., 2017) and high expression of this gene is associated with "less-differentiated PCs" in MM (Paiva et al., 2017). Moreover, in other cases despite not overlapping, multiple candidate enhancers are found upstream of the proposed region regulating the COBLL1 gene.

Other studies have identified MM enhancers using different approaches: using 11 primary MM samples (but none with a t(14;16) MAF translocation) and *in vitro* differentiated memory B-cells as PC, approximately 120,000 H3K27ac candidate enhancers were determined and approximately 20,000 MM vs. PC differential enhancers were assigned to target genes within 200Kb up and downstream of each enhancer confined to CTCF boundaries (Jin et al., 2018). Furthermore, MM super enhancers were determined from H3K27ac and ZFP36, PRDM1 or FLI1 TFs associated with MM super enhancers and enhancers were observed through motif enrichment. Additionally, integration of ATAC-seq data using TF footprinting, ChIP-seq, RNA-seq yielded a TF network for MM including TF such as IRF4, FLI1, MYC, IKZF1, RUNX, ETS or MEF2 and relating the found TFs to different samples cytogenetic features for the subgroups available (Jin et al., 2018). Apart from the fact that this study was using a proxy for PCs, subgroup specific enhancer – promoter contacts and pathways were not determined beyond detecting an increase in H3K27ac signal at the translocation breakpoints of t(11;14) involving CCND1 and t(4;14) involving MMSET. Other studies have studied particular enhancers and variations to associate its effect to gene expression such as the rs6877329 allele affecting expression of the ELL2 gene (Li et al., 2017).

Of particular importance is IRF4, a gene encoding a biologically relevant TF in PC and MM (Young et al., 2013), IRF4 is in the top 0.2% in terms of average of the normalized expression from all PC and MM primary samples in the study. Found to be under the control of super-enhancers (Young et al., 2013) and with an inferred key role in the MM TF-gene expression network (Jin et al., 2018). IRF4 is recurrently highlighted throughout this thesis having enrichment for binding sites in MM enhancers regulating OEMM protein coding genes (see section 3.3.11) and SMM enhancers regulating OESMM genes for the HD and MMSET subgroups (see section 4.3.13). Since IRF4 is not significantly overexpressed in any of the MM subgroups compared to normal PCs, it is possible that, at least in the MMSET and HD

subgroups, opening of chromatin at IRF4 bounded sites is activating elucidated SMM enhancers and regulating overexpressing OESMM genes. Repression of IRF4 in lymphoma has been found to significantly elevate abundance of IFI44 (Wang et al., 2014) a gene more expressed in MM in the present study (although not significantly) and having an LF1 weight indicating MM activation. Within MM subgroups, IFI44 is not significantly OE but on the border of being for the CCND1 and MAF subgroups. Taken together, perhaps the mechanism that is present in IRF4-motif enriched SMM enhancers (in MMSET and HD subgroups) is preventing IFI44 overexpression downstream, occurring in the remaining subgroups.

To my knowledge, this study is the first to tackle exhaustively gene regulation in MM subgroups in terms of unique enhancer and promoter interactions from a completely unsupervised standpoint and shows that the cytogenetic subgroups are biologically relevant. This analysis compliments the one performed in chapter 4 in different ways, first, the supervised analysis isolates features in one MM subtype compared with the normal state, while MOFA obtains unique subtype features versus all the rest (extending the requirements on the supervised analysis), adding a weight loading as an importance in the interactions to the separation. Furthermore, it includes non-coding RNAs (proven critical to obtain the separations created by the current model) as RP11-70D24.2. This consistent with the literature: in the present study its associated enhancer chr16:79,283,625 – 79,283,930 forms an interaction active in the MAF subgroup and inactive in CCND1 translocated samples. This is coherent with the downregulation in CCND1 translocated samples and correlation with MAF expression found (Ronchetti et al., 2018).

Also, critical subgroup interactions included here might not be included in the MM subgroup specific interactions for example because they are considered to have regions that are accessible enough to have enhancer functionality in PC although having a significant openness enrichment in a particular subgroup, such is the case of PRR15 regulating chr7:30,237,120-30,237,762 region. Moreover, the choice of thresholds in the analysis in chapter 4 yields some regions included in MOFA but not in the results in chapter 4, such is the case of the regions: chr8:106,745,254-106,745,782, chr8:107,382,945-107,383,638 and chr8:106,597,737-106,598,615 having a MM vs. PC $\log_2$foldChange slightly below the used threshold.

Additionally, there are also candidate interactions in the supervised analysis that are not obtained in the MOFA analysis. This is the case with the CLEC11A gene, having an extreme negative LF3 weight (MMSET active) that ranks in the top 3% by absolute value but not having a MOFA accessible feature nearby. In this case, two candidate enhancer regions are found in

the subgroup supervised analysis, one of them (chr19:51,621,245-51,621,560) being more than 5-fold more accessible in MMSET compared to PC, perhaps the regions associated to this gene are not variable enough to be in the MOFA features and again advocates for maybe including more accessible features in the MOFA analysis.

6. Chapter 6: Discussion

This work aims at uncovering and understanding enhancer biology and how it relates to expression in MM and its different subgroups based on cytogenetic properties: recurrent translocations involving the IgH enhancer with MAF t(14;16), CCND1 t(11;14) and MMSET/NSD2 t(4;14) and the Hyperdiploid condition (HD). It compares 31 quality samples from bone marrow comprising of 28 MM primary samples with 3 PC samples (with CD19-receptor positive and negative variations for two of them) and 5 MM CLs. 9,527 MMPC enhancers and 806 DEMM genes are discovered altered between MM and the normal state, forming 2,698 candidate protein-coding interactions. Moreover, from these, a subset considered MM active only is presented, consisting of 1,462 MM enhancers, 548 OEMM genes and 311 interactions between them. The work further provides novel subgroup resolution into MM gene deregulation biology elucidating in total for all MM subgroups 6,897 DASMM enhancers and 2,749 DESMM genes generating 6,090 interactions and 2,801 SMM enhancers, 1,664 OESMM and 1,419 gene – enhancer pairs. The enhancer effect from multiple candidate regions from the analysis is tested on the key gene CCND2 in a MM CL harboring the MAF t(14;16) feature, showing evidence of enhancer activity. Additionally MM and subgroup-specific candidate TFs associated with the enhancer collections and possible gene pathways deregulated were determined. The combined effects of enhancer and gene expression biology are shown to genuinely separate cancer and healthy samples and distinguish the cytogenetic clinical subgroups involving primary translocation events. Such effects are also quantified to infer their relative importance for future testing.

6.1. Genes other than through translocation events may be deregulated in MM through cis-enhancer activation

Enhancers have previously been identified using their properties such as high chromatin accessibility, TF binding and sequence conservation, histone modifications, general DNA methylation, distance and orientation to the target genes (see section 1.2). Two previous investigations have tried to unveil enhancers and characterize their action in MM. One of them used whole-exome sequencing (WES) and whole-genome sequencing (WGS) for 804 and 765 MM tumor-normal primary pairs of samples respectively and it is based on recurring mutations in MM. It studied enhancer effects in altered gene expression and used naïve B-cell (not PC) interactions as a proxy for MM interactions (Hoang et al., 2018). A second compared 11 primary MM samples (none reported to include MAF translocations) to *in vitro* memory B-cells differentiated into PCs (Jin et al., 2018) through histone modifications.

Hoang and colleagues revealed 221,380 interactions in naïve B-cells (through promoter capture Hi-C), which were distilled to 114 recurrently mutated regulatory regions thought to interact (within 1Mb) with 271 genes in MM. Despite having high starting numbers, when relating the regulatory regions' mutational burden to the target gene expression, only six regions were found to have a linked gene that was significantly regulated in MM and even then occurring in a very small percentage of samples: 1-6%.

The study comparing MM samples to *in vitro* differentiated Plasmablasts (referred to as "PC" from here on) used differential H3K27ac signal to identify "approximately 20,000 enhancers with altered activity", when comparing MM to memory B-cells or PC (Jin et al., 2018). Genes with altered expression between primary MM and PC were also obtained but to my knowledge, the number of deregulated genes is not provided, neither are comprehensive lists for deregulated enhancers and genes. Genes with altered expression within 200Kb of the regions and CTCF bound limits (delimiting TADs) were used to determine functional interactions but to my knowledge, the number of interactions or a comprehensive list of the interactions identified are not reported.

In Chapter 3, enhancers are determined on the basis of chromatin accessibility after removing TSS (except single-exon unannotated TSS, which could be enhancer RNAs). Around 0.1% of the provided MMPC enhancers and MM enhancers are found to overlap the regions intervening in the interactions defined by Hoang and colleagues. Furthermore, less than 1% of the DEMM and OEMM genes determined are recapitulated by the genes contained in the interactions (Hoang et al., 2018). Moreover, candidate enhancer regions determined in the presented work are related to genes within 1Mb linear distance having condition (healthy or cancer) specific altered expression coherent with the corresponding enhancer accessibility to reduce the false positive interactions. None of the proposed MM interactions by Hoang *et al.* are previously recapitulated to my knowledge, however, enhancers associated with COBLL1, PAX5 and ATP13A2 genes are found up or downstream of the previously proposed interacting regions, perhaps appending or replacing the MM enhancer repertoire. In addition, from the 12 offered interaction examples by Jin *et al.* only the gene HGF was recapitulated in this thesis. The corresponding MM deregulated genes are enriched for multiple ontology categories, some already elucidated such as: angiogenesis or cell migration (Hoang et al., 2018; Jin et al., 2018) and new relevant ones like the extracellular matrix structural constituent, which is part of the bone marrow microenvironment, and is of known importance in MM pathogenesis (Glavey et al., 2017). HGF, which shows enrichment in B-cell interactions with proposed enhancers in my work, is thought to be involved in these processes through interaction with osteoblasts and

bone marrow stromal cells (Strømme et al., 2019; Ullah, 2019) and already has been aimed for in treatment (Rao et al., 2018).

Together, these results and previous studies show different mechanisms of action on candidate enhancer regions, whether mutations (Hoang et al., 2018), acetylation (Jin et al., 2018) or accessibility (current study), yielding different enhancer sets (as far as the limited provided data in previous studies was compared) and targeting orthogonal deregulation pathways. Different possibilities might be explaining this, it is possible that the enhancer-activity altering mechanisms are independent, a hypothesis previously backed by studies showing that enhancer sets obtained through different enhancer properties are not significantly overlapping (Kleftogiannis et al., 2015). Perhaps, simultaneous enhancer chromatin accessibility and histone acetylation is not a requirement for enhancer activity, despite the fact that there was some correlation between normalized H3K27ac and accessibility (ATAC-seq) signals in the cell types studied: MM, memory B-cells, plasma blasts and differentiated memory B-cells into PCs (Jin et al., 2018). Maybe new histone modifications may emerge as creating different enhancer sets as it happened recently (Henriques et al., 2018) with some histone modifications being complementary to accessibility. Furthermore, it is reasonable to assume that the *in vitro* differentiated cells do not fully resemble the PCs in this study.

Another possibility explaining the differences in results when considering the enhancer – promoter interactions could be due to the methodology. Previous studies examining effects of enhancer deregulation have used a distance based approach such as the one in this thesis by, for example, employing coordinated changes in accessibility and expression within a certain range in B-cell lymphoma (Koues et al., 2015) or using nearest enhancer – gene pairs in T-cell Leukemia/Lymphoma (Ishida et al., 2017). In this work, interactions have been determined on the basis of non-promoter areas with alteration in chromatin accessibility being within 1Mb in linear DNA range of changed expression promoters. Earlier analyses have determined that 90% of these interactions occur within the used range (Javierre et al., 2016) and 60% of them might be occurring within 20Kb (Kumasaka et al., 2018). Together, this points to my study likely covering most genuine MM vs. PC interactions but also incurring in a high false positive rate of interactions due to the distance-based association. This, however, is mitigated by the fact that only same sign changes between accessibility and gene expression amongst conditions are considered. Jin and colleagues and Hoang *et. al* used more restrictive methods of linking elements than my study: CTCF-determined TADs and B-cell promoter capture Hi-C respectively. If this was a critical factor in explaining differences and the interactions to be

determined were very similar in all three studies, the prior two should be largely contained in my study and this was not the case from the examples compared. Furthermore, on average, I showed that genes thought to be regulated have more candidate regulatory regions in this range than genes in general, pointing at a genuine regulatory effect of the interactions proposed. Despite this, the interactions proposed here should be refined. Another key aspect when considering these studies is that, in contrast to Jin et al, the analysis presented in this thesis also incorporates a broader range of samples including those with the IgH-MAF translocation, while still being limited with respect to the Hoang *et. al* study, particularly in the number of PC samples (only 3 patient samples available).

There are 18,339 DAMM regions, but strikingly only about 2% are more chromatin accessible in PC, pointing at a general MM opening of chromatin. One possible reason explaining this could be the difference in the number of samples used to produce the peaks: 28 MM vs. 5 PC, despite keeping sample sequencing depth constant, the total sequencing material used to produce each condition's consensus peak set is more than 5 times higher in the MM condition. This difference in condition-specific sequencing depth could account for a much higher number of MM-specific peaks driving MM over accessibility whether overlapping or not PC consensus peaks. Another possible option explaining these results could be that PC samples could be more homogeneous in terms of chromatin accessible areas compared with MM samples. This is something feasible given that the unsupervised MOFA analysis in chapter 5 has around 40% of its variability explaining LFs producing sample splitting within the cancer state. The per sample number of broad peaks called for PC is higher than for primary MM samples: 75 and 60 million respectively and peak calling signals have been visually verified for various cases. This makes it unlikely that bone marrow PC samples suffer from more complex ATAC-seq signal distribution (perhaps diffused) due to biological or technical artefacts but this could also be in effect. To further investigate this, a condition specific sample – consensus peak saturation analysis would provide additional insights regarding each sample's heterogeneity and contribution to an estimated maximum number of accessible peaks. Ultimately, a future experiment adding PC sequencing depth prior to peakcalling to balance out cancer and healthy state accessibility would be desirable, in the hopes of deepening the knowledge on PC over accessible regions.

This work elucidates *in vivo* and *de novo* Myeloma vs. PC and MM subgroup specific enhancer – promoter interactions essential for the oncogenic state, as well as their corresponding activation proteins and pathways enriched. The deregulated enhancers provided in my thesis represent to my knowledge, the first comprehensive and extensive list of its kind. This

catalogue lays a foundation to further incorporate other information regarding enhancer characteristics such as acetylation, observing possible overlaps and testing the candidates through one of the methods outlined in section 1.4.

Interactions in the PC – MM transition can also benefit from further improvements that could be applied to provide the interactions with more confidence such as considering using only interactions within TAD boundaries as it was previously done in MM using CTCF as a proxy (Jin et al., 2018). Furthermore, another option would be addition of 3C promoter interactions in B-cells data (Hoang et al., 2018) as it was carried out with the CCND2 enhancer region (see section 4.3.12) until PC or MM interaction data is available. Moreover, this could also be reinforced by providing statistical correlation between accessibility and gene expression, something previously employed to determine enhancer effects in different cancers (Corces et al., 2018) and ependymoma (Pajtler et al., 2019), which may provide more than 62% interaction Hi-C validation when using other assays (Mckeown et al., 2017). Future analysis should therefore include only ATAC-seq and RNA-seq statistically significant correlations within Bcell TADs, for example using inTAD software (Okonechnikov et al., 2019).

Testing of the elucidated interactions through 3C or CRISPR-dCas9 repression should be the next step to determine the validity of the hypothesis. Clinical relevance of the suggested features should be studied through MM cell line viability after intervention. The reason for this being that 3C techniques require a high number of cells to obtain cell averages of interactions and primary samples do not always comply with the requirements. Alternatively, single cell assays can be used to test the findings.

6.2. MM developmental enhancers

Once the condition-specific enhancers in Chapter 3 are elucidated, they are next observed in the context of other cell types and in particular the B-cell lineage during PC formation. In the area of MM enhancer development in the B-cell lineage, first an investigation found 794 regions with active enhancer histone modifications and DNA accessibility in B-cells and having potential for activation in PC, while being hyper methylated, inactive regions in MM and ESCs (Agirre et al., 2015). Methylation of these candidate enhancers was inversely correlated with associated gene expression. Jin *et al.* also studied the chromatin state in the GM12878 (normal B-cell line), memory B-cells, plasma-blasts and PCs. The main finding was that there was an enrichment for possible activation of MM enhancers in B-cell heterochromatin, although no comprehensive list of regions was provided. Another prior study tried to profile candidate

enhancers at 16 developmental hematopoietic stages showing that nearly the totality of the discovered enhancers altered the state during development and a high conservation of enhancers was also found within the lineage (Lara-Astiaso et al., 2014). Furthermore, enhancer activation was primarily occurring in the early stages of differentiation.

Different enhancer sets in the B-cell lineage were determined in my study from the catalogue of 1,959 unique regions obtained from the novel 2,698 pairs of MMPC enhancers regulating DEMM protein coding genes in 173 other cell types (healthy and diseased). The assigned chromatin states on these MMPC enhancers grouped samples by cell tissues, a phenomenon supporting the cell type specific enhancer programs in haematopoiesis (Lara-Astiaso et al., 2014) and suggesting the relevance of these candidate regions. Furthermore, in line with Jin and colleagues' findings, most of the MM enhancers studied in healthy tissues were inactive in B-cells, with around one third of enhancers unique to MM. Out of the set of regions that were simultaneously functional in PC, B-cells, MM and other cell healthy cell types, a subset was marked as regulatory in neutrophils, in line with MM morphological characteristics previously found to overlap (Wei Wang and Shimin Hu, 2015). Additionally, other regulatory programs overlapping with MM occur in multiple other tissues such as macrophages within the myeloid lineage or T-cells. Around half of all studied MM enhancers are novel in terms of the disease tissues studied, with about a quarter of the enhancers active in most studied disease tissues and a small proportion appearing like they could be shared between PC and MM.

Together, the analysis suggests that in general the candidate regulatory regions have chromatin state reflecting MM enhancer activation. Within this context, there is a predominance of MM enhancers that are inactive on other cell types, as it was found to occur in early stages of the normal hematopoiesis process. Furthermore, MM enhancers being active on other cell types might advocate for overlapping causal mechanisms at this stage of blood cell type differentiation, considering MM as a pluripotent-state-like in the blood lineage. Moreover, a significant number of regulatory programs seem to be shared with other cell types, providing novel insights into the mechanisms for possible known gene expression pathways. Another possibility is that the same enhancers in multiple cell types target different genes, for example through TAD-boundary disturbance. This should be further investigated linking the different types of enhancers to their target genes using methodologies previously mentioned (see section 1.6.2) and comparing them to the expression programs of neutrophils, T-cells and other cell types sharing enhancers.

These results expand the current MM enhancer knowledge in terms of the spectrum of tissue types investigated in the context of the enhancer set produced. This region set however is comprised primarily of MM more accessible regions (98%). The low abundance of PC only enhancers (33 out of 1,959) in this analysis means that the loss of enhancer activity in MM is probably under represented and should be addressed in future work. This is reflected by the fact that no visible B-cell activated enhancers appear decommissioned in MM as it was shown in a prior study (Agirre et al., 2015). In this regard, it is thereby advisable to follow the recommendations (referred to in the previous section) in future studies to elucidate more PC active regions. Furthermore, since the chromatin states are derived from histone modification data, the suggested hypothesis and candidate regions should be tested for enhancer activity in the pertinent tissues activity through CAGE and perhaps linked to target genes using Chromosome Conformation Capture.

6.3. The role of promoter accessibility in PC and MM gene expression

Previous publications have suggested that additionally to promoter accessibility, other mechanisms are required for gene expression (Klemm et al., 2019). Such means include TF binding, enhancer assisted transcription (Ampuja et al., 2017), histone modifications (de la Torre-Ubieta et al., 2018; Lara-Astiaso et al., 2014), DNA methylation or mRNA processing mechanisms such as alternative polyadenylation (Misiewicz-Krzeminska et al., 2016) and play a pivotal role. Moreover, a recent paper (Starks et al., 2019), linked different types of correlation between promoter accessibility and high gene expression to housekeeping (high accessibility) and tissue-specific genes (medium or low accessibility) in multiple cell types and tissues, and identified cell fate decisions based on promoter repression. Together, this suggests that, at least on a significant number of cases, promoter accessibility is necessary but not sufficient for gene expression and in part of the remaining situations, it may not be necessary at all.

Apart from studying enhancer-driven gene expression in MM and other tissues, my work tries to quantify the impact of promoter accessibility in this process in Chapter 3 and determine whether its activity is critical regarding gene deregulation. In this context, this thesis shows that overall, more genes have an accessible promoter in MM, consistent with the general opening of chromatin at enhancer regions in the cancerous state. For example, for both MM and PC, there is a substantial difference between accessible promoters and those considered to have a high expression. Furthermore, the great majority of genes having low expression in both conditions have an accessible promoter. Moreover, genes in open chromatin tend to be

lowly expressed in the majority of cases. Together, this suggests that open chromatin at the promoter is not sufficient for high gene expression, advocating for other mechanisms such as enhancer assisted transcription (Ampuja et al., 2017), which was studied in this thesis.

When studying the 264 protein coding OEMM genes, the majority of promoters are accessible both in PC and MM, with 20% of genes having repressed chromatin at the promoter. Additionally, there is no significant difference in the proportion of these genes being non-accessible (in each condition individually) compared to any gene (MM OE or not). Proposing that protein coding MM overexpression does not make a gene more likely to have the promoter accessible either in PC or in MM. Likewise, PC promoter accessibility is not significantly indicative of the gene becoming upregulated in the cancer state. In fact, from the 264 OEMM genes studied, only 38 have gained MM chromatin accessibility. In line with the previous work mentioned in the field, this advocates for further mechanisms as a requisite for an important fraction of the PC and MM genes to be expressed. Collectively, general gene promoter accessibility is, in the majority of cases, a requirement for high gene expression, despite being a number of exceptions. This could further be validated with available DNase I accessibility and gene expression data for the U-266 MM CL (Agirre et al., 2015) additionally providing a means to evaluate the degree to which surveying of chromatin accessibility through the ATAC-seq assay is representative of the underlying biology.

Additionally, a very small proportion of genes in closed chromatin are also highly expressed. Proposing that promoter closed chromatin generally represses gene expression but not in all cases. According to Starks *et al.* conclusions, these genes may correspond to cell type specific genes and should be verified if there is an enrichment in activated enhancers regulating them in future work. Along the lines of this hypothesis, it would also be desirable to elucidate whether other additional mechanisms may be at play regulating genes, promoter TF binding is a good candidate to be tested with already available data. This could be attained by TF footprinting and then performing ChIP-seq experiments targeting the candidate TFs. It could be further complemented by incorporating different types of data such as methylation and performing unsupervised analysis simultaneously on them and gene expression to get a broader picture in terms of MM gene regulation and cell fate.

Certain limitations regarding the analysis have to be taken into account. For example, a strict threshold of 5 TPM was used to determine if a gene was expressed or not, this threshold is arbitrary and does not necessarily imply that genes below or over that threshold are functionally repressed or expressed respectively. In the same vein, chromatin accessibility peak

calling converts a spectrum of accessibility into a binary outcome. As it has been previously mentioned in this section Starks and colleagues considered promoter accessibility divided into high, medium and low. Furthermore, while some peaks for primary samples in the higher and lower end of sequencing depth have been visually verified, the method used does not have 100% sensitivity at detecting open chromatin. Lastly, when considering OEMM genes, a minimum expression criteria was not applied, being possible that genes are having an extreme low expression in PCs transferring to a higher one (but still low) in MM. As mentioned, future studies establishing promoter contacts through 3C techniques should help to isolate enhancer from promoter effects.

6.4. MM has subgroup specific mechanisms extending the driver initiating event

Once the general Myelomagenesis gene pathways and associated altered enhancers are obtained and compared to other tissues in Chapter 3, the specific mechanisms creating clinical subgroup phenotypes are determined in Chapter 4. Previous attempts at subclassifying MM into subgroups have been performed, mainly using cytogenetics. For example, the translocation/cyclin D (TC) based on IgH translocation partners and cyclin D expression yielded 8 subgroups (Bergsagel et al., 2005). Also, the University of Arkansas for Medical Science (UAMS) classified MM into 7 distinct groups, 4 of them overlapping the cytogenetic classification in this chapter: MMSET [t(4;14)], MAF [t(14;16)] (also including MAFB [t(14;20)]), CCND1 [t(11;14)] and HD/HY (Zhan et al., 2006) later adding 4 supplementary groups (Broyl et al., 2010).

This led to gene expression data being used to classify MM molecular subtypes (Zhou et al., 2012). More recently, two studies in particular have created actionable gene sets for MM subgroups, the first one providing the top 10 over and under expressed genes isolating clusters of 320 bone marrow MM patients into clinical subgroups (Broyl et al., 2010). The second one postulating 71 genes differentially expressed in MMSET translocated MM patients compared with non-MMSET translocated (Wu et al., 2016). Furthermore, other genetic and epigenetic MM classifications have emerged such as DNA mutations (Kuijjer et al., 2018) or methylation patterns (Kumar Mishra and Guda, 2017). In terms of MM enhancers and its interactions at subgroup resolution, only two interactions being significantly occurring in MM subgroups based on translocations were suggested (Hoang et al., 2018). These are chr3:187,635,970-187,636,359 linked to TPRG1 in a very low MM samples percentage (2% of the HD and 3% of MYC translocated subtype samples) and chr3:186,739,608-186,745,052 with ST6GAL1 in 4/109

(4%) of MYC translocated samples (Hoang et al., 2018). Furthermore, Jin *et al.* studied the biological pathways linked to genes associated with inactive B-cell regions thought to become active in MM. It must be noted however, that this study did not contain MAF translocated MM samples. There is a gap in knowledge to be elucidated in MM subgroups at the enhancer, interactions, TF and biological processes level, more so in the MAF subgroup where the knowledge is specially lacking.

Chapter 4, proposes 6,897 novel DASMM enhancers for one or multiple MM subgroups. MAF is the subgroup deviating most from healthy state in terms of accessibility: DASMM enhancers are enriched in MAF-specific regions with around one quarter (1,798), MMSET 297; CCND1 213 and HD 509. In terms of transcriptomic profiles, the MAF subgroup also has the highest fold variation compared to PC in both extremes and MMSET has the most modest. Additionally, accessibility is more subgroup exclusive than gene expression, with this effect being even more pronounced when restricting the analysis to MM subgroup activated regions and genes (compared with PC). This reinforces the idea that to a degree, enhancer biology is more subgroup discriminative than gene expression and thus underpins the clinical importance of studying enhancer action beyond the IgH enhancer in MM. In this vein, MM subgroup changed regions are associated with altered genes in 6,090 interactions. Moreover, 2,801 SMM enhancers are found, with nearly half being MAF specific (1,265), HD having 321, 136 exclusive to CCND1 and 218 to MMSET. When linked to OESMM genes, 1,419 gene – enhancer pairs are revealed.

When comparing with the two interactions proposed by Hoang and colleagues none are recapitulated in my study. Perhaps, since both are associated with a MYC translocated subgroup, this subgroup has samples in different subgroups in my study and there is not a strong enough MM subgroup-specific effect for it to be significant. Another possibility is that since these enhancer – gene connections are based on a small percentage of samples, they may not be representative enough to generate an effect in my study, which contains less than an order of magnitude fewer samples. Finally, it is also possible that orthogonal enhancer sets are created with different regulatory programs. For example, mutations may be enabling regions to bind different TFs (Hoang et al., 2018) and unmodified genomic sequences may be becoming active in my study.

Importantly, the altered enhancer – protein coding promoter interactions between the different subgroups and PCs generate a separate enhancer and gene signature that correctly categorizes samples with prior cytogenetics information into their corresponding MM

subgroups. Furthermore, it also assigns the cytogenetically unannotated A17.5/RS_2.1 sample to the MMSET subgroup, where it is thought to belong given the extremely high NSD2/MMSET gene expression as a probable effect of t(4;14) of the IgH enhancer, thereby suggesting the validity of the signature. Notably, the expression profile expands this information proposing other genes that separate this and other MM subgroups and equally relevantly, chromatin active regions contributing to the separations and their relationship to the genes can be used for future reference. In this context, there is a subgroup specific enrichment of multiple pathways for DESMM genes, some novel like vasculogenesis, which is only found overrepresented in the MAF-translocated samples. Others already known like the cAMP (cyclic adenosine monophosphate) mediated signalling in all subgroups but MMSET subgroup. This agrees with previous findings from the study by Jin and colleagues, which show enrichment in around half of the analysed MM samples (none of them including the MMSET translocation). Moreover, my study extends this information being positive for the MAF translocated samples (lacking in the Jin *et. al* study), with potential therapeutical implications.

Collectively, the results presented in this work describe an exhaustive collection of candidate regulatory regions and their MM subgroup specific interactions with target genes covering a gap in current knowledge and capable of correctly classifying samples. This assumes, however, that all MM subgroup chromatin over accessible areas non-overlapping TSS represent enhancers which is known to not be the case. Similarly to the future work in section 6.1, it is important to perform future validation of the true positive regulatory enhancers and their interactions with promoters through the means proposed. In addition, this assessment should also help to uncover whether there is enhancer-mediated activation of cAMP and other pathways or propose alternative mechanisms to be studied such as TF binding at the promoter through point mutations. Since very few subgroup interactions are available in the literature to compare to the current work, it is important to complement the regions provided in this work with data such as acetylation and other histone modifications found at active enhancers, prior to performing MM subgroup specific CRISPR-dCas9 repression screens. Once the subtype critical regions are confirmed, panels based on enhancer – gene activity, as well as survival analysis and personalized treatment options directed at different pathways can be tested. These should incorporate prior elucidated MM genes already found to affect patient survival such as integrins (reviewed in Brown and Marshall, 2019) or ROBO1 in non-MMSET myeloma patients (Wu et al., 2016).

6.5. The MAF subgroup CCND2 enhancer region

CCND2 is a gene very heavily studied in PCs (Nahar et al., 2011; Tooze, 2013) and MM samples (Bergsagel et al., 2005; Nahar et al., 2011; Tooze, 2013). In general, virtually all MM samples are thought to express either high cyclin CCND1, CCND2 or CCND3, the latter only in a very small percentage of MM patients. While the mechanisms for CCND1 and CCND3 OE have been tracked down (Shah et al., 2018), CCND2 and its regulation mechanism has not been conclusively defined, making it a very attractive target. In this regard, it has been previously proposed in MM CLs that two CCND2 mRNA isoforms are produced and, via alternative polyadenylation, CCND2 overexpression is attained by shortened 3'UTR at the mRNA level that losses miRNA binding sites, thereby preventing repression (Misiewicz-Krzeminska et al., 2016). Furthermore, a reported CCND2 enhancer overlapping the CCND2 gene TSS was reported about 100Kb downstream of the novel elucidated CCND2 enhancer regions found in the present work (Young et al., 2013). Moreover, Misiewicz-Krzeminska *et al.* also concluded that the CCND2 promoter is more methylated in MM CLs with t(11;14) compared to MM CLs without the translocation but expressing CCND2 (including U266 co-expressing CCND1 and CCND2), suggesting that this could, at least partially explain how CCND1 (and perhaps CCND3) completely repress CCND2 expression.

Given the biological relevance of CCND2 in the MM context, after elucidating MM subgroup interactions in Chapter 4, my work investigates in further detail examples of enhancers that might be regulating this critical gene in subgroup-specific manner. In this vein, the genomic area previously reported by Young *et al.* is considered an extension of the CCND2 TSS and this thesis proposes instead interaction of the regulatory enhancer chr12:4,142,636-4,143,378 (part of a larger CCND2 superenhancer with location chr12:4,103,242-4,177,985) with the CCND2 gene which is very highly expressed in MAF samples (and to a lesser extent, some HD and MMSET). The study in this thesis has provided evidence from multiple angles that suggests this novel candidate enhancer might be at least partly responsible for CCND2 overexpression in the corresponding samples. For example, there is a high correlation between its accessibility (which is also relatively high in some HD and MMSET samples) and CCND2 expression throughout all samples in the study and B cell line (GM12878) Hi-C data reflects contacts between the CCND2 promoter and enhancer. More compellingly, it has been shown by Alexia Katsarou (Department of Haematology, Hammersmith Hospital, Imperial College Healthcare NHS Foundation Trust in London, *pers. comms*.), that CRISPR-dCas9 repression of this region in a MAF translocated cell line (to my knowledge JJN-3) yields significantly lower relative CCND2 expression akin to CCND2 promoter repression (p-value is less than 0.01). Moreover, super-

enhancer calling using H3K27ac ChIP-Seq data in this cell line also identified this region as significant. This region has been experimentally confirmed to be correlated to CCND2 expression in MAF translocated cell lines. The next step to further confirm the finding could be to obtain similar 3C data for the MM MAF-translocated CLs to identify contacts between the suggested region and the CCND2 promoter and to further assess regulatory function in primary MM samples harbouring the MAF translocation. It is important to note, that although Misiewicz-Krzeminska *et al.* also found the high CCND2 expression to be a predominantly MAF-related, the MM primary samples analysed in this thesis, also show that this may be occurring in samples from other subgroups such as HD and MMSET. Furthermore, given the fact that only two MAF translocated samples are included in the present study, it should be confirmed whether the CCND2 overexpression occurs in all MAF samples.

Combining the current knowledge provided by Misiewicz-Krzeminska and colleagues as well as the one provided in this thesis, it is possible that an additive regulatory effect occurs in MAF translocated MM CLs augmenting CCND2 expression. This may consist on the increased expression of all CCND2 isoforms by means of the suggested enhancer region, further amplified by a reduced repression of the shorter CCND2 mRNA isoform that evades suppression by miRNAs. This could be tested with the data in this thesis by analysis of the different CCND2 isoforms produced and their ratios for the different subgroups, observing whether a significant difference exists. Furthermore, since CCND1 and CCND2 expression appear to be mutually exclusive in MM and hence, relevant targets, it would be of high interest to know the mechanism explaining why simultaneous expression of both genes doesn't co-occur. In this regard, in chapter 5 it is discussed how MMSET and MAF subgroups might converge in over-expression of CCND2 through common processes such as activation of lncRNA ENSG00000278058/RP11-70D24.2 with MMSET and MAF SMM enhancers. Also, through different routes such as BMP4 activation occurring in MMSET but not in MAF. Validation and quantification of these mechanisms of action in MM primary cells subtypes including MAF, MMSET and HD should therefore provide valuable clinical insights. BMP4 has already been validated to have 3C contacts in B-cells with candidate enhancers presented in this work.

6.6. A subgroup-specific TF network

After having revised the general MM and specific MM subgroup deregulated enhancers and their effects on gene expression in Chapter 3 and Chapter 4 respectively, a probable means of

enabling such processes: through enhancer TF binding, is examined in their corresponding chapters. Earlier studies have tried to address the key TFs involved in MM, for example in MM CLs BRD4 was found to bind at enhancers (Fulciniti et al., 2018). To my knowledge, the only analysis to extensively study TF binding in MM and its transition from PC was done by Jin et al. In this investigation, enrichment of IRF-IRF (Interferon Regulatory Factor) composite, AP-1, E-box, ETS, and MEF2 motifs was found in MM gained enhancers. All but the latter and the OCT TF were also found enriched in MM more accessible regions that were also repressed in B-cells (a proxy for novel MM enhancers). Furthermore, MM super-enhancers that were found in most of the samples were associated with 55 TFs clustered into 4 groups, showing super-enhancer per sample presence in PC, MM, memory B-cells and/or plasma blasts. All clusters contained TFs corresponding to primary MM samples, additionally, 2 of the clusters totalling 27 TFs were also found in *in vitro* differentiated PCs. Moreover, the clusters of TFs also contained MM sample subgroup information for the subgroups: MMSET t(4;14) having only one sample, CCND1 t(11;14) 3 samples, HD 8 samples with one overlapping also with 17p arm deletion (del17p/-17). Despite this, no clear subgroup TF patterns emerged in terms of the clusters elucidated. Jin and colleagues, also built MM and differentiated PCs TF networks based on TF footprinting data (from ATAC-seq), inferring a less interconnected network of TFs in MM. Furthermore, *in silico* integration of TF footprinting on determined super-enhancers and target gene expression data was executed to obtain an extensive regulatory network. Apart from the TFs enriched in gained MM enhancers, other MM relevant TFs such as the IRF family, MYC or IKZF1 were suggested. Also gene families such as RUNX, Fos, Jun or and several ETS factors appeared enriched. To verify the role of IRF4 and FLI1 (from the ETS TF family) in MM CLs, ChIP-seq was performed, showing critical simultaneous binding at candidate enhancers for key MM genes such as IRF4 and PRDM1.

Using the unique candidate MM enhancers thought to interact with OEMM genes, enrichment for 15 TFs binding in these regions was obtained. The TFs found in the present work recapitulate the findings of Jin et al. in the case of the IRF family, POU2F2 or CTCF and extending previously found cases (Jin et al., 2018) for example with CPEB1, BCL6, STAT1, AHCTF1, POU2F1 and CTCFL. BRD4, a previously found TF binding MM CLs (Fulciniti et al., 2018) was not recapitulated by either my investigation or by Jin *et. al.* and perhaps is only detectable in MM CLs. Interestingly, the novel TFs found to be enriched in MM enhancers have a moderate to high expression in the healthy and cancer condition with no significant differences among them. This suggests that if the binding is occurring, a possible enabling mechanism is through gain of chromatin accessibility at DNA regions containing the

corresponding motif sites and not TF availability increase. Experimental testing of this hypothesis (for example through TF ChIP-seq) and if binding is occurring, whether the chromatin accessibility is an effect of TF binding or a consequence (TFs acting as activator proteins) would be necessary to validate the developments in the MM TF network from my and Jin et al. work.

Using each subgroup's activated SMM enhancers interactions with OESMM genes, 38 TFs were found to be enriched in a subgroup specific manner, overlapping previously found primary and MM CL biology (Hideshima et al., 2010; Jin et al., 2018; Knief et al., 2017), for example: PRDM1, RXRA, CTCF, ETV or proteins of the IRF family. In cases such as MAZ, where it is known to activate MYC in a subset of MM samples (Zhan et al., 2002), this thesis provides further information regarding this activation occurring in HD samples. Despite being expected, there is no enrichment found in the MAF TF at MAF activated candidate enhancers. Possible reasons explaining this might be insufficient ratio of MAF bound enhancers from all the MAF activated regions or the fact that as it happens with the CCND2 gene, MAF binds to the promoter (Hurt et al., 2004) and perhaps not the enhancer. Moreover, most enriched TFs occur in HD and MMSET, with only BCL6 binding significantly in the MAF subgroup specifically. This suggests that the SMM enhancers analysed for CCND1 and MAF do not have particular TFs binding to them or they might be very heterogeneous in terms of TF binding at chromatin accessible areas, not yielding significant results. It is also possible that this method for obtaining TF candidates has low sensitivity particularly for the mentioned subgroups since it uses a hardcoded significance threshold. In the case of MAF, perhaps over accessibility calling is less reliable than in the rest of the subgroups due to only having two samples. Furthermore, it is also observed that each subgroup's and PC enriched TF genes are more highly expressed than random TFs suggesting a general trend commissioning enhancers with motifs which are already available in the PC cell.

In conjunction with the *in silico* MM and PC analysis by Jin and colleagues, my thesis provides new subgroup-specific insights about key activator candidates with therapeutical implications. The IRF family and particularly IRF4 has been previously found to have high biological relevance in PC and MM thought to be under the control of super enhancers (Jin et al., 2018; Young et al., 2013). The previous findings in this thesis confirm this idea, with IRF4 expression being in the top 0.2% in PC and MM samples but not overexpressed in MM or MM subgroups with respect to normal, it is possible that the general opening of chromatin in MM favours IRF4 binding at enhancer regions. Moreover, it is hypothesized that IFI44 overexpression via IRF4 repression found in lymphoma (Wang et al., 2014) may also be occurring in MAF and CCND1

subgroups where IFI44 is overexpressed (albeit nearly significantly) through lack of significant IRF4 binding at enhancers.

My work has obtained TF enrichment through TF DNA motif comparison. There are at least two limitations in using this method, the first is that TF motifs consensus sequences and their affinities suffer from errors and also the fact that even if significant enrichment for one TF occurs, it doesn't imply that the TF protein is necessarily bound to DNA and involved in transcription. Future work, should first try to validate the results from Jin *et. al* and my study in MM and MM subgroup samples respectively by using PC samples as control. Adding to genome-wide Chromosome Conformation Capture to determine enhancer promoter interactions, ChIP-seq to elucidate key TFs could be an option. Furthermore, perhaps mutations from MM Whole-Genome Sequencing (WGS) (Hoang et al., 2018), could be used to study the creation of novel or disruption of TF binding sites at chromatin accessible regions altering the expression of correlated genes as it was previously done (Corces et al., 2018). Another point to be addressed as it was mentioned previously (see The role of promoter accessibility in PC and MM gene expression, section 6.3), would be to observe TF binding at the promoters as it was done with the enhancers. Moreover, it is not known however, whether binding of the TFs at MM subgroups is a consequence of chromatin opening or a subset of the TFs found enriched at enhancer regions are remodelling chromatin acting as activators. To address this, depletion of key TFs could be performed followed by ATAC-seq to determine whether TF binding is making chromatin accessible. In this context, as mentioned in section 6.2, it has previously been detected in MM that B-cell active enhancers known to bind key TF in their lineage become methylated in MM (but not in PCs), thereby affecting binding of such TFs (Agirre et al., 2015). Given this fact, it would be interesting to consider methylation data of enhancer regions at the MM subtype level causing the TF patterns observed. As it is explained in section 6.4, confirmation of subtype-specific activator or TF activity would yield valuable information to proceed into investigating the repercussion in terms of diagnosis, prognosis and treatment in the clinic.

6.7. MOFA Unsupervised analysis – elucidates novel MM subgroup biology and weighting of features

Previous attempts in the literature have tried to subclassify MM samples in an unsupervised way based on microarray RNA expression (Mattioli et al., 2005; Shaughnessy et al., 2007), snoRNAs (Ronchetti et al., 2012), lncRNA signature (Ronchetti et al., 2018; Zhou et al., 2015),

miRNA-mRNA relationships (Liu et al., 2019) and MM disease progression applied to single cell RNA-seq (Jang et al., 2019). Additionally, studies have also used a combination of multiple data types to assess MM related aspects, for example CL resistance with DNA accessibility, methylation and gene expression (Dimopoulos et al., 2018); WGS, WES and RNA-seq for sample classifying (Barwick et al., 2019); DNA accessibility and ChIP-seq to determine super enhancers (Barwick et al., 2019) and E2F-DP1 effects (Fulciniti et al., 2018). Moreover, two previously mentioned studies (see Genes other than through translocation events may be deregulated in MM through cis-enhancer activation, section 6.1) used combinations of data, despite not using them in an unsupervised way, to propose enhancer – promoter interactions (Hoang et al., 2018; Jin et al., 2018). Jin and colleagues used unsupervised clustering to assign MM super-enhancer associated TFs to 4 groups separating MM from PC and PB but not obtaining a MM subgroup specific signature.

Chapter 5 makes use of MOFA on the top 5,000 most overall variable PC and MM chromatin accessible candidate enhancers and genes in an unsupervised way, creating independent LFs and having features associated different importance in each LF. For relevant separations of samples, enhancers, gene promoters, interactions of the two former, as well as involved TFs and gene pathways similarly as in Chapters 3 and 4 are elucidated. The fact that the technique is unsupervised, allows recapitulating findings previously found in this thesis and the literature, but importantly, it establishes novel interactions and their significance: one of the major aims in this thesis. Consequently, MOFA produces a model with dimensions discriminating the normal from the cancer state and the different MM subtypes based on IgH translocation partner: MAF, CCND1 and MMSET. LF1 and LF2 separate PC from tumour samples, LF3 parts the MMSET samples from the rest and LF5 creates an axis dividing the samples into MAF and CCND1 at each of the extremes. For completeness, deregulated gene examples from each separation are presented considering whether they are previously found in this thesis' supervised analysis, associated to MM and cancerous states in the literature or they are novel. Furthermore, addition of LF-specific criteria is also taken into account, examples with extreme and negligible gene loadings and overlapping or not LF-context specific reference gene-sets are provided. In this vein, the following gene sets are pondered: MM – gene association score (Piñero et al., 2020) for LF1 and LF2, differential genes between MMSET and non-MMSET MM samples (Wu et al., 2016) for LF3 and top differentially expressed genes isolating individual MAF and CCND1 subgroup MM clusters of samples from the rest (Broyl et al., 2010) for LF5.

6.7.1. Combined ATAC-seq and RNA-seq data classify subgroups

LF1, LF3 and LF5 (but not LF2) explain a significant ratio of both accessibility and gene expression variability, suggesting that the contribution of both variables is critical. Reinforcing this idea, incorporating ATAC-seq and RNA-seq data is more effective at unsupervised MM subgroup separation (for example, in terms of separating CCND1 translocated samples) than equivalent analysis of either gene expression or chromatin accessibility data by itself. This suggests that accessibility (a proxy for enhancers) and associated changes combined provide a more complete picture of the difference between different MM cytogenetic subgroups. Moreover, in terms of accessibility and gene profiles and their interactions, while the supervised analysis provides MM subgroups vs. PC differences, MOFA elucidates how particular MM subgroups are distinct from all other samples (whether PC or MM). Thereby treating each subgroup like a different disease from a distinct perspective.

Another finding is that ATAC-seq and RNA-seq variability are different, among other factors, because of the scale: genes have a 5-fold larger range in values than accessible regions, this hints at changes in accessibility being more subtle. In turn helping to explain why in some cases MM and subtype enhancers with importance in MOFA might be disregarded in the supervised analysis for overlapping a region that, despite being considered accessible, it may not surpass the accessibility threshold to be active in PC (see Other general limitations of the work, section 6.8). Despite the nature of ATAC-seq and RNA-seq, more variance in the data is explained by the former than the latter. LF1 (MM vs. PC separation) explains nearly 60% of the total ATAC-seq variance in the data. In terms of the number of features significantly distinguishing the cancer from healthy states, there were between two to more than ten-fold more DNA regions than genes (9,527 MMPC enhancers compared with 806 DEMM genes and 6,897 DASMM enhancers vs. 2,749 DESMM genes). Although the criteria was less stringent for regions than genes: fold change of 2 compared to 2.83 respectively. Together, this suggests that MOFA may be capturing more subgroup specific genes than ATAC-seq features. Perhaps explaining why despite the fact that on the supervised analysis there are proportionally less common candidate enhancers to all MM subgroups than DE genes (accessibility being more subgroup discriminative than gene expression), in MOFA, ATAC-seq by itself classifies samples into subgroups inferiorly than RNA-seq alone but combining both data types outperforms the two models.

Collectively, this proposes the importance of considering enhancer – promoter deregulation in MM and particularly its subgroups. Albeit being pondered in previous studies (Hoang et al., 2018; Jin et al., 2018), it has never been done in a non-directed way (not explicitly relating

regulatory regions to target genes) where it can provide a more complete picture. Furthermore, as explained in more detail in section 6.7.3, the number of samples in each MM subgroup and PCs is highly variable and can create significant bias as to how the variability is explained by the different LFs. It would therefore be desirable to utilize the MOFA framework on data with similar subtype sample numbers in the future and compare the results with the ones in this thesis.

6.7.2. Different LF separations and their meaning

As mentioned, the first two LFs separate MM from PCs. LF1 establishes a chromatin and gene expression driven axis emphasizing a general opening of chromatin in MM candidate enhancers which may be leading in some cases to MM gene overexpression. There are also MM under expressed genes (perhaps tumor suppressor genes), which may be occurring through a combination of mechanisms such as methylation, for example in the CCND2 promoter of MM CLs with CCND1 translocation (Misiewicz-Krzeminska et al., 2016), mutations (Hoang et al., 2018) or perhaps indirect effects such as enhancer activating genes that suppress other genes. Another way in which the PC genes may become active is by triggering of PC enhancers through means different to gain in accessibility such as TF binding or acetylation.

Importantly, MOFA delineates critical genes, some of which are significantly associated with MM disease (Piñero et al., 2020) and their cognate pathways dictating the LF1 division. Examples of these pathways regarding MM activation and/or PC repression recapitulate previous findings such as BMP signaling via interferon genes (Takaoka et al., 2008) or positive regulation of proteolysis which has been subject of directed treatment in MM using proteasome inhibitors (Csizmadia et al., 2016). Another example is epigenetic and chromatin remodelling categories (Baxter et al., 2014), interestingly, this occurs in LF1 (and not LF2), which explains close to 60% of all chromatin accessibility variance. Hence, apart from MM enhancers augmenting expression, these genes may in turn be activating DNA in the cancer state, as the features reflect it: through chromatin modifiers. These may include activators, histone methyltransferases as it has been previously found in MM with EZH2 (Goldsmith et al., 2019) and NSD2 (Lhoumaud et al., 2019), or elements like CTCF involved in loop formation (Braccioli and de Wit, 2019): already found to be enriched in SMM enhancers (see section 3.3.11). Furthermore, post-transcriptional gene repression also appears to be significantly enriched in this set of genes, prompting the hypothesis that a subset of the genes are repressed in the MM state (and hence appear PC OE) by these mechanisms. This could be the case of MIR425, a miRNA OEMM gene already found to be DE in extracellular RNA in the

plasma of MM patients with prognostic value (Chen et al., 2019), despite not being found to interact with mRNA in MM MMSET translocated compared with non-translocated patients (Liu et al., 2019). Importantly, the genes causing enrichments in these pathways are identified and they can be further studied in MM and PC.

Of all the MOFA candidate enhancer regions, 40% overlap with MMPC enhancers obtained in Chapter 3. Furthermore, the latter regions are very significantly enriched in the MOFA MM activated genomic areas, suggesting a MM and PC differential enhancer signature overlapping both analyses to be verified with possible clinical value. Candidate enhancers and genes within 1Mb are paired in order to study the 15,903 interactions obtained. Interestingly, one of these interactions regarding the genomic area chr9:37,374,916-37,375,220 recapitulates a previously proposed interaction between a putative enhancer at chr9:37,375,172-37,395,282 and PAX5 based on mutation of the regulatory region being linked to an altered expression (Hoang et al., 2018). Consistently, the loading for LF1 on the region reflects MM activation, which is confirmed by increased region accessibility in MM, however, the PAX5 gene is not OE in all MM samples. Perhaps pointing at a subgroup specific effect.

LF2 is a MM – PC separation dominated by gene expression with negligible chromatin accessibility variability explained. It is highly correlated with MM vs. PC gene changes (Spearman correlation coefficient -0.83) suggesting it is very paired with a subset of the Myelomagenesis expression changes. Importantly, in terms of transcriptome, the genes with top MM activation reflected in this axis and their ontology are mostly non-overlapping with the analogous genes for the LF1 separation, implying independent regulatory mechanisms. Correspondingly, LF2 reveals MM activated gene enrichment for neuron and synaptic pathway-related genes, perhaps through genes such as the very MM active Reelin (RELN) involved in neuron migration (Lin et al., 2017a) or the cadherin 2 gene (CDH2), a valuable treatment target in MM (Mrozik et al., 2015). Consistently, both RELN and CDH2 already have been associated with MM (Jin et al., 2018; Piñero et al., 2020) having a high significance in the LF2 separation but a minor one in LF1. Differently to other interactions, both genes contain interactions with multiple enhancers having a very low LF2 implication. Together, this suggests alternative activating mechanisms such as the already suggested RELN OE is promoter hypomethylation (Lin et al., 2017a).

LF1 and LF2 are therefore two axes found sharing a very similar separation, hence why a certain degree of correlation exists between them (see Figure 5-4 in MOFA separates samples into their constituent subgroups, section 5.3.1). Despite this, it is thought that they do not

merge into a unique LF after running the analysis with different starting seeds for two main reasons. The first is that they explain very distinct volumes of ATAC-seq variability, with LF1 being predominantly chromatin variant. The second is that they both explain different transcriptomic pathways.

LF3 separates MMSET translocated samples (including A17.5 which is unknown but thought to be) from the rest, consequently having MMSET changes with respect to PC (see Chapter 4) only partially explaining this axis. The A26.1 HD sample is also placed on the MMSET cluster, perhaps, this sample exhibiting a MMSET-like translocation phenotype also shares similar chromatin and gene expression changes during Myelomagenesis. LF3 division is consistent with previously key genes (mostly MMSET OE) found in MMSET vs. Non-MMSET patients (Wu et al., 2016), as well as with FGFR3 which is activated in most MMSET samples as a translocation effect (Stewart et al., 2004) and MAP1B (Mirabella et al., 2014). Also with MM CLs genes such as GJB2, COCH, IGFBP6, F12, KRT86, CEBPA, RNF157, ADCY1 (Martinez-Garcia et al., 2011).

Moreover, apart from recapitulating previous literature regarding the MMSET subgroup, enrichment on MOFA determined exclusively activated MMSET genes in MOFA reveals chromatin binding enrichment, which is consistent with the fact that NSD2/MMSET has been previously thought to be involved with transcriptional elongation (Kuo et al., 2011; Martinez-Garcia et al., 2011) through H3K36 methylation. This may allow SETD2 and other proteins such as the histone methyltransferase EZH2 to bind to DNA (Popovic et al., 2014). LF3 highly weighted genes and the regulatory mechanisms and effects of the chromatin binding proteins should be further studied for example through ChIP-seq.

Different novel gene interactions with enhancers are proposed to be distinctive in MMSET, some regarding genes already known to have implications in MM biology and elucidated in Chapter 4 such as CDC42BPA. A gene only overexpressed with respect to PC and altered in MMSET and non-MMSET subgroup MM patients (Wu et al., 2016), it is proposed to form a pathway interacting with CLEC11A (also marked active in LF3) in two t(4;14) MM CLs (Laganà et al., 2018). Furthermore, MAP1B, MYRIP and FGFR3 are also proposed to cooperate with CDC42BPA (Liu et al., 2019), the latter, a gene deregulated in approximately 75% of MMSET cases. Importantly, LF3 delineates a very active candidate region chr1:227,191,697-227,192,560 thought to augment CDC42BPA expression and suggesting amplification of this pathway only in MMSET. Other with novel genes in MMSET are elucidated such as FOXA1

which is found to be recurrently mutated in prostate cancer (Barbieri et al., 2012) and in prognosis of breast cancer (summarized in Hu et al., 2014).

Jointly, the LF3 separation recapitulates widespread mechanisms of chromatin remodelling and prior subgroup signature genes while also extending them and proposing associated MMSET altered enhancers. Importantly, this axis also properly classifies samples with unknown MMSET translocation state but with equivalent gene expression to t(4;14) samples as it occurs with the signature subgroup differential regions and genes in Chapter 4.

LF5 represents a scale in which each of the ends are associated with CCND1 and MAF translocated samples and thought to be driven by the latter. Consistent with this, DE genes isolating MAF and CCND1 samples into individual clusters (Broyl et al., 2010) have coherent expression correlated with LF5 loadings only for the MAF subgroup (not for CCND1-translocated samples). Furthermore, apart from novel genes, other important known signature genes for this separation also are reliably classified, these include MAF, ENSG00000258776/AL161757.4: one of the top lncRNA significantly deregulated in MAF translocated vs. non MAF translocated samples (Ronchetti et al., 2018), or CCND2 particularly overexpressed in the MAF subgroup (Shah et al., 2018).

In line with the previously mentioned CCND1 - CCND2 dichotomy (see The MAF subgroup CCND2 enhancer region 6.5), these two genes are assigned opposing loadings in this axis suggesting that LF5 may be explanatory of this phenomenon as well as other pathways. In this context, the region chr12:4,148,394-4,148,872 (part of the tested CCND2 super-enhancer) is assigned extreme weights coherent with suggesting strong activation in the MAF subgroup. Furthermore, MOFA also allows proposing common and different mechanisms between MMSET (the other subgroup having CCND2 with a moderately high expression) and MAF translocated samples. For example, the MAF and MMSET subgroups activate the chr16:79,283,625-79,283,930 enhancer which may be regulating overexpression of the lncRNA ENSG00000278058/RP11-70D24.2 gene, a gene known to be correlated with MAF expression in CCND1 translocated MM samples (Ronchetti et al., 2018). Moreover, there are also possibilities of divergence in subgroup specific activation of different pathways, for example, when considering the chr14:53,968,790-53,969,329 enhancer activating BMP4 in MMSET or MAF-specific gene repression of the CPED1 thought to be associated with the enhancer at chr7:120,032,813-120,033,781.

Notably, the LF5 separation provides novel subgroup resolution for already elucidated MM vs. PC interactions. Such is the case with the PAX5 gene and chr9:37,375,172-37,395,282,

indicating activity in the CCND1 group only, suggesting its importance for the MAF – CCND1 subgroup distinction. Since the PAX5 gene is important in the PC and MM differentiation state (Manier et al., 2017; Paiva et al., 2017), this pathway may be occurring only in CCND1 translocated samples, providing prognosis and personalized therapeutical knowledge. Moreover, associated processes linked to the top genes involved in LF5 are under-represented for integrin binding and signalling receptor binding which could be explained by lack of MMSET-specific genes such as NSD2 (with negligible LF5 loadings). It is known that integrins have important roles in oncogenic processes in MM cells; one suggested way in which this occurs is by adopting an active conformation (Hosen, 2020), perhaps this mechanism is not employed in this axis. Together, LF5 provides a robust axis differentiating MAF samples (particularly from CCND1 samples) and recapitulating critical related gene biology to further explore novel genes and the suggested pathways and interactions with enhancers.

Accompanied by MOFA LF5, the gene ontology for the OEMM and OESMM (chapter 3 and 4 respectively) also show lack of integrin binding which is involved in signalling pathways with the extra-cellular matrix for different cell processes (Brown and Marshall, 2019; Hynes R O., 2002). Furthermore, SPP1 a MAF subgroup very OE gene is involved in bone remodelling and deregulated in multiple haematological cancers (Bastos et al., 2017). Moreover, in LF1, Reelin appears highly expressed in MM samples, with associated cancerous properties in the literature promoted through a Reelin/β1 integrin pathway such as bone marrow invasion (Lin et al., 2017b, 2016; Neri et al., 2011). Additionally, different BMPs are also particularly OE and repressed in the MMSET and MAF subgroups respectively (Grčević et al., 2010), enriched in MOFA LF1 separations. Together this points at the processes between the bone marrow, MM and the extra-cellular matrix being a recurrent theme in this work as it has previously been shown (Glavey et al., 2017). These mechanisms can now be studied with a MM subgroup perspective.

Despite previous attempts to classify MM samples into the different subgroups, to my knowledge, this is the first time that an exhaustive unsupervised classification based on simultaneous integration of enhancer biology and target gene expression has been performed. Collectively, MOFA proposes previously elucidated and novel MM independent (LF1 and LF2) or MM subgroup-specific (LF3 and LF5) pathways through enhancer-mediated activation, but also other possible mechanisms to be explored when other assays such as H3K27Ac or methylation of primary bone marrow MM cells are available. These mechanisms may be acting in a mutually exclusive or additive way, at the enhancer and/or promoter level of tumor suppressing or proto-oncogenes. Furthermore, possible feedback loops whereby altered gene

expression may in turn be altering chromatin state and other gene networks are also propositioned. These pathways should be validated through experimental testing, while the enhancer – promoter interactions proposed can be explored through means previously explained (see section 1.4) as it has been done with the CCND2 enhancer. Moreover, future addition of other paired orthogonal data such as DNA mutations could complement the MOFA analysis and reinforce the different acting pathways in the MM state.

It is also important to evaluate whether non-coding transcripts are key to the MM and subgroup processes. This could be done by comparing additional MOFA models combining chromatin accessibility with separate versions for protein coding and non-protein coding genes. I have already compared combinations of genomic features including TSS with protein coding or all genes obtaining different separations (not shown). I have also performed a run with MOFA having all inputs equal to the run analysed in this work, except for using regions including TSS, resulting in no appreciable changes in terms of LF separations. Predictably, using only enhancer regions with different combinations of protein coding and non-protein coding genes should yield different separations, elucidating that the non-coding transcripts are key to the MM and subgroup processes.

Importantly, MOFA, as opposed to other methods such as the ones involving the supervised analysis in Chapters 3 and 4 have the added benefit of providing an additional layer establishing the importance of each element forming interactions in each LF biological separation. This weighting is based on an unbiased data-driven approach and can be used advantageously to prioritize the validation of the findings. Moreover, the divisions created by this analysis provide a novel angle in viewing MM subgroups as independent conditions, not only separating from PC (as with the analysis in Chapter 4) but also from other MM subgroups as well. Together, similarly as in section 6.4, validating these insights can provide knowledge of the enhancer network, affected genes, pathways and the network of TFs involved. Given that currently Myeloma is an incurable cancer, these elements can be further examined in the context of diagnosis, prognosis and treatment stratifying into subgroups. The latter especially, given the recent advances in tumour directed, *in vivo* genome editing, capable of knocking out specific targets and significantly suppressing tumour growth (Guo et al., 2019).

6.7.3. Quality control considerations and limitations in MOFA

The result of applying the MOFA model on the data was verified for robustness, confirming that the current analysis performs equally well than a model with twice the number of features (10,000) and the results were the same with different random seed trials. Moreover,

the meaningfulness of the input variables was also verified observing that a proportion of the features used in the model had insignificant means across all samples (seeming a larger problem in gene expression) and thus appearing like a more parsimonious model with less features could perform equally well. Despite this, the number of samples in the MM subgroups and PC is very variable: for example, there are only 2 MAF translocated compared to 13 HD samples. Therefore an argument could be made that a biologically distinguishing feature in one of the subgroups containing fewer samples would yield a low overall variance across all samples and thus might not be selected if a stricter threshold with higher minimum feature variance was used.

Furthermore, given that the number of samples per subgroup was different, a compromise had to be taken between generating a completely unsupervised model (without PC or MM subgroup labels) based on overall feature variability and taking into account inter subgroup variability but not being purely unsupervised in terms of feature selection. The former was chosen, but this means that as mentioned, PC or subgroup specific expressed or repressed features will have a mutable overall variability depending on subgroup sample size and this affects their inclusion or exclusion as MOFA features. Perhaps for this reason, all the MOFA features found to overlap differential features in the MM vs. PC analysis in Chapter 3 are more accessible in cancer (but probably have a high variability within cancer samples).

Given the limitations in different sample numbers for the different subgroups, a future study with more balanced numbers should be performed to make sure that the per subgroup variance as a function of the total is more even. Moreover, the number of features used should have probably been higher for accessibility and lower for gene expression to represent an equal ratio from the starting features (2% and 9% respectively). Also, this would mitigate the presence of low variability genes and provide additional enhancer candidates to MOFA genes present in supervised interactions such as CLEC11A but lacking MOFA candidate regions.

Despite the fact that all the MM subgroups based on IgH enhancer translocations are separated by the different LFs, the HD group doesn't appear to be. Multiple reasons could exist for this, for example, nearly half of the total samples in the study are HD, therefore the total variability is probably very dictated by the within variability of the HD group, leading to LFs creating partitions dividing the HD samples. Another factor to consider is heterogeneity, in general, MM is a very heterogeneous disease (Egan et al., 2012; Lohr et al., 2014), as can be seen in section 1.7.2. Furthermore, in particular, within the hyperdiploid group there may exist critical translocations as well as different chromosomal abnormalities dictating MM biology.

For example, it has been previously found that even in MM patients presenting other critical abnormalities such as del(17p) and t(4;14), trisomy 3 and/or 5 improved MM survival, while trisomy 21 worsened it (Chretien et al., 2015). Moreover, it is possible that since the driving initiating event is not IgH enhancer related, the enhancer – promoter profile is less distinctive and more heterogeneous than in the other MM subgroups.

In terms of the MOFA analysis, the HD group is the least analysed subgroup, given the significant number of samples provided in this study, this subgroup can benefit from further analysis and classification into subdivisions. One of the possible improvements in the current study already mentioned would be to analyse how HD samples saturate accessibility features to determine heterogeneity. Future work should begin by doing unsupervised clustering of enhancer and promoter features only within the HD group. Furthermore, insights from the supervised analysis can also be combined with additional data such as gene copy number to determine critical pathways.

6.8. Other general limitations of the work

Some restrictions regarding samples have affected the work, for example, despite an ATAC-seq depth of 30 to 50 million reads (Ackermann et al., 2016; Buenrostro et al., 2015; Neph et al., 2012) being recommended. A lower than desirable minimum sample ATAC-seq read depth of 28,231,242 single end reads for primary samples and less than 10 million for cell lines was used, this meant that cell line information had to be incorporated in a qualitative way. Furthermore, a common minimum read depth per sample was used and the samples for each MM subgroup and PC were pooled together to call peaks and then subgroup peaks merged. Firstly, since the minimum sample read depth is less than recommended, it is possible that a significant fraction of condition specific peaks was not recalled. Secondly, since subgroups had different number of samples, this also meant that per subgroup sequencing depth was not common across subgroups (and PC) when calling subgroup peaks. The result of this is that, assuming equal MM subgroup-specific features, subgroups with more samples such as HD should be closer to saturating the specific accessible features than other subgroups like MAF (as observed by the enrichment in the number of HD exclusive consensus peaks for PC and MM subgroups). This compromise was chosen since for chromatin accessible peak calling a minimum starting material per sample is necessary to produce accessibility signal peaks differentiating them from noise. If per subgroup constant read depth was employed, for subgroups containing many samples, the per sample sequencing depth would be too low to

attain proper peak calling. Another important point is that subgroups may have different degrees of heterogeneity, for example, as it was mentioned (see Quality control considerations and limitations in MOFA, section 6.7.3) it is likely that HD is more heterogeneous and may not contain a genuine set of consensus accessible regions for it. It is therefore important that accessibility comparisons between MM subgroups and PC take this into account. In this regard, it would be informative to produce subgroup specific consensus peaks per added sample saturation curves to analyse how close the data is to identifying a theoretical upper limit.

Another issue regarding the quality control thresholds used for the ATAC-seq and RNA-seq samples is that this caused some of the samples being removed from the study. This resulted in compromises to be taken, for example, PC donor CD19 status, donor id, MM subgroups and batches were confounding variables, which meant that all variables' effect sizes could not be taken into account simultaneously. This was mitigated by using a conservative approach: considering only significant effect sizes simultaneously overlapping each individual covariate comparison. Similarly, replicates were only available for one PC sample and should have been added in general to validate and increase the confidence in the biological findings. Furthermore, after quality control removal of samples, some batches contained only one sample and to mitigate this, all this samples in this situation were considered to belong to the same batch, assuming that they all were under the same batch effects. Ideally, a future study where batch effects and other covariates can be modelled simultaneously would be desirable to check the results obtained.

In the supervised analysis in chapters 3 and 4, MM and MM subgroup exclusively active regions were determined on the basis of removing PC accessible regions from significantly accessible regions in the cancer state. Since the thresholds used for chromatin accessibility peak calling determine what is considered an accessible active region, it is possible that some exclusive MM (or MM subgroup) active regions were disregarded for having non-genuine PC accessibility (despite surpassing peak calling thresholds). This may have been the case with the proposed interaction in chapter 5: chr15:33,915,253-33,916,443 with the FMN1 gene, which is not considered overlapped by a SMM enhancer in chapter 3 for the reason stated. Despite the fact that the peak calling process was visually verified in different signal scenarios, a more conservative threshold on the PC peak calling would reduce the false positives. Moreover, it must be noted that *in silico* determined candidate enhancers to be tested were visually verified first.

7. References

Abe, M., Hiura, K., Wilde, J., Shioyasono, A., Moriyama, K., Hashimoto, T., Kido, S., Oshima, T., Shibata, H., Ozaki, S., Inoue, D., Matsumoto, T., 2004. Osteoclasts enhance myeloma cell growth and survival via cell-cell contact: A vicious cycle between bone destruction and myeloma expansion. Blood. https://doi.org/10.1182/blood-2003-11-3839

Ackermann, A.M., Wang, Z., Schug, J., Naji, A., Kaestner, K.H., 2016. Integration of ATAC-seq and RNA-seq identifies human alpha cell and beta cell signature genes. Mol. Metab. 5, 233–244. https://doi.org/10.1016/j.molmet.2016.01.002

Affer, M., Chesi, M., Chen, W.D., Keats, J.J., Demchenko, Y.N., Tamizhmani, K., Garbitt, V.M., Riggs, D.L., Brents, L.A., Roschke, A. V, Van Wier, S., Fonseca, R., Bergsagel, P.L., Kuehl, W.M., 2014. Promiscuous MYC locus rearrangements hijack enhancers but mostly super-enhancers to dysregulate MYC expression in multiple myeloma. Leukemia 28, 1725–1735. https://doi.org/10.1038/leu.2014.70

Aggarwal, R., Ghobrial, I.M., Roodman, G.D., 2006. Chemokines in multiple myeloma. Exp. Hematol. https://doi.org/10.1016/j.exphem.2006.06.017

Agirre, X., Castellano, G., Pascual, M., Heath, S., Kulis, M., Segura, V., Bergmann, A., Esteve, A., Merkel, A., Raineri, E., Agueda, L., Blanc, J., Richardson, D., Clarke, L., Datta, A., Russiñol, N., Queirós, A.C., Beekman, R., Rodríguez-Madoz, J.R., José-Enériz, E.S., Fang, F., Gutiérrez, N.C., García-Verdugo, J.M., Robson, M.I., Schirmer, E.C., Guruceaga, E., Martens, J.H.A., Gut, M., Calasanz, M.J., Flicek, P., Siebert, R., Campo, E., San Miguel, J.F., Melnick, A., Stunnenberg, H.G., Gut, I.G., Prosper, F., Martín-Subero, J.I., 2015. Whole-epigenome analysis in multiple myeloma reveals DNA hypermethylation of B cell-specific enhancers. Genome Res. 25, 478–487. https://doi.org/10.1101/gr.180240.114

Agnarelli, A., Chevassut, T., Mancini, E.J., 2018. IRF4 in multiple myeloma—Biology, disease and therapeutic target. Leuk. Res. https://doi.org/10.1016/j.leukres.2018.07.025

Aiden, E.L., Casellas, R., 2015. Somatic Rearrangement in B Cells: It's (Mostly) Nuclear Physics. Cell 162, 708–11. https://doi.org/10.1016/j.cell.2015.07.034

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A., 2017. Genetic effects on chromatin accessibility foreshadow gene expression changes in macrophage immune response. bioRxiv. https://doi.org/10.1101/102392

Albrecht, F., List, M., Bock, C., Lengauer, T., 2016. DeepBlue epigenomic data server: programmatic data retrieval and analysis of epigenome region sets. Nucleic Acids Res. https://doi.org/10.1093/nar/gkw211

Alvarez-Benayas, J., 2020a. Data for Computational analysis of enhancer deregulation in Multiple Myeloma [WWW Document]. URL https://doi.org/10.15131/shef.data.12273149

Alvarez-Benayas, J., 2020b. pipeline_atac_consensus_balanced_peaks for "Computational analysis of enhancer deregulation in Multiple Myeloma" [WWW Document]. URL https://doi.org/10.15131/shef.data.12280904

Alvarez-Benayas, J., 2019. atac_seq_pipeline for Thesis "Computational analysis of enhancer deregulation in Multiple Myeloma."

Ampuja, M., Rantapero, T., Rodriguez-Martinez, A., Palmroth, M., Alarmo, E.L., Nykter, M., Kallioniemi, A., 2017. Integrated RNA-seq and DNase-seq analyses identify phenotype-specific BMP4 signaling in breast cancer. BMC Genomics 18, 1–15. https://doi.org/10.1186/s12864-016-3428-1

Andrews, S., 2010. FASTQC. A quality control tool for high throughput sequence data.

Angelica, M.D., Fong, Y., 2008. In search of the determinants of enhancer–promoter interaction specificity. October 141, 520–529. https://doi.org/10.1016/j.surg.2006.10.010.Use

Aplan, P.D., 2006. Causes of oncogenic chromosomal translocation. Trends Genet. 22, 46–55. https://doi.org/10.1016/j.tig.2005.10.002

Aran, D., Hellman, A., 2013. DNA Methylation of Transcriptional Enhancers and Cancer Predisposition. Cell 154, 11–13. https://doi.org/10.1016/J.CELL.2013.06.018

Aran, D., Sabato, S., Hellman, A., 2013. DNA methylation of distal regulatory sites characterizes dysregulation of cancer genes. Genome Biol. https://doi.org/10.1186/gb-2013-14-3-r21

Argelaguet, R., Velten, B., Arnol, D., Dietrich, S., Zenz, T., Marioni, J.C., Buettner, F., Huber, W., Stegle, O., 2018. Multi-Omics Factor Analysis-a framework for unsupervised integration of multi-omics data sets. Mol. Syst. Biol. https://doi.org/10.15252/msb.20178124

Arnold, 2011. GAwk.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., Stark, A., 2013. Genome-wide

quantitative enhancer activity maps identified by STARR-seq. Science (80-.). https://doi.org/10.1126/science.1232542

Bailey, T., Noble, W.S., n.d. The MEME Suite [WWW Document]. URL http://meme-suite.org/doc/meme.html

Bailey, T.L., 2011. DREME: Motif discovery in transcription factor ChIP-seq data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btr261

Bailey, T.L., Elkan, C., 1994. Fitting a mixture model by expectation maximization to discover motifs in biopolymers. Proceedings. Int. Conf. Intell. Syst. Mol. Biol.

Barbieri, C.E., Baca, S.C., Lawrence, M.S., Demichelis, F., Blattner, M., Theurillat, J.P., White, T.A., Stojanov, P., Van Allen, E., Stransky, N., Nickerson, E., Chae, S.S., Boysen, G., Auclair, D., Onofrio, R.C., Park, K., Kitabayashi, N., MacDonald, T.Y., Sheikh, K., Vuong, T., Guiducci, C., Cibulskis, K., Sivachenko, A., Carter, S.L., Saksena, G., Voet, D., Hussain, W.M., Ramos, A.H., Winckler, W., Redman, M.C., Ardlie, K., Tewari, A.K., Mosquera, J.M., Rupp, N., Wild, P.J., Moch, H., Morrissey, C., Nelson, P.S., Kantoff, P.W., Gabriel, S.B., Golub, T.R., Meyerson, M., Lander, E.S., Getz, G., Rubin, M.A., Garraway, L.A., 2012. Exome sequencing identifies recurrent SPOP, FOXA1 and MED12 mutations in prostate cancer. Nat. Genet. https://doi.org/10.1038/ng.2279

Barwick, B.G., Neri, P., Bahlis, N.J., Nooka, A.K., Dhodapkar, M. V., Jaye, D.L., Hofmeister, C.C., Kaufman, J.L., Gupta, V.A., Auclair, D., Keats, J.J., Lonial, S., Vertino, P.M., Boise, L.H., 2019. Multiple myeloma immunoglobulin lambda translocations portend poor prognosis. Nat. Commun. https://doi.org/10.1038/s41467-019-09555-6

Bastos, A.C.S.F., Blunck, C.B., Emerenciano, M., Gimba, E.R.P., 2017. Osteopontin and their roles in hematological malignancies: Splice variants on the new avenues. Cancer Lett. https://doi.org/10.1016/j.canlet.2017.08.022

Battula, V.L., Le, P.M., Sun, J.C., Nguyen, K., Yuan, B., Zhou, X., Sonnylal, S., McQueen, T., Ruvolo, V., Michel, K.A., Ling, X., Jacamo, R., Shpall, E., Wang, Z., Rao, A., Al-Atrash, G., Konopleva, M., Davis, R.E., Harrington, M.A., Cahill, C.W., Bueso-Ramos, C., Andreeff, M., 2017. AML-induced osteogenic differentiation in mesenchymal stromal cells supports leukemia growth. JCI insight. https://doi.org/10.1172/jci.insight.90036

Baxter, E., Windloch, K., Gannon, F., Lee, J.S., 2014. Epigenetic regulation in cancer progression. Cell Biosci. https://doi.org/10.1186/2045-3701-4-45

Benabdallah, N.S., Williamson, I., Illingworth, R.S., Boyle, S., Grimes, G.R., Therizols, P., Bickmore, W., 2017. PARP mediated chromatin unfolding is coupled to long-range enhancer activation. bioRxiv. https://doi.org/10.1101/155325

Bergsagel, P., Kuehl, W., 2005. Molecular pathogenesis and a consequent classification of multiple myeloma.

Bergsagel, P.L., Kuehl, W.M., Zhan, F., Sawyer, J., Barlogie, B., Shaughnessy, J., 2005. Cyclin D dysregulation: An early and unifying pathogenic event in multiple myeloma. Blood. https://doi.org/10.1182/blood-2005-01-0034

Bhagwat, A.S., Lu, B., Vakoc, C.R., 2018. Enhancer dysfunction in leukemia. Blood 131, 1795–1804. https://doi.org/10.1182/blood-2017-11-737379

Bian, Z., Zhang, Jiwei, Li, M., Feng, Y., Wang, X., Zhang, Jia, Yao, S., Jin, G., Du, J., Han, W., Yin, Y., Huang, S., Fei, B., Zou, J., Huang, Z., 2018. LncRNA-FEZF1-AS1 promotes tumor proliferation and metastasis in colorectal cancer by regulating PKM2 signaling. Clin. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-17-2967

Bianchi, G., Munshi, N.C., 2015. Pathogenesis beyond the cancer clone(s) in multiple myeloma. Blood 125, 1–11. https://doi.org/10.1182/blood-2014-11-568881.BLOOD

Bickmore, W.A., 2013. The Spatial Organization of the Human Genome. https://doi.org/10.1146/annurev-genom-091212-153515

Bila, J., Suvajdzic, N., Elezovic, I., Colovic, M., Boskovic, D., 2007. Systemic lupus Erythematosus and IgA multiple myeloma: A rare association? Med. Oncol. https://doi.org/10.1007/s12032-007-0047-3

Birshtein, B.K., 2014. Epigenetic regulation of individual modules of the immunoglobulin heavy chain locus 3' regulatory region. Front. Immunol. 5, 1–9. https://doi.org/10.3389/fimmu.2014.00163

Blow, M.J., Mcculley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-frick, I., Shoukry, M., Wright, C., Chen, F., Bristow, J., Ren, B., Black, B.L., Rubin, E.M., Visel, A., 2011. ChIP-seq Identification of Weakly Conserved Heart Enhancers 42, 806–810. https://doi.org/10.1038/ng.650.ChIP-seq

Blueprint_project, 2016. Segmentation of ChIP-Seq data blueprint portal [WWW Document]. URL ftp://ftp.ebi.ac.uk/pub/databases/blueprint/releases/current_release/homo_sapiens/sec

ondary_analysis/Segmentation_of_ChIP-Seq_data/

Bong, I.P.N., Ng, C.C., Baharuddin, P., Zakaria, Z., 2017. MicroRNA expression patterns and target prediction in multiple myeloma development and malignancy. Genes and Genomics. https://doi.org/10.1007/s13258-017-0518-7

Bose, D.A., Donahue, G., Reinberg, D., Shiekhattar, R., Bonasio, R., Berger, S.L., 2017. RNA Binding to CBP Stimulates Histone Acetylation and Transcription. Cell. https://doi.org/10.1016/j.cell.2016.12.020

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., Crawford, G.E., 2008. High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell 132, 311–322. https://doi.org/10.1016/j.cell.2007.12.014.High-Resolution

Braccioli, L., de Wit, E., 2019. CTCF: a Swiss-army knife for genome organization and transcription regulation. Essays Biochem. https://doi.org/10.1042/EBC20180069

Broad_Institute, 2018. Picard Tools.

Brown, C.R., Kennedy, C.J., Delmar, V.A., Forbes, D.J., Silver, P.A., 2008. Global histone acetylation induces functional genomic reorganization at mammalian nuclear pore complexes 627–639. https://doi.org/10.1101/gad.1632708.4

Brown, N.F., Marshall, J.F., 2019. Integrin-Mediated TGFβ Activation Modulates the Tumour Microenvironment. Cancers (Basel). 11, 1221. https://doi.org/10.3390/cancers11091221

Broyl, A., Hose, D., Lokhorst, H., De Knegt, Y., Peeters, J., Jauch, A., Bertsch, U., Buijs, A., Stevens-Kroef, M., Beverloo, H.B., Vellenga, E., Zweegman, S., Kersten, M.J., Van Der Holt, B., El Jarari, L., Mulligan, G., Goldschmidt, H., Van Duin, M., Sonneveld, P., 2010. Gene expression profiling for molecular classification of multiple myeloma in newly diagnosed patients. Blood. https://doi.org/10.1182/blood-2009-12-261032

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., Greenleaf, W.J., 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods. https://doi.org/10.1038/nmeth.2688

Buenrostro, J.D., Wu, B., Chang, H.Y., Greenleaf, W.J., 2015. ATAC-seq: A method for assaying chromatin accessibility genome-wide. Curr. Protoc. Mol. Biol. 2015, 21.29.1-21.29.9. https://doi.org/10.1002/0471142727.mb2129s109

Bulger, M., Groudine, M., 2011. Functional and mechanistic diversity of distal transcription enhancers. Cell 144, 327–339. https://doi.org/10.1016/j.cell.2011.01.024

Cannavò, E., Khoueiry, P., Garfield, D.A., Geeleher, P., Zichner, T., Gustafson, E.H., Ciglar, L., Korbel, J.O., Furlong, E.E.M., 2016. Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. Curr. Biol. 26, 38–51. https://doi.org/10.1016/j.cub.2015.11.034

Canver, M.C., Smith, E.C., Sher, F., Pinello, L., Sanjana, N.E., Shalem, O., Chen, D.D., Schupp, P.G., Vinjamur, D.S., Garcia, S.P., Luc, S., Kurita, R., Nakamura, Y., Fujiwara, Y., Maeda, T., Yuan, G.C., Zhang, F., Orkin, S.H., Bauer, D.E., 2015. BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. Nature. https://doi.org/10.1038/nature15521

Carlson, M., 2018. org.Hs.eg.db: Genome wide annotation for Human [WWW Document]. Bioconductor.

Cavalcante, R.G., Sartor, M.A., 2017. Annotatr: Genomic regions in context. Bioinformatics. https://doi.org/10.1093/bioinformatics/btx183

CGAT_Developers, 2018. CGATOxford Github [WWW Document]. URL https://github.com/CGATOxford/CGATPipelines

Chaudhuri, J., Basu, U., Zarrin, A., Yan, C., Franco, S., Perlot, T., Vuong, B., Wang, J., Phan, R.T., Datta, A., Manis, J., Alt, F.W., 2007. Evolution of the Immunoglobulin Heavy Chain Class Switch Recombination Mechanism.

Chauhan, D., Singh, A. V., Ciccarelli, B., Richardson, P.G., Palladino, M.A., Anderson, K.C., 2010. Combination of novel proteasome inhibitor NPI-0052 and lenalidomide trigger in vitro and in vivo synergistic cytotoxicity in multiple myeloma. Blood. https://doi.org/10.1182/blood-2009-03-213009

Chavan, S.S., He, J., Tytarenko, R., Deshpande, S., Patel, P., Bailey, M., Stein, C.K., Stephens, O., Weinhold, N., Petty, N., Steward, D., Rasche, L., Bauer, M., Ashby, C., Peterson, E., Ali, S., Ross, J., Miller, V.A., Stephens, P., Thanendrarajan, S., Schinke, C., Zangari, M., Van Rhee, F., Barlogie, B., Mughal, T.I., Davies, F.E., Morgan, G.J., Walker, B.A., 2017. Bi-allelic inactivation is more prevalent at relapse in multiple myeloma, identifying RB1 as an independent prognostic marker. Blood Cancer J. https://doi.org/10.1038/bcj.2017.12

Chen, D.H., Yu, J.W., Jiang, B.J., 2015. Contactin 1: A potential therapeutic target and biomarker in gastric cancer. World J. Gastroenterol.

https://doi.org/10.3748/wjg.v21.i33.9707

Chen, F., Chen, C., Yang, S., Gong, W., Wang, Y., Cianflone, K., Tang, J., Wang, D.W., 2012. Let-7b inhibits human cancer phenotype by targeting cytochrome P450 epoxygenase 2J2. PLoS One. https://doi.org/10.1371/journal.pone.0039197

Chen, H., Levo, M., Barinov, L., Fujioka, M., Jaynes, J.B., Gregor, T., 2018. Dynamic interplay between enhancer–promoter topology and gene activity. Nat. Genet. 50, 1296–1303. https://doi.org/10.1038/s41588-018-0175-z

Chen, J., OuYang, H., An, X., Liu, S., 2018. Vault RNAs partially induces drug resistance of human tumor cells MCF-7 by binding to the RNA/DNA-binding protein PSF and inducing oncogene GAGE6. PLoS One. https://doi.org/10.1371/journal.pone.0191325

Chen, M., Mithraprabhu, S., Ramachandran, M., Choi, K., Khong, T., Spencer, A., 2019. Utility of Circulating Cell-Free RNA Analysis for the Characterization of Global Transcriptome Profiles of Multiple Myeloma Patients. Cancers (Basel). https://doi.org/10.3390/cancers11060887

Chen, N., He, S., Geng, J., Song, Z.J., Han, P.H., Qin, J., Zhao, Z., Song, Y.C., Wang, H.X., Dang, C.X., 2018. Overexpression of Contactin 1 promotes growth, migration and invasion in Hs578T breast cancer cells. BMC Cell Biol. https://doi.org/10.1186/s12860-018-0154-3

Chen, R., Zhao, H., Wu, D., Zhao, C., Zhao, W., Zhou, X., 2016. The role of SH3GL3 in myeloma cell migration/invasion, stemness and chemo-resistance. Oncotarget. https://doi.org/10.18632/oncotarget.12231

Chesi, M., Bergsagel, P.L., 2011. Many multiple myelomas: making more of the molecular mayhem. Hematology Am. Soc. Hematol. Educ. Program. https://doi.org/10.1182/asheducation-2011.1.344

Chi, J., Ballabio, E., Chen, X.H., Kušec, R., Taylor, S., Hay, D., Tramonti, D., Saunders, N.J., Littlewood, T., Pezzella, F., Boultwood, J., Wainscoat, J.S., Hatton, C.S.R., Lawrie, C.H., 2011. MicroRNA expression in multiple myeloma is associated with genetic subtype, isotype and survival. Biol. Direct. https://doi.org/10.1186/1745-6150-6-23

Chiles, T.C., 2004. Regulation and Function of Cyclin D2 in B Lymphocyte Subsets. J. Immunol. https://doi.org/10.4049/jimmunol.173.5.2901

Choi, J.W., Han, S.W., Kwon, K.T., Kim, G.W., 2010. Early onset multiple myeloma in a patient with systemic lupus erythematosus: A case report and literature review. Clin. Rheumatol.

https://doi.org/10.1007/s10067-010-1417-3

Chretien, M.L., Corre, J., Lauwers-Cances, V., Magrangeas, F., Cleynen, A., Yon, E., Hulin, C., Leleu, X., Orsini-Piocelle, F., Blade, J.S., Sohn, C., Karlin, L., Delbrel, X., Hebraud, B., Roussel, M., Marit, G., Garderet, L., Mohty, M., Rodon, P., Voillat, L., Royer, B., Jaccard, A., Belhadj, K., Fontan, J., Caillot, D., Stoppa, A.M., Attal, M., Facon, T., Moreau, P., Minvielle, S., Avet-Loiseau, H., 2015. Understanding the role of hyperdiploidy in myeloma prognosis: Which trisomies really matter? Blood. https://doi.org/10.1182/blood-2015-06-650242

Clapier, C.R., Cairns, B.R., 2009. The Biology of Chromatin Remodeling Complexes. Annu. Rev. Biochem. https://doi.org/10.1146/annurev.biochem.77.062706.153223

Cohen, Y., Gutwein, O., Garach-Jehoshua, O., Bar-Haim, A., Kornberg, A., 2014. The proliferation arrest of primary tumor cells out-of-niche is associated with widespread downregulation of mitotic and transcriptional genes. Hematology. https://doi.org/10.1179/1607845413Y.0000000125

Comet, I., Schuettengruber, B., Sexton, T., Cavalli, G., 2011. A chromatin insulator driving three-dimensional Polycomb response element (PRE) contacts and Polycomb association with the chromatin fiber. Proc. Natl. Acad. Sci. U. S. A. 108, 2294–2299. https://doi.org/10.1073/pnas.1002059108

Condomines, M., Hose, D., Reme, T., Requirand, G., Hundemer, M., Schoenhals, M., Goldschmidt, H., Klein, B., 2009. Gene Expression Profiling and Real-Time PCR Analyses Identify Novel Potential Cancer-Testis Antigens in Multiple Myeloma. J. Immunol. https://doi.org/10.4049/jimmunol.0803298

Corces, M.R., Granja, J.M., Shams, S., Louie, B.H., Seoane, J.A., Zhou, W., Silva, T.C., Groeneveld, C., Wong, C.K., Cho, W., Satpathy, A.T., Mumbach, M.R., Hoadley, K.A., Robertson, A.G., Sheffield, N.C., Felau, I., Castro, M.A.A., Berman, B.P., Staudt, L.M., Zenklusen, J.C., Laird, P.W., 2018. The chromatin accessibility landscape of primary human cancers 1898. https://doi.org/10.1126/science.aav1898

Cornish, A.J., Hoang, P.H., Dobbins, S.E., Law, P.J., Chubb, D., Orlando, G., Houlston, R.S., 2019. Identification of recurrent noncoding mutations in B-cell lymphoma using capture Hi-C. Blood Adv. https://doi.org/10.1182/bloodadvances.2018026419

Cramer, P., 2019. Organization and regulation of gene transcription. Nature.

https://doi.org/10.1038/s41586-019-1517-4

Csizmadia, V., Hales, P., Tsu, C., Ma, J., Chen, J., Shah, P., Fleming, P., Senn, J.J., Kadambi, V.J., Dick, L., Wolenski, F.S., 2016. Proteasome inhibitors bortezomib and carfilzomib used for the treatment of multiple myeloma do not inhibit the serine protease HtrA2/Omi. Toxicol. Res. (Camb). https://doi.org/10.1039/c6tx00220j

Cunningham, T.J., Lancman, J.J., Berenguer, M., Dong, P.D.S., Duester, G., 2018. Genomic Knockout of Two Presumed Forelimb Tbx5 Enhancers Reveals They Are Nonessential for Limb Development. Cell Rep. https://doi.org/10.1016/j.celrep.2018.05.052

Dairaghi, D.J., Oyajobi, B.O., Gupta, A., McCluskey, B., Miao, S., Powers, J.P., Seitz, L.C., Wang, Y., Zeng, Y., Zhang, P., Schall, T.J., Jaen, J.C., 2012. CCR1 blockade reduces tumor burden and osteolysis in vivo in a mouse model of myeloma bone disease. Blood. https://doi.org/10.1182/blood-2011-10-384784

Dalla-Favera, R., Bregni, M., Erikson, J., Patterson, D., Gallo, R.C., Croce, C.M., 1982. Human c-myc onc gene is located on the region of chromosome 8 that is translocated in Burkitt lymphoma cells. Proc. Natl. Acad. Sci.

Davie, K., Jacobs, J., Atkins, M., Potier, D., Christiaens, V., Halder, G., Aerts, S., 2015. Discovery of Transcription Factors and Regulatory Regions Driving In Vivo Tumor Development by ATAC-seq and FAIRE-seq Open Chromatin Profiling. PLoS Genet. 11, 1–24. https://doi.org/10.1371/journal.pgen.1004994

de la Torre-Ubieta, L., Stein, J.L., Won, H., Opland, C.K., Liang, D., Lu, D., Geschwind, D.H., 2018. The Dynamic Landscape of Open Chromatin during Human Cortical Neurogenesis. Cell 172, 289-304.e18. https://doi.org/10.1016/j.cell.2017.12.014

de Wit, E., Vos, E.S.M., Holwerda, S.J.B., Valdes-Quezada, C., Verstegen, M.J.A.M., Teunissen, H., Splinter, E., Wijchers, P.J., Krijger, P.H.L., de Laat, W., 2015. CTCF Binding Polarity Determines Chromatin Looping. Mol. Cell 60, 676–684. https://doi.org/10.1016/j.molcel.2015.09.023

Decker, J.M., n.d. Northern Arizona University - B Cell Development (Immunology) [WWW Document]. URL http://www2.nau.edu/~fpm/immunology/Exams/Bcelldevelopment-401.html

Dekker, J., 2002. Capturing Chromosome Conformation. Science (80-.). 295, 1306–1311. https://doi.org/10.1126/science.1067799

Dekker, J., Mirny, L., 2016. The 3D Genome as Moderator of Chromosomal Communication. Cell 164, 1110–1121. https://doi.org/10.1016/j.cell.2016.02.007

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P.D., Dean, A., 2012. Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor. Cell 149, 1233–1244. https://doi.org/10.1016/j.cell.2012.03.051

Dent, A.L., Shaffer, A.L., Yu, X., Allman, D., Staudt, L.M., 1997. Control of inflammation, cytokine expression, and germinal center formation by BCL-6. Science (80-.). https://doi.org/10.1126/science.276.5312.589

Dettling, S., Warta, R., Rapp, C., Pocha, K., Roesch, S., Jungk, C., Unterberg, A., Herold-Mende, C., Stamova, S., Rathinasamy, A., Beckhove, P., Schnölzer, M., Warnken, U., Reuss, D., von Deimling, A., Eckstein, V., Grabe, N., Schramm, C., Weigand, M.A., 2018. Identification of CRKII, CFL1, CNTN1, NME2, and TKT as novel and frequent T-cell targets in human IDH-mutant glioma. Clin. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-17-1839

Dimopoulos, K., Søgaard Helbo, A., Fibiger Munch-Petersen, H., Sjö, L., Christensen, J., Sommer Kristensen, L., Asmar, F., Hermansen, N.E.U., O'Connel, C., Gimsing, P., Liang, G., Grønbæk, K., 2018. Dual inhibition of DNMTs and EZH2 can overcome both intrinsic and acquired resistance of myeloma cells to IMiDs in a cereblon-independent manner. Mol. Oncol. https://doi.org/10.1002/1878-0261.12157

Ding, M., Liu, Yuhan, Liao, X., Zhan, H., Liu, Yuchen, Huang, W., 2018. Enhancer RNAs (eRNAs): New Insights into Gene Transcription and Disease Treatment. J. Cancer 9, 2334–2340. https://doi.org/10.7150/jca.25829

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., Ren, B., 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485, 376–380. https://doi.org/10.1038/nature11082

Dorighi, K.M., Swigut, T., Henriques, T., Bhanu, N. V., Scruggs, B.S., Nady, N., Still, C.D., Garcia, B.A., Adelman, K., Wysocka, J., 2017. Mll3 and Mll4 Facilitate Enhancer RNA Synthesis and Transcription from Promoters Independently of H3K4 Monomethylation. Mol. Cell. https://doi.org/10.1016/j.molcel.2017.04.018

Douds, J.J., Long, D.J., Kim, A.S., Li, S., 2014. Diagnostic and prognostic significance of CD200 expression and its stability in plasma cell myeloma. J. Clin. Pathol. https://doi.org/10.1136/jclinpath-2014-202421

Dunham, I., Kundaje, A., Aldred, S.F., Collins, P.J., Davis, C.A., Doyle, F., Epstein, C.B., Frietze, S., Harrow, J., Kaul, R., Khatun, J., Lajoie, B.R., Landt, S.G., Lee, B.K., Pauli, F., Rosenbloom, K.R., Sabo, P., Safi, A., Sanyal, A., Shoresh, N., Simon, J.M., Song, L., Trinklein, N.D., Altshuler, R.C., Birney, E., Brown, J.B., Cheng, C., Djebali, S., Dong, X., Ernst, J., Furey, T.S., Gerstein, M., Giardine, B., Greven, M., Hardison, R.C., Harris, R.S., Herrero, J., Hoffman, M.M., Iyer, S., Kellis, M., Kheradpour, P., Lassmann, T., Li, Q., Lin, X., Marinov, G.K., Merkel, A., Mortazavi, A., Parker, S.C.J., Reddy, T.E., Rozowsky, J., Schlesinger, F., Thurman, R.E., Wang, J., Ward, L.D., Whitfield, T.W., Wilder, S.P., Wu, W., Xi, H.S., Yip, K.Y., Zhuang, J., Bernstein, B.E., Green, E.D., Gunter, C., Snyder, M., Pazin, M.J., Lowdon, R.F., Dillon, L.A.L., Adams, L.B., Kelly, C.J., Zhang, J., Wexler, J.R., Good, P.J., Feingold, E.A., Crawford, G.E., Dekker, J., Elnitski, L., Farnham, P.J., Giddings, M.C., Gingeras, T.R., Guigó, R., Hubbard, T.J., Kent, W.J., Lieb, J.D., Margulies, E.H., Myers, R.M., Stamatoyannopoulos, J.A., Tenenbaum, S.A., Weng, Z., White, K.P., Wold, B., Yu, Y., Wrobel, J., Risk, B.A., Gunawardena, H.P., Kuiper, H.C., Maier, C.W., Xie, L., Chen, X., Mikkelsen, T.S., Gillespie, S., Goren, A., Ram, O., Zhang, X., Wang, L., Issner, R., Coyne, M.J., Durham, T., Ku, M., Truong, T., Eaton, M.L., Dobin, A., Tanzer, A., Lagarde, J., Lin, W., Xue, C., Williams, B.A., Zaleski, C., Röder, M., Kokocinski, F., Abdelhamid, R.F., Alioto, T., Antoshechkin, I., Baer, M.T., Batut, P., Bell, I., Bell, K., Chakrabortty, S., Chrast, J., Curado, J., Derrien, T., Drenkow, J., Dumais, E., Dumais, J., Duttagupta, R., Fastuca, M., Fejes-Toth, K., Ferreira, P., Foissac, S., Fullwood, M.J., Gao, H., Gonzalez, D., Gordon, A., Howald, C., Jha, S., Johnson, R., Kapranov, P., King, B., Kingswood, C., Li, G., Luo, O.J., Park, E., Preall, J.B., Presaud, K., Ribeca, P., Robyr, D., Ruan, X., Sammeth, M., Sandhu, K.S., Schaeffer, L., See, L.H., Shahab, A., Skancke, J., Suzuki, A.M., Takahashi, H., Tilgner, H., Trout, D., Walters, N., Wang, Huaien, Hayashizaki, Y., Reymond, A., Antonarakis, S.E., Hannon, G.J., Ruan, Y., Carninci, P., Sloan, C.A., Learned, K., Malladi, V.S., Wong, M.C., Barber, G.P., Cline, M.S., Dreszer, T.R., Heitner, S.G., Karolchik, D., Kirkup, V.M., Meyer, L.R., Long, J.C., Maddren, M., Raney, B.J., Grasfeder, L.L., Giresi, P.G., Battenhouse, A., Sheffield, N.C., Showers, K.A., London, D., Bhinge, A.A., Shestak, C., Schaner, M.R., Kim, S.K., Zhang, Z.Z., Mieczkowski, P.A., Mieczkowska, J.O., Liu, Z., McDaniell, R.M., Ni, Y., Rashid, N.U., Kim, M.J., Adar, S., Zhang, Zhancheng, Wang, T., Winter, D., Keefe, D., Iyer, V.R., Zheng, M., Wang, P., Gertz, J., Vielmetter, J., Partridge, E.C., Varley, K.E., Gasper, C., Bansal, A., Pepke, S., Jain, P., Amrhein, H., Bowling, K.M., Anaya, M., Cross, M.K., Muratet, M.A., Newberry, K.M., McCue, K., Nesmith, A.S., Fisher-Aylor, K.I., Pusey, B., DeSalvo, G., Parker, S.L., Balasubramanian, Sreeram, Davis, N.S., Meadows, S.K., Eggleston, T., Newberry, J.S., Levy, S.E., Absher, D.M., Wong, W.H., Blow, M.J., Visel, A., Pennachio,

L.A., Petrykowska, H.M., Abyzov, A., Aken, B., Barrell, D., Barson, G., Berry, A., Bignell, A., Boychenko, V., Bussotti, G., Davidson, C., Despacio-Reyes, G., Diekhans, M., Ezkurdia, I., Frankish, A., Gilbert, J., Gonzalez, J.M., Griffiths, E., Harte, R., Hendrix, D.A., Hunt, T., Jungreis, I., Kay, M., Khurana, E., Leng, J., Lin, M.F., Loveland, J., Lu, Z., Manthravadi, D., Mariotti, M., Mudge, J., Mukherjee, G., Notredame, C., Pei, B., Rodriguez, J.M., Saunders, G., Sboner, A., Searle, S., Sisu, C., Snow, C., Steward, C., Tapanari, E., Tress, M.L., Van Baren, M.J., Washietl, S., Wilming, L., Zadissa, A., Zhang, Zhengdong, Brent, M., Haussler, D., Valencia, A., Addleman, N., Alexander, R.P., Auerbach, R.K., Balasubramanian, Suganthi, Bettinger, K., Bhardwaj, N., Boyle, A.P., Cao, A.R., Cayting, P., Charos, A., Cheng, Y., Eastman, C., Euskirchen, G., Fleming, J.D., Grubert, F., Habegger, L., Hariharan, M., Harmanci, A., Iyengar, S., Jin, V.X., Karczewski, K.J., Kasowski, M., Lacroute, P., Lam, H., Lamarre-Vincent, N., Lian, J., Lindahl-Allen, M., Min, R., Miotto, B., Monahan, H., Moqtaderi, Z., Mu, X.J., O'Geen, H., Ouyang, Z., Patacsil, D., Raha, D., Ramirez, L., Reed, B., Shi, M., Slifer, T., Witt, H., Wu, L., Xu, X., Yan, K.K., Yang, X., Struhl, K., Weissman, S.M., Penalva, L.O., Karmakar, S., Bhanvadia, R.R., Choudhury, A., Domanus, M., Ma, L., Moran, J., Victorsen, A., Auer, T., Centanin, L., Eichenlaub, M., Gruhl, F., Heermann, S., Hoeckendorf, B., Inoue, D., Kellner, T., Kirchmaier, S., Mueller, C., Reinhardt, R., Schertel, L., Schneider, S., Sinn, R., Wittbrodt, B., Wittbrodt, J., Jain, G., Balasundaram, G., Bates, D.L., Byron, R., Canfield, T.K., Diegel, M.J., Dunn, D., Ebersol, A.K., Frum, T., Garg, K., Gist, E., Hansen, R.S., Boatman, L., Haugen, E., Humbert, R., Johnson, A.K., Johnson, E.M., Kutyavin, T. V., Lee, K., Lotakis, D., Maurano, M.T., Neph, S.J., Neri, F. V., Nguyen, E.D., Qu, H., Reynolds, A.P., Roach, V., Rynes, E., Sanchez, M.E., Sandstrom, R.S., Shafer, A.O., Stergachis, A.B., Thomas, S., Vernot, B., Vierstra, J., Vong, S., Wang, Hao, Weaver, M.A., Yan, Y., Zhang, M., Akey, J.M., Bender, M., Dorschner, M.O., Groudine, M., MacCoss, M.J., Navas, P., Stamatoyannopoulos, G., Beal, K., Brazma, A., Flicek, P., Johnson, N., Lukk, M., Luscombe, N.M., Sobral, D., Vaquerizas, J.M., Batzoglou, S., Sidow, A., Hussami, N., Kyriazopoulou-Panagiotopoulou, S., Libbrecht, M.W., Schaub, M.A., Miller, W., Bickel, P.J., Banfai, B., Boley, N.P., Huang, H., Li, J.J., Noble, W.S., Bilmes, J.A., Buske, O.J., Sahu, A.D., Kharchenko, P. V., Park, P.J., Baker, D., Taylor, J., Lochovsky, L., 2012. An integrated encyclopedia of DNA elements in the human genome. Nature. https://doi.org/10.1038/nature11247

Ebert, A., Hill, L., Busslinger, M., 2015. Spatial Regulation of V-(D)J Recombination at Antigen Receptor Loci.

Egan, J.B., Shi, C.X., Tembe, W., Christoforides, A., Kurdoglu, A., Sinari, S., Middha, S., Asmann,

Y., Schmidt, J., Braggio, E., Keats, J.J., Fonseca, R., Bergsagel, P.L., Craig, D.W., Carpten, J.D., Stewart, A.K., 2012. Whole-genome sequencing of multiple myeloma from diagnosis to plasma cell leukemia reveals genomic initiating events, evolution, and clonal tides. Blood 120, 1060–1066. https://doi.org/10.1182/blood-2012-01-405977

ENCODE_UCSC, 2011. ENCODE UCSC Low mappability regions [WWW Document]. URL http://hgdownload.cse.ucsc.edu/goldenPath/hg19/encodeDCC/wgEncodeMapability/wgEncodeDukeMapabilityRegionsExcludable.bed.gz

ENCODE, n.d. ENCODE blacklist regions [WWW Document]. URL https://www.encodeproject.org/files/ENCFF419RSJ/@@download/ENCFF419RSJ.bed.gz

Ernst, J., Kellis, M., 2017. Chromatin-state discovery and genome annotation with ChromHMM. Nat. Protoc. https://doi.org/10.1038/nprot.2017.124

Erwin, G.D., Oksenberg, N., Truty, R.M., Kostka, D., Murphy, K.K., Ahituv, N., Pollard, K.S., Capra, J.A., 2014. Integrating Diverse Datasets Improves Developmental Enhancer Prediction. PLoS Comput. Biol. https://doi.org/10.1371/journal.pcbi.1003677

Eyboulet, F., Wydau-Dematteis, S., Eychenne, T., Alibert, O., Neil, H., Boschiero, C., Nevers, M.C., Volland, H., Cornu, D., Redeker, V., Werner, M., Soutourina, J., 2015. Mediator independently orchestrates multiple steps of preinitiation complex assembly in vivo. Nucleic Acids Res. 43, 9214–9231. https://doi.org/10.1093/nar/gkv782

Fang, Yaping, Wang, Y., Zhu, Q., Wang, J., Li, G., Erokhin, M., Vassetzky, Y., Georgiev, P., Chetverina, D., Pott, S., Lieb, J.D., Zhang, Y.B., Ishii, H., Kadonaga, J.T., Ren, B., Espinoza, C.A., Ren, B., Dixon, J.R., Mansour, M.R., Hnisz, D., Miguel-Escalada, I., Pasquali, L., Ferrer, J., Shlyueva, D., Stampfel, G., Stark, A., Andersson, R., Yue, F., Zhu, Y., Kim, T.K., Shiekhattar, R., Baumann, K., Rajagopal, N., Boyle, A.P., Visel, A., Lee, D., Fletez-Brant, C., Lee, D., McCallion, A.S., Beer, M.A., Ghandi, M., Lee, D., Mohammad-Noori, M., Beer, M.A., Podsiadlo, A., Wrzesien, M., Paja, W., Rudnicki, W., Wilczynski, B., Taher, L., Smith, R.P., Kim, M.J., Ahituv, N., Ovcharenko, I., Erwin, G.D., Whitaker, J.W., Nguyen, T.T., Zhu, Y., Wildberg, A., Wang, W., Majewski, J., Ott, J., Ritchie, M.D., Holzinger, E.R., Li, R., Pendergrass, S.A., Kim, D., Diaz-Uriarte, R., Vaquerizas, J.M., Kummerfeld, S.K., Teichmann, S.A., Luscombe, N.M., Whyte, W.A., Meysman, P., Altmann, A., Tolosi, L., Sander, O., Lengauer, T., Rivera, C.M., Ren, B., Pelish, H.E., Lin, C., Garruss, A.S., Luo, Z., Guo, F., Shilatifard, A., Wu, H., Zhang, Y., Uchimura, Y., Cuadrado, A., Remeseiro, S., Grana, O., Pisano, D.G., Losada, A., Roy, S., Shah, M., Rennoll, S.A., Raup-Konsavage,

W.M., Yochum, G.S., Tang, Z., Lee, D., Karchin, R., Beer, M.A., Ernst, J., Kellis, M., Boer, C.G. de, Liu, T., Rosenbloom, K.R., Thomas-Chollier, M., Pinello, L., Xu, J., Orkin, S.H., Yuan, G.C., Zhang, Y., Wang, X.H., Kang, L., L, B., Fang, Y., Gao, S., Tai, D., Middaugh, C.R., Fang, J., Li, Y., Fang, Y., Fang, J., Fernandez-Delgado, M., Cernadas, E., Barro, S., Amorim, D., 2016. In silico identification of enhancers on the basis of a combination of transcription factor binding motif occurrences. Sci. Rep. 6, 32476. https://doi.org/10.1038/srep32476

Feng, Y., Zhang, Y., Wei, X., Zhang, Q., 2019. Correlations of DKK1 with pathogenesis and prognosis of human multiple myeloma. Cancer Biomarkers. https://doi.org/10.3233/CBM-181909

Feuk, L., Carson, A.R., Scherer, S.W., 2006. Structural variation in the human genome. Nat. Rev. Genet. https://doi.org/10.1038/nrg1767

Ford, E., Grimmer, M.R., Stolzenburg, S., Bogdanovic, O., Mendoza, A. de, Farnham, P.J., Blancafort, P., Lister, R., 2017. Frequent lack of repressive capacity of promoter DNA methylation identified through genome-wide epigenomic manipulation. bioRxiv. https://doi.org/10.1101/170506

Fritz, Andrew J, Sehgal, N., Pliss, A., Xu, J., Berezney, R., 2019. Chromosome territories and the global regulation of the genome. Genes, Chromosom. Cancer 0. https://doi.org/10.1002/gcc.22732

Fritz, Andrew J., Sehgal, N., Pliss, A., Xu, J., Berezney, R., 2019. Chromosome territories and the global regulation of the genome. Genes, Chromosom. Cancer. https://doi.org/10.1002/gcc.22732

Fu, Y., Tessneer, K.L., Li, C., Gaffney, P.M., 2018. From association to mechanism in complex disease genetics: The role of the 3D genome 06 Biological Sciences 0604 Genetics. Arthritis Res. Ther. 20, 1–10. https://doi.org/10.1186/s13075-018-1721-x

Fudenberg, G., Imakaev, M., Lu, C., Goloborodko, A., Abdennur, N., Mirny, L.A., 2015. Formation of Chromosomal Domains by Loop Extrusion. bioRxiv 024620. https://doi.org/10.1101/024620

Fukuda, N., Saitoh, M., Kobayashi, N., Miyazono, K., 2006. Execution of BMP-4-induced apoptosis by p53-dependent ER dysfunction in myeloma and B-cell hybridoma cells. Oncogene. https://doi.org/10.1038/sj.onc.1209393

Fukuda, T., Yoshida, T., Okada, S., Hatano, M., Miki, T., Ishibashi, K., Okabe, S., Koseki, H., Hirosawa, S., Taniguchi, M., Miyasaka, N., Tokuhisa, T., 1997. Disruption of the Bcl6 gene results in an impaired germinal center formation. J. Exp. Med. https://doi.org/10.1084/jem.186.3.439

Fulciniti, M., Lin, C.Y., Samur, M.K., Lopez, M.A., Singh, I., Lawlor, M.A., Szalat, R.E., Ott, C.J., Avet-Loiseau, H., Anderson, K.C., Young, R.A., Bradner, J.E., Munshi, N.C., 2018. Non-overlapping Control of Transcriptome by Promoter- and Super-Enhancer-Associated Dependencies in Multiple Myeloma. Cell Rep. https://doi.org/10.1016/j.celrep.2018.12.016

Ganji, M., Shaltiel, I.A., Bisht, S., Kim, E., Kalichava, A., Haering, C.H., Dekker, C., 2018. Real-time imaging of DNA loop extrusion by condensin. Science (80-.). https://doi.org/10.1126/science.aar7831

García-Carpizo, V., Sarmentero, J., Han, B., Graña, O., Ruiz-Llorente, S., Pisano, D.G., Serrano, M., Brooks, H.B., Campbell, R.M., Barrero, M.J., 2016. NSD2 contributes to oncogenic RAS-driven transcription in lung cancer cells through long-range epigenetic activation. Sci. Rep. https://doi.org/10.1038/srep32952

García-González, E., Escamilla-Del-Arenal, M., Arzate-Mejía, R., Recillas-Targa, F., 2016. Chromatin remodeling effects on enhancer activity. Cell. Mol. Life Sci. https://doi.org/10.1007/s00018-016-2184-3

Gasperini, M., Hill, A.J., McFaline-Figueroa, J.L., Martin, B., Kim, S., Zhang, M.D., Jackson, D., Leith, A., Schreiber, J., Noble, W.S., Trapnell, C., Ahituv, N., Shendure, J., 2019. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. Cell. https://doi.org/10.1016/j.cell.2018.11.029

Ghafouri-Fard, S., Seifi-Alan, M., Shamsi, R., Esfandiary, A., 2015. Immunotherapy in multiple myeloma using cancer-testis antigens. Int. J. Cancer Manag. https://doi.org/10.17795/ijcp-3755

Ghavi-Helm, Y., Klein, F.A., Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., Furlong, E.E.M., 2014. Enhancer loops appear stable during development and are associated with paused polymerase. Nature 512, 96–100. https://doi.org/10.1038/nature13417

Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., Dekker, J., 2010. Cell-type-specific long-range looping interactions identify distant

regulatory elements of the CFTR gene 38, 4325–4336. https://doi.org/10.1093/nar/gkq175

Gibcus, J.H., Dekker, J., 2013. The Hierarchy of the 3D Genome. Mol. Cell 49, 773–782. https://doi.org/10.1016/j.molcel.2013.02.011

Gilcrease, M.Z., 2015. How Many Etiological Subtypes of Breast Cancer: Two, Three, Four, or More? Breast Dis. https://doi.org/10.1016/j.breastdis.2015.10.037

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., Lieb, J.D., 2007. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. https://doi.org/10.1101/gr.5533506

Glavey, S. V., Naba, A., Manier, S., Clauser, K., Tahri, S., Park, J., Reagan, M.R., Moschetta, M., Mishima, Y., Gambella, M., Rocci, A., Sacco, A., O'Dwyer, M.E., Asara, J.M., Palumbo, A., Roccaro, A.M., Hynes, R.O., Ghobrial, I.M., 2017. Proteomic characterization of human multiple myeloma bone marrow extracellular matrix. Leukemia 31, 2426–2434. https://doi.org/10.1038/leu.2017.102

Goldsmith, S.R., Fiala, M.A., O'Neal, J., Souroullas, G.P., Toama, W., Vij, R., Schroeder, M.A., 2019. EZH2 Overexpression in Multiple Myeloma: Prognostic Value, Correlation With Clinical Characteristics, and Possible Mechanisms. Clin. Lymphoma, Myeloma Leuk. https://doi.org/10.1016/j.clml.2019.08.010

Goloborodko, A., Imakaev, M. V, Marko, J.F., Mirny, L., 2016. Compaction and segregation of sister chromatids via active loop extrusion 1–20. https://doi.org/10.1101/038281

Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S., 1999. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. Science (80-.). https://doi.org/10.1126/science.286.5439.531

Gómez-Marín, C., Tena, J.J., Acemel, R.D., López-Mayorga, M., Naranjo, S., de la Calle-Mustienes, E., Maeso, I., Beccari, L., Aneas, I., Vielmas, E., Bovolenta, P., Nobrega, M. a., Carvajal, J., Gómez-Skarmeta, J.L., 2015. Evolutionary comparison reveals that diverging CTCF sites are signatures of ancestral topological associating domains borders. Proc Natl Acad Sci U S A 112, 201505463. https://doi.org/10.1073/pnas.1505463112

Gorkin, D.U., Leung, D., Ren, B., 2014. The 3D genome in transcriptional regulation and pluripotency. Cell Stem Cell 14, 771–775. https://doi.org/10.1016/j.stem.2014.05.017

Grab, A.L., Seckinger, A., Horn, P., Hose, D., Cavalcanti-Adam, E.A., 2019. Hyaluronan hydrogels delivering BMP-6 for local targeting of malignant plasma cells and osteogenic differentiation of mesenchymal stromal cells. Acta Biomater. https://doi.org/10.1016/j.actbio.2019.07.018

Grčević, D., Kušec, R., Kovačić, N., Lukić, A., Lukić, I.K., Ivčević, S., Nemet, D., Seiwerth, R.S., Ostojić, S.K., Croucher, P.I., Marušić, A., 2010. Bone morphogenetic proteins and receptors are over-expressed in bone-marrow cells of multiple myeloma patients and support myeloma cells by inducing ID genes. Leuk. Res. https://doi.org/10.1016/j.leukres.2009.10.016

Gröschel, S., Sanders, M.A., Hoogenboezem, R., De Wit, E., Bouwman, B.A.M., Erpelinck, C., Van Der Velden, V.H.J., Havermans, M., Avellino, R., Van Lom, K., Rombouts, E.J., Van Duin, M., Döhner, K., Beverloo, H.B., Bradner, J.E., Döhner, H., Löwenberg, B., Valk, P.J.M., Bindels, E.M.J., De Laat, W., Delwel, R., 2014. A single oncogenic enhancer rearrangement causes concomitant EVI1 and GATA2 deregulation in Leukemia. Cell. https://doi.org/10.1016/j.cell.2014.02.019

Gu, X., Coates, P.J., Boldrup, L., Wang, L., Krejci, A., Hupp, T., Fahraeus, R., Norberg-Spaak, L., Sgaramella, N., Wilms, T., Nylander, K., 2018. Copy number variation: A prognostic marker for young patients with squamous cell carcinoma of the oral tongue. J. Oral Pathol. Med. 0. https://doi.org/10.1111/jop.12792

Guo, P., Yang, J., Huang, J., Auguste, D.T., Moses, M.A., 2019. Therapeutic genome editing of triple-negative breast tumors using a noncationic and deformable nanolipogel. Proc. Natl. Acad. Sci. 116, 18295–18303. https://doi.org/10.1073/PNAS.1904697116

Gupta, S., Stamatoyannopoulos, J.A., Bailey, T.L., Noble, W.S., 2007. Quantifying similarity between motifs. Genome Biol. https://doi.org/10.1186/gb-2007-8-2-r24

Hait, T.A., Amar, D., Shamir, R., Elkon, R., 2018. FOCS: A novel method for analyzing enhancer and gene activity patterns infers an extensive enhancer-promoter map. Genome Biol. https://doi.org/10.1186/s13059-018-1432-2

Hansen, A.S., Cattoglio, C., Darzacq, X., Tjian, R., 2017. Recent evidence that TADs and chromatin loops are dynamic structures. Nucleus 1034, 1–13. https://doi.org/10.1080/19491034.2017.1389365

Hanssen, L.L.P., Kassouf, M.T., Oudelaar, A.M., Biggs, D., Preece, C., Downes, D.J., Gosden, M.,

Sharpe, J.A., Sloane-Stanley, J.A., Hughes, J.R., Davies, B., Higgs, D.R., 2017. Tissue-specific CTCF-cohesin-mediated chromatin architecture delimits enhancer interactions and function in vivo. Nat. Cell Biol. https://doi.org/10.1038/ncb3573

Heinig, M., Colomé-Tatché, M., Taudt, A., Rintisch, C., Schafer, S., Pravenec, M., Hubner, N., Vingron, M., Johannes, F., 2015. histoneHMM: Differential analysis of histone modifications with broad genomic footprints. BMC Bioinformatics. https://doi.org/10.1186/s12859-015-0491-6

Heller, G., Schmidt, W.M., Ziegler, B., Holzer, S., Müllauer, L., Bilban, M., Zielinski, C.C., Drach, J., Zöchbauer-Müller, S., 2008. Genome-wide transcriptional response to 5-Aza-2'-deoxycytidine and trichostatin A in multiple myeloma cells. Cancer Res. https://doi.org/10.1158/0008-5472.CAN-07-2531

Henriques, T., Scruggs, B.S., Inouye, M.O., Muse, G.W., Williams, L.H., Burkholder, A.B., Lavender, C.A., Fargo, D.C., Adelman, K., 2018. Widespread transcriptional pausing and elongation control at enhancers. Genes Dev. 32, 26–41. https://doi.org/10.1101/gad.309351.117

Hernández-García, S., San-Segundo, L., González-Méndez, L., Corchete, L.A., Misiewicz-Krzeminska, I., Martín-Sánchez, M., López-Iglesias, A.A., Algarín, E.M., Mogollón, P., Díaz-Tejedor, A., Paíno, T., Tunquist, B., Mateos, M.V., Gutiérrez, N.C., Díaz-Rodriguez, E., Garayoa, M., Ocio, E.M., 2017. The kinesin spindle protein inhibitor filanesib enhances the activity of pomalidomide and dexamethasone in multiple myeloma. Haematologica. https://doi.org/10.3324/haematol.2017.168666

Herranz, D., Ambesi-Impiombato, A., Palomero, T., Schnell, S.A., Belver, L., Wendorff, A.A., Xu, L., Castillo-Martin, M., Llobet-Navás, D., Cordon-Cardo, C., Clappier, E., Soulier, J., Ferrando, A.A., 2014. A NOTCH1-driven MYC enhancer promotes T cell development, transformation and acute lymphoblastic leukemia. Nat. Med. 20, 1130–1137. https://doi.org/10.1038/nm.3665

Heuston, E.F., Keller, C.A., Lichtenberg, J., Giardine, B., Anderson, S.M., Hardison, R.C., Bodine, D.M., 2018. Establishment of regulatory elements during erythro-megakaryopoiesis identifies hematopoietic lineage-commitment points. Epigenetics and Chromatin. https://doi.org/10.1186/s13072-018-0195-z

Hideshima, T., Mitsiades, C., Ikeda, H., Chauhan, D., Raje, N., Gorgun, G., Hideshima, H., Munshi, N.C., Richardson, P.G., Carrasco, D.R., Anderson, K.C., 2010. A proto-oncogene

BCL6 is up-regulated in the bone marrow microenvironment in multiple myeloma cells. Blood. https://doi.org/10.1182/blood-2010-02-270082

Hoang, P.H., Dobbins, S.E., Cornish, A.J., Chubb, D., Law, P.J., Kaiser, M., Houlston, R.S., 2018. Whole-genome sequencing of multiple myeloma reveals oncogenic pathways are targeted somatically through multiple mechanisms. Leukemia 32, 1–12. https://doi.org/10.1038/s41375-018-0103-3

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J.A., Birney, E., Hardison, R.C., Dunham, I., Kellis, M., Noble, W.S., 2013. Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. 41, 827–841. https://doi.org/10.1093/nar/gks1284

Holstein, S.A., Avet-Loiseau, H., Hahn, T., Ho, C.M., Lohr, J.G., Munshi, N.C., Paiva, B., Pasquini, M.C., Tario, J.D., Usmani, S.Z., Wallace, P.K., Weisel, K., McCarthy, P.L., 2018. BMT CTN Myeloma Intergroup Workshop on Minimal Residual Disease and Immune Profiling: Summary and Recommendations from the Organizing Committee. Biol. Blood Marrow Transplant. https://doi.org/10.1016/j.bbmt.2017.12.774

Hope, C., Foulcer, S., Jagodinsky, J., Chen, S.X., Jensen, J.L., Patel, S., Leith, C., Maroulakou, I., Callander, N., Miyamoto, S., Hematti, P., Apte, S.S., Asimakopoulos, F., 2016. Immunoregulatory roles of versican proteolysis in the myeloma microenvironment. Blood. https://doi.org/10.1182/blood-2016-03-705780

Hosen, N., 2020. Integrins in multiple myeloma. Inflamm. Regen. https://doi.org/10.1186/s41232-020-00113-y

Hou, C., Dale, R., Dean, A., 2010. Cell type specificity of chromatin organization mediated by CTCF and cohesin. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.0912087107

Hu, Q., Luo, Z., Xu, T., Zhang, J.Y., Zhu, Y., Chen, W.X., Zhong, S.L., Zhao, J.H., Tang, J.H., 2014. FOXA1: A promising prognostic marker in breast cancer. Asian Pacific J. Cancer Prev. https://doi.org/10.7314/APJCP.2014.15.1.11

Huang, X., Wang, Y., Nan, X., He, S., Xu, X., Zhu, X., Tang, J., Yang, X., Yao, L., Wang, X., Cheng, C., 2014. The role of the orphan G protein-coupled receptor 37 (GPR37) in multiple myeloma cells. Leuk. Res. https://doi.org/10.1016/j.leukres.2013.11.007

Hurt, E.M., Wiestner, A., Rosenwald, A., Shaffer, A.L., Campo, E., Grogan, T., Bergsagel, P.L., Kuehl, W.M., Staudt, L.M., 2004. Overexpression of c-maf is a frequent oncogenic event

in multiple myeloma that promotes proliferation and pathological interactions with bone marrow stroma. Cancer Cell. https://doi.org/10.1016/S1535-6108(04)00019-4

Hussain, S., Bedekovics, T., Chesi, M., Bergsagel, P.L., Galardy, P.J., 2015. UCHL1 is a biomarker of aggressive multiple myeloma required for disease progression. Oncotarget. https://doi.org/10.18632/oncotarget.5727

Hynes R O., 2002. Integrins: bidirectional, allosteric signaling machines. Cell.

Iborra, F.J., Pombo, A., Jackson, D.A., Cook, P.R., 1996. Active RNA polymerases are localized within discrete transcription "factories'' in human nuclei." J. Cell Sci. https://doi.org/10.1021/acs.analchem.8b00162

Ishida, T., Asamitsu, K., Wong, R.W.J., Zhang, T., Sanda, T., Yam, A.W.Y., Ueda, R., Gray, N.S., Iida, S., Okamoto, T., Leong, W.Z., Ngoc, P.C.T., 2017. Enhancer profiling identifies critical cancer genes and characterizes cell identity in adult T-cell leukemia. Blood 130, 2326–2338. https://doi.org/10.1182/blood-2017-06-792184

Jacobs, J., Atkins, M., Davie, K., Imrichova, H., Romanelli, L., Christiaens, V., Hulselmans, G., Potier, D., Wouters, J., Taskiran, I.I., Paciello, G., González-Blas, C.B., Koldere, D., Aibar, S., Halder, G., Aerts, S., 2018. The transcription factor Grainy head primes epithelial enhancers for spatiotemporal activation by displacing nucleosomes. Nat. Genet. https://doi.org/10.1038/s41588-018-0140-x

Jang, J.S., Li, Y., Mitra, A.K., Bi, L., Abyzov, A., van Wijnen, A.J., Baughn, L.B., Van Ness, B., Rajkumar, V., Kumar, S., Jen, J., 2019. Molecular signatures of multiple myeloma progression through single cell RNA-Seq. Blood Cancer J. https://doi.org/10.1038/s41408-018-0160-x

Javierre, B.M., Sewitz, S., Cairns, J., Wingett, S.W., V??rnai, C., Thiecke, M.J., Freire-Pritchett, P., Spivakov, M., Fraser, P., Burren, O.S., Cutler, A.J., Todd, J.A., Wallace, C., Wilder, S.P., Kreuzhuber, R., Kostadima, M., Zerbino, D.R., Stegle, O., Kreuzhuber, R., Burden, F., Farrow, S., Rehnstr??m, K., Downes, K., Grassi, L., Kostadima, M., Ouwehand, W.H., Frontini, M., Kreuzhuber, R., Burden, F., Farrow, S., Rehnstr??m, K., Downes, K., Grassi, L., Kostadima, M., Ouwehand, W.H., Frontini, M., Hill, S.M., Wang, F., Wallace, C., Stunnenberg, H.G., Ouwehand, W.H., Frontini, M., Ouwehand, W.H., Wallace, C., Martens, J.H., Kim, B., Sharifi, N., Janssen-Megens, E.M., Yaspo, M.L., Linser, M., Kovacsovics, A., Clarke, L., Richardson, D., Datta, A., Flicek, P., 2016. Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene

Promoters. Cell 167, 1369-1384.e19. https://doi.org/10.1016/j.cell.2016.09.037

Jiang, Y., Saga, K., Miyamoto, Y., Kaneda, Y., 2016. Cytoplasmic calcium increase via fusion with inactivated Sendai virus induces apoptosis in human multiple myeloma cells by downregulation of c-Myc oncogene. Oncotarget. https://doi.org/10.18632/oncotarget.9105

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C., Schmitt, A.D., Espinoza, C.A., Ren, B., 2013. A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature 503, 290–294. https://doi.org/10.1038/nature12644

Jin, J., Wang, T., Wang, Y., Chen, S., Li, Z., Li, X., Zhang, J., Wang, J., 2017. SRC3 expressed in BMSCs promotes growth and migration of multiple myeloma cells by regulating the expression of Cx43. Int. J. Oncol. https://doi.org/10.3892/ijo.2017.4171

Jin, Y., Chen, K., De Paepe, A., Hellqvist, E., Krstic, A.D., Metang, L., Gustafsson, C., Davis, R.E., Levy, Y.M., Surapaneni, R., Wallblom, A., Nahi, H., Mansson, R., Lin, Y.C., 2018. Active enhancer and chromatin accessibility landscapes chart the regulatory network of primary multiple myeloma. Blood 131, 2138–2150. https://doi.org/10.1182/blood-2017-09-808063

Johnson, D.C., Corthals, S.L., Walker, B.A., Ross, F.M., Gregory, W.M., Dickens, N.J., Lokhorst, H.M., Goldschmidt, H., Davies, F.E., Durie, B.G.M., Van Ness, B., Child, J.A., Sonneveld, P., Morgan, G.J., 2011. Genetic factors underlying the risk of thalidomide-related neuropathy in patients with multiple myeloma. J. Clin. Oncol. https://doi.org/10.1200/JCO.2010.28.0792

Joshi, N., Fass, J., 2011. sickle - A windowed adaptive trimming tool for FASTQ files using quality. (Version 1.33). https://doi.org/10.1088/0022-3727/13/9/001

Kaiser, M.F., Walker, B.A., Hockley, S.L., Begum, D.B., Wardell, C.P., Gonzalez, D., Ross, F.M., Davies, F.E., Morgan, G.J., 2013. A TC classification-based predictor for multiple myeloma using multiplexed real-time quantitative PCR. Leukemia. https://doi.org/10.1038/leu.2013.12

Kalff, A., Spencer, A., 2012. The t(4;14) translocation and FGFR3 overexpression in multiple myeloma: Prognostic implications and current clinical strategies. Blood Cancer J. https://doi.org/10.1038/bcj.2012.37

Kalverda, B., Fornerod, M., 2010. Characterization of genome-nucleoporin interactions in

Drosophila links chromatin insulators to the nuclear pore complex. Cell Cycle 9, 4812–4817. https://doi.org/10.4161/cc.9.24.14328

Kang, Y., Kim, Y.W., Kang, J., Yun, W.J., Kim, A.R., 2017. Erythroid specific activator GATA-1-dependent interactions between CTCF sites around the β-globin locus. Biochim. Biophys. Acta - Gene Regul. Mech. https://doi.org/10.1016/j.bbagrm.2017.01.013

Kassambara, A., Hose, D., Moreaux, J., Walker, B.A., Rotopopov, A., Reme, T., Pellestor, F., Pantesco, V., Jauch, A., Morgan, G., Goldschmidt, H., Klein, B., 2012. Genes with a spike expression are clustered in chromosome (sub) bands and spike (sub) bands have a powerful prognostic value in patients with multiple Myeloma. Haematologica. https://doi.org/10.3324/haematol.2011.046821

Katagiri, T., Watabe, T., 2016. Bone Morphogenetic Proteins. Cold Spring Harb. Perspect. Biol. https://doi.org/10.1101/cshperspect.a021899

Katainen, R., Dave, K., Pitkänen, E., Palin, K., Kivioja, T., Välimäki, N., Gylfe, A.E., Ristolainen, H., Hänninen, U.A., Cajuso, T., Kondelin, J., Tanskanen, T., Mecklin, J.P., Järvinen, H., Renkonen-Sinisalo, L., Lepistö, A., Kaasinen, E., Kilpivaara, O., Tuupanen, S., Enge, M., Taipale, J., Aaltonen, L.A., 2015. CTCF/cohesin-binding sites are frequently mutated in cancer. Nat. Genet. https://doi.org/10.1038/ng.3335

Katayama, S., Suzuki, M., Yamaoka, A., Keleku-Lukwete, N., Katsuoka, F., Otsuki, A., Kure, S., Engel, J.D., Yamamoto, M., 2017. GATA2 haploinsufficiency accelerates EVI1-driven leukemogenesis. Blood. https://doi.org/10.1182/blood-2016-12-756767

Kheradpour, P., Stark, A., Roy, S., Kellis, M., 2007. Reliable prediction of regulator targets using 12 Drosophila genomes 1919–1931. https://doi.org/10.1101/gr.7090407.

Kim, D., Langmead, B., Salzberg, S.L., 2015. HISAT: A fast spliced aligner with low memory requirements. Nat. Methods. https://doi.org/10.1038/nmeth.3317

Kim, S., Yu, N.-K., Kaang, B.-K., 2015. CTCF as a multifunctional protein in genome regulation and gene expression. Exp. Mol. Med. 47, e166. https://doi.org/10.1038/emm.2015.33

Kim, S.Y., Min, H.J., Park, H.K., Oh, B., Kim, T.Y., She, C.J., Hwang, S.M., Kim, M., Kim, H.K., Kim, I., Yoon, S.S., Park, S., Kim, B.K., Lee, J.H., Lee, D.S., 2011. Increased Copy Number of the Interleukin-6 Receptor Gene Is Associated with Adverse Survival in Multiple Myeloma Patients Treated with Autologous Stem Cell Transplantation. Biol. Blood Marrow Transplant. https://doi.org/10.1016/j.bbmt.2011.01.002

Kleftogiannis, D., Kalnis, P., Bajic, V.B., 2015. Progress and challenges in bioinformatics approaches for enhancer identification. Brief. Bioinform. bbv101-. https://doi.org/10.1093/bib/bbv101

Klein, U., Casola, S., Cattoretti, G., Shen, Q., Lia, M., Mo, T., Ludwig, T., Rajewsky, K., Dalla-Favera, R., 2006. Transcription factor IRF4 controls plasma cell differentiation and class-switch recombination. Nat. Immunol. https://doi.org/10.1038/ni1357

Klemm, S.L., Shipony, Z., Greenleaf, W.J., 2019. Chromatin accessibility and the regulatory epigenome. Nat. Rev. Genet. https://doi.org/10.1038/s41576-018-0089-8

Knief, J., Reddemann, K., Gliemroth, J., Brede, S., Bartscht, T., Thorns, C., 2017. ERG expression in multiple myeloma—A potential diagnostic pitfall. Pathol. Res. Pract. https://doi.org/10.1016/j.prp.2016.10.014

Kodzius, R., Kojima, M., Nishiyori, H., Nakamura, M., Fukuda, S., Tagami, M., Sasaki, D., Imamura, K., Kai, C., Harbers, M., Hayashizaki, Y., Carninci, P., 2006. CAGE: cap analysis of gene expression. Nat. Methods 3, 211.

Koenecke, N., Johnston, J., Gaertner, B., Natarajan, M., Zeitlinger, J., 2016. Genome-wide identification of Drosophila dorso-ventral enhancers by differential histone acetylation analysis. Genome Biol. https://doi.org/10.1186/s13059-016-1057-2

Korthauer, K., Irizarry, R.A., 2018. Genome-wide repressive capacity of promoter DNA methylation is revealed through epigenomic manipulation. bioRxiv. https://doi.org/10.1101/381145

Koues, O.I., Kowalewski, R.A., Chang, L.W., Pyfrom, S.C., Schmidt, J.A., Luo, H., Sandoval, L.E., Hughes, T.B., Bednarski, J.J., Cashen, A.F., Payton, J.E., Oltz, E.M., 2015. Enhancer Sequence Variants and Transcription-Factor Deregulation Synergize to Construct Pathogenic Regulatory Circuits in B-Cell Lymphoma. Immunity 42, 186–198. https://doi.org/10.1016/j.immuni.2014.12.021

Kubota, S., Tokunaga, K., Umezu, T., Yokomizo-Nakano, T., Sun, Y., Oshima, M., Tan, K.T., Yang, H., Kanai, A., Iwanaga, E., Asou, N., Maeda, T., Nakagata, N., Iwama, A., Ohyashiki, K., Osato, M., Sashida, G., 2019. Lineage-specific RUNX2 super-enhancer activates MYC and promotes the development of blastic plasmacytoid dendritic cell neoplasm. Nat. Commun. 10, 1653. https://doi.org/10.1038/s41467-019-09710-z

Kuehl, W., Bergsagel, P., 2012. Molecular pathogenesis of multiple myeloma and its

premalignant precursor. J. Clin. Invest. 122, 3456–3463. https://doi.org/10.1172/JCI61188.3456

Kuijjer, M.L., Paulson, J.N., Salzman, P., Ding, W., Quackenbush, J., 2018. Cancer subtype identification using somatic mutation data. Br. J. Cancer. https://doi.org/10.1038/s41416-018-0109-7

Kulakovskiy, I. V., Vorontsov, I.E., Yevshin, I.S., Soboleva, A. V., Kasianov, A.S., Ashoor, H., Ba-Alawi, W., Bajic, V.B., Medvedeva, Y.A., Kolpakov, F.A., Makeev, V.J., 2016. HOCOMOCO: Expansion and enhancement of the collection of transcription factor binding sites models. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv1249

Kumar Mishra, N., Guda, C., 2017. Genome-wide DNA methylation analysis reveals molecular subtypes of pancreatic cancer. Oncotarget. https://doi.org/10.18632/oncotarget.15993

Kumasaka, N., Knights, A.J., Gaffney, D.J., 2018. High-resolution genetic mapping of putative causal interactions between regions of open chromatin. Nat. Genet. 51. https://doi.org/10.1038/s41588-018-0278-6

Kundaje, A. (Stanford U., 2019a. ENCODE atac-seq guidelines [WWW Document]. URL https://www.encodeproject.org/atac-seq/

Kundaje, A. (Stanford U., 2019b. Kundaje lab ATAC-seq pipeline guidelines [WWW Document]. URL https://www.encodeproject.org/pipelines/ENCPL792NWO/

Kundaje, A. (Stanford U., 2016. Removing multimappers script.

Kuo, A.J., Cheung, P., Chen, K., Zee, B.M., Kioi, M., Lauring, J., Xi, Y., Park, B.H., Shi, X., Garcia, B.A., Li, W., Gozani, O., 2011. NSD2 Links Dimethylation of Histone H3 at Lysine 36 to Oncogenic Programming. Mol. Cell. https://doi.org/10.1016/j.molcel.2011.08.042

Laganà, A., Perumal, D., Melnekoff, D., Readhead, B., Kidd, B.A., Leshchenko, V., Kuo, P.Y., Keats, J., DeRome, M., Yesil, J., Auclair, D., Lonial, S., Chari, A., Cho, H.J., Barlogie, B., Jagannath, S., Dudley, J.T., Parekh, S., 2018. Integrative network analysis identifies novel drivers of pathogenesis and progression in newly diagnosed multiple myeloma. Leukemia. https://doi.org/10.1038/leu.2017.197

Lagler, C., El-Mesery, M., Kübler, A.C., Müller-Richter, U.D.A., Stühmer, T., Nickel, J., Müller, T.D., Wajant, H., Seher, A., 2017. The anti-myeloma activity of bone morphogenetic protein 2 predominantly relies on the induction of growth arrest and is apoptosis-independent. PLoS One. https://doi.org/10.1371/journal.pone.0185720

Landgren, O., Kyle, R. a, Pfeiffer, R.M., Katzmann, J. a, Caporaso, N.E., Richard, B., Dispenzieri, A., Kumar, S., Clark, R.J., Baris, D., Hoover, R., Vincent, S., Hayes, R.B., Rajkumar, S.V., 2014. consistently precedes multiple myeloma : a prospective study Monoclonal gammopathy of undetermined significance (MGUS) consistently precedes multiple myeloma : a prospective study 113, 5412–5417. https://doi.org/10.1182/blood-2008-12-194241

Langmead, B., Salzberg, S.L., 2012. Fast gapped-read alignment with Bowtie 2. Nat. Methods. https://doi.org/10.1038/nmeth.1923

Lara-Astiaso, D., Weiner, A., Lorenzo-Vivas, E., Zaretsky, I., Jaitin, D.A., David, E., Keren-Shaul, H., Mildner, A., Winter, D., Jung, S., Friedman, N., Amit, I., 2014. Chromatin state dynamics during blood formation. Science. https://doi.org/10.1126/science.1256271

Lea, A.J., Vockley, C.M., Johnston, R.A., Del Carpio, C.A., Barreiro, L.B., Reddy, T.E., Tung, J., 2017. Genome-wide quantification of the effects of DNA methylation on human gene regulation. bioRxiv 146829. https://doi.org/10.1101/146829

Leblanc, B., Hoichman, M., Sexton, T., Yaffe, E., Kenigsberg, E., Parrinello, H., Tanay, A., Cavalli, G., 2012. Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome. https://doi.org/10.1016/j.cell.2012.01.010

Lei, Z., Shi, H., Li, W., Yu, D., Shen, F., Yu, X., Lu, D., Sun, C., Liao, K., 2018. MiR-185 inhibits non-small cell lung cancer cell proliferation and invasion through targeting of SOX9 and regulation of Wnt signaling. Mol. Med. Rep. https://doi.org/10.3892/mmr.2017.8050

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., de Graaff, E., 2003. A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. 12, 1725–1735. https://doi.org/10.1093/hmg/ddg180

Lhoumaud, P., Badri, S., Rodriguez-Hernaez, J., Sakellaropoulos, T., Sethia, G., Kloetgen, A., Cornwell, M.I., Bhattacharyya, S., Ay, F., Bonneau, R., Tsirigos, A., Skok, J.A., 2019. NSD2 overexpression drives clustered chromatin and transcriptional changes in a subset of insulated domains. Nat. Commun. https://doi.org/10.1038/s41467-019-12811-4

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R., 2009. The Sequence Alignment/Map format and SAMtools. Bioinformatics. https://doi.org/10.1093/bioinformatics/btp352

Li, H.B., Ohno, K., Gui, H., Pirrotta, V., 2013. Insulators Target Active Genes to Transcription Factories and Polycomb-Repressed Genes to Polycomb Bodies. PLoS Genet. 9. https://doi.org/10.1371/journal.pgen.1003436

Li, N., Johnson, D.C., Weinhold, N., Kimber, S., Dobbins, S.E., Mitchell, J.S., Kinnersley, B., Sud, A., Law, P.J., Orlando, G., Scales, M., Wardell, C.P., Försti, A., Hoang, P.H., Went, M., Holroyd, A., Hariri, F., Pastinen, T., Meissner, T., Goldschmidt, H., Hemminki, K., Morgan, G.J., Kaiser, M., Houlston, R.S., 2017. Genetic Predisposition to Multiple Myeloma at 5q15 Is Mediated by an ELL2 Enhancer Polymorphism. Cell Rep. 20, 2556–2564. https://doi.org/10.1016/j.celrep.2017.08.062

Li, Q. yu, Chen, L., Hu, N., Zhao, H., 2018. Long non-coding RNA FEZF1-AS1 promotes cell growth in multiple myeloma via miR-610/Akt3 axis. Biomed. Pharmacother. https://doi.org/10.1016/j.biopha.2018.04.094

Li, X.Y., MacArthur, S., Bourgon, R., Nix, D., Pollard, D.A., Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., Hou, C.C., Ogawa, N., Inwood, W., Sementchenko, V., Beaton, A., Weiszmann, R., Celniker, S.E., Knowles, D.W., Gingeras, T., Speed, T.P., Eisen, M.B., Biggin, M.D., 2008. Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. 6, 0365–0388. https://doi.org/10.1371/journal.pbio.0060027

Lieber, M.R., 1996. Mechanistic constraints on diversity in human V (D) J recombination . These include : Mechanistic Constraints on Diversity in Human V (D) J Recombination 16, 258–269.

Lieberman-Aiden, E., Berkum, N. van, 2009. Comprehensive mapping of long range interactions reveals folding principles of the human genome. Science (80-.). 326, 289–293. https://doi.org/10.1126/science.1181369.Comprehensive

Liiu, Y., Ding, M., Gao, Q., He, A., Liu, Y., Mei, H., 2018. Current Advances on the Important Roles of Enhancer RNAs in Gene Regulation and Cancer. Biomed Res. Int. 2018. https://doi.org/10.1155/2018/2405351

Lin, L., Wang, P., Liu, X., Zhao, D., Zhang, Y., Hao, J., Liang, X., Huang, X., Lu, J., Ge, Q., 2017a. Epigenetic regulation of reelin expression in multiple myeloma. Hematol. Oncol. https://doi.org/10.1002/hon.2311

Lin, L., Yan, F., Zhao, D., Lv, M., Liang, X., Dai, H., Qin, X., Zhang, Y., Hao, J., Sun, X., Yin, Y.,

Huang, X., Zhang, J., Lu, J., Ge, Q., 2016. Reelin promotes the adhesion and drug resistance of multiple myeloma cells via integrin β1 signaling and STAT3. Oncotarget. https://doi.org/10.18632/oncotarget.7151

Lin, L., Zhang, X., Cao, L., An, Q., Hao, J., Zhang, Y., Jin, R., Chang, Y., Huang, X., Lu, J., Ge, Q., 2017b. Reelin promotes adhesion of multiple myeloma cells to bone marrow stromal cells via integrin β1 signaling. J. Cancer. https://doi.org/10.7150/jca.18808

Lin, Y.C., Jhunjhunwala, S., Benner, C., Heinz, S., Welinder, E., Mansson, R., Sigvardsson, M., Hagman, J., Espinoza, C.A., 2011. A global network of transcription factors, involving E2A, EBF1 and Foxo1, that orchestrates the B cell fate 11, 635–643. https://doi.org/10.1038/ni.1891.A

Lis, J.T., 2015. Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers 46, 1311–1320. https://doi.org/10.1038/ng.3142.Analysis

Liu, H., Wang, G., Huang, Y., Zhao, C., Chen, J., Wang, X., 2019. Identification specific miRNA in t(4;14) multiple myeloma based on miRNA-mRNA expressing profile correlation analysis. J. Cell. Biochem. https://doi.org/10.1002/jcb.27537

Lohr, J.G., Stojanov, P., Carter, S.L., Cruz-Gordillo, P., Lawrence, M.S., Auclair, D., Sougnez, C., Knoechel, B., Gould, J., Saksena, G., Cibulskis, K., McKenna, A., Chapman, M.A., Straussman, R., Levy, J., Perkins, L.M., Keats, J.J., Schumacher, S.E., Rosenberg, M., Consortium, T.M.M.R., Anderson, K.C., Richardson, P., Krishnan, A., Lonial, S., Kaufman, J., Siegel, D.S., Vesole, D.H., Roy, V., Rivera, C.E., Rajkumar, S.V., Kumar, S., Fonseca, R., Ahmann, G.J., Bergsagel, P.L., Stewart, A.K., Hofmeister, C.C., Efebera, Y.A., Jagannath, S., Chari, A., Trudel, S., Reece, D., Wolf, J., Martin, T., Zimmerman, T., Rosenbaum, C., Jakubowiak, A.J., Lebovic, D., Vij, R., Stockerl-Goldstein, K., Getz, G., Golub, T.R., 2014. Widespread Genetic Heterogeneity in Multiple Myeloma: Implications for Targeted Therapy. Cancer Cell 25, 91–101.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., Axel, R., 2006. Interchromosomal Interactions and Olfactory Receptor Choice 403–413. https://doi.org/10.1016/j.cell.2006.06.035

Long, M.D., Smiraglia, D.J., Campbell, M.J., 2017. The genomic impact of DNA CpG methylation on gene expression; relationships in prostate cancer. Biomolecules 7, 1–20. https://doi.org/10.3390/biom7010015

Lu, R., Wang, P., Parton, T., Zhou, Y., Chrysovergis, K., Rockowitz, S., Chen, W.Y., Abdel-Wahab, O., Wade, P.A., Zheng, D., Wang, G.G., 2016. Epigenetic Perturbations by Arg882-Mutated DNMT3A Potentiate Aberrant Stem Cell Gene-Expression Program and Acute Leukemia Development. Cancer Cell. https://doi.org/10.1016/j.ccell.2016.05.008

Lu, Y., Wang, Y., Xu, H., Shi, C., Jin, F., Li, W., 2018. Profilin 1 induces drug resistance through Beclin1 complex-mediated autophagy in multiple myeloma. Cancer Sci. https://doi.org/10.1111/cas.13711

Ludwig, M.Z., Manu, Kittler, R., White, K.P., Kreitman, M., 2011. Consequences of eukaryotic enhancer architecture for gene expression dynamics, development, and fitness. PLoS Genet. 7. https://doi.org/10.1371/journal.pgen.1002364

Lun, A.T.L., Smyth, G.K., 2014. De novo detection of differentially bound regions for ChIP-seq data using peaks and windows: Controlling error rates correctly. Nucleic Acids Res. https://doi.org/10.1093/nar/gku351

Lupiáñez, D.G., Kraft, K., Heinrich, V., Krawitz, P., Brancati, F., Klopocki, E., Horn, D., Kayserili, H., Opitz, J.M., Laxova, R., Santos-Simarro, F., Gilbert-Dussardier, B., Wittler, L., Borschiwer, M., Haas, S.A., Osterwalder, M., Franke, M., Timmermann, B., Hecht, J., Spielmann, M., Visel, A., Mundlos, S., 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. Cell 161, 1012–1025. https://doi.org/10.1016/j.cell.2015.04.004

Maechler, M., Struyf, A., Hubert, M., Hornik, K., Studer, M., Roudier, P., 2016. cluster: Cluster Analysis Basics and Extensions, In: R Package Version 2.0.4. https://doi.org/ISBN 0-387-95457-0

Maes, K., De Smedt, E., Kassambara, A., Hose, D., Seckinger, A., Van Valckenborgh, E., Menu, E., Klein, B., Vanderkerken, K., Moreaux, J., De Bruyne, E., 2015. *In vivo* treatment with epigenetic modulating agents induces transcriptional alterations associated with prognosis and immunomodulation in multiple myeloma. Oncotarget. https://doi.org/10.18632/oncotarget.3207

Magnani, L., Stoeck, A., Zhang, X., Lanczky, A., Mirabella, A.C., Wang, T.-L., Gyorffy, B., Lupien, M., 2013. Genome-wide reprogramming of the chromatin landscape underlies endocrine therapy resistance in breast cancer. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.1219992110

Mahtouk, K., Cremer, F.W., Rème, T., Jourdan, M., Baudard, M., Moreaux, J., Requirand, G., Fiol, G., De Vos, J., Moos, M., Quittet, P., Goldschmidt, H., Rossi, J.F., Hose, D., Klein, B., 2006. Heparan sulphate proteoglycans are essential for the myeloma cell growth activity of EGF-family ligands in multiple myeloma. Oncogene. https://doi.org/10.1038/sj.onc.1209699

Mahtouk, K., Moreaux, J., Hose, D., Rème, T., Meißner, T., Jourdan, M., Rossi, J.F., Pals, S.T., Goldschmidt, H., Klein, B., 2010. Growth factors in multiple myeloma: A comprehensive analysis of their expression in tumor cells and bone marrow environment using Affymetrix microarrays. BMC Cancer. https://doi.org/10.1186/1471-2407-10-198

Manier, S., Salem, K.Z., Park, J., Landau, D.A., Getz, G., Ghobrial, I.M., 2017. Genomic complexity of multiple myeloma and its clinical implications. Nat. Rev. Clin. Oncol. 14, 100–113. https://doi.org/10.1038/nrclinonc.2016.122

Manolio, T.A., Brooks, L.D., Collins, F.S., 2008. A HapMap harvest of insights into the genetics of common disease. J. Clin. Invest. 118, 1590–1605. https://doi.org/10.1172/JCI34772

Mansour, A., Wakkach, A., Blin-Wakkach, C., 2017. Emerging roles of osteoclasts in the modulation of bone microenvironment and immune suppression in multiple myeloma. Front. Immunol. https://doi.org/10.3389/fimmu.2017.00954

Mantovani, F., Collavin, L., Del Sal, G., 2019. Mutant p53 as a guardian of the cancer cell. Cell Death Differ. https://doi.org/10.1038/s41418-018-0246-9

Marcel, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal. https://doi.org/10.14806/ej.17.1.200

Marchica, V., Toscani, D., Corcione, A., Bolzoni, M., Storti, P., Vescovini, R., Ferretti, E., Dalla Palma, B., Vicario, E., Accardi, F., Mancini, C., Martella, E., Ribatti, D., Vacca, A., Pistoia, V., Giuliani, N., 2019. Bone Marrow CX3CL1/Fractalkine is a New Player of the Pro-Angiogenic Microenvironment in Multiple Myeloma Patients. Cancers (Basel). 11, 321. https://doi.org/10.3390/cancers11030321

Markaki, Y., Gunkel, M., Schermelleh, L., 2011. Functional Nuclear Organization of Transcription and DNA Replication : A Topographical Marriage between Chromatin Domains and the Interchromatin Compartment Functional Nuclear Organization of Transcription and DNA Replication A Topographical Marriage betw LXXV. https://doi.org/10.1101/sqb.2010.75.042

Market, E., Papavasiliou, F.N., 2003. V(D)J Recombination and the Evolution of the Adaptive Immune System. PLoS Biol. https://doi.org/10.1371/journal.pbio.0000016

Martinez-Garcia, E., Popovic, R., Min, D.J., Sweet, S.M.M., Thomas, P.M., Zamdborg, L., Heffner, A., Will, C., Lamy, L., Staudt, L.M., Levens, D.L., Kelleher, N.L., Licht, J.D., 2011. The MMSET histone methyl transferase switches global histone methylation and alters gene expression in t(4;14) multiple myeloma cells. Blood. https://doi.org/10.1182/blood-2010-07-298349

Mathelier, A., Fornes, O., Arenillas, D.J., Chen, C.Y., Denay, G., Lee, J., Shi, W., Shyr, C., Tan, G., Worsley-Hunt, R., Zhang, A.W., Parcy, F., Lenhard, B., Sandelin, A., Wasserman, W.W., 2016. JASPAR 2016: A major expansion and update of the open-access database of transcription factor binding profiles. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv1176

Matthews, A.J., Zheng, S., DiMenna, L.J., Chaudhuri, J., 2014. Regulation of immunoglobulin class-switch recombination: Choreography of noncoding transcription, targeted DNA deamination, and long-range DNA repair, Advances in Immunology. https://doi.org/10.1016/B978-0-12-800267-4.00001-8

Mattioli, M., Agnelli, L., Fabris, S., Baldini, L., Morabito, F., Bicciato, S., Verdelli, D., Intini, D., Nobili, L., Cro, L., Pruneri, G., Callea, V., Stelitano, C., Maiolo, A.T., Lombardi, L., Neri, A., 2005. Gene expression profiling of plasma cell dyscrasias reveals molecular patterns associated with distinct IGH translocations in multiple myeloma. Oncogene. https://doi.org/10.1038/sj.onc.1208447

Mckeown, M.R., Corces, M.R., Eaton, M.L., Fiore, C., Lee, E., Lopez, J.T., Chen, M.W., Smith, D., Chan, S.M., Koenig, J.L., Austgen, K., Guenther, M.G., Orlando, D.A., Lovén, J., Fritz, C.C., Majeti, R., Christian, C., 2017. Superenhancer Analysis Defines Novel Epigenomic Subtypes of Non-APL AML, Including an RARα Dependency Targetable by SY-1425, a Potent and Selective RARα Agonist. Cancer Discov. 7, 1136–1153. https://doi.org/10.1158/2159-8290.CD-17-0399

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., Bejerano, G., 2010. GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. https://doi.org/10.1038/nbt.1630

Meunier, D., Patra, K., Smits, R., Hägebarth, A., Lüttges, A., Jaussi, R., Wieduwilt, M.J., Quintanilla-Fend, L., Himmelbauer, H., Fodde, R., Fundele, R.H., 2011. Expression analysis

of proline rich 15 (Prr15) in mouse and human gastrointestinal tumors. Mol. Carcinog. https://doi.org/10.1002/mc.20692

Meyering, J., n.d. Grep.

Mikhaylichenko, O., Bondarenko, V., Harnett, D., Schor, I.E., Males, M., Viales, R.R., Furlong, E.E.M., 2018. The degree of enhancer or promoter activity is reflected by the levels and directionality of eRNA transcription. Genes Dev. https://doi.org/10.1101/gad.308619.117

Mirabella, F., Murison, A., Aronson, L.I., Wardell, C.P., Thompson, A.J., Hanrahan, S.J., Fok, J.H.L., Pawlyn, C., Kaiser, M.F., Walker, B.A., Davies, F.E., Morgan, G.J., 2014. A novel functional role for MMSET in RNA processing based on the link between the REIIBP isoform and its interaction with the SMN complex. PLoS One. https://doi.org/10.1371/journal.pone.0099493

Misiewicz-Krzeminska, I., Sarasquete, M.E., Vicente-Dueñas, C., Krzeminski, P., Wiktorska, K., Corchete, L.A., Quwaider, D., Rojas, E.A., Corral, R., Martín, A.A., Escalante, F., Bárez, A., García, J.L., Sánchez-García, I., García-Sanz, R., San Miguel, J.F., Gutiérrez, N.C., 2016. Post-transcriptional modifications contribute to the upregulation of cyclin D2 in multiple myeloma. Clin. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-14-2796

Mittrücker, H.W., Matsuyama, T., Grossman, A., Kündig, T.M., Potter, J., Shahinian, A., Wakeham, A., Patterson, B., Ohashi, P.S., Mak, T.W., 2017. Requirement for the transcription factor LSIRF/IRF4 for mature B and T lymphocyte function. J. Immunol. https://doi.org/10.1126/science.275.5299.540

Molist, R., Remvikos, Y., Dutrillaux, B., Muleris, M., 2004. Characterization of a new cytogenetic subtype of ductal breast carcinomas. Oncogene. https://doi.org/10.1038/sj.onc.1207799

Mrozik, K.M., Cheong, C.M., Hewett, D., Chow, A.W.S., Blaschuk, O.W., Zannettino, A.C.W., Vandyke, K., 2015. Therapeutic targeting of N-cadherin is an effective treatment for multiple myeloma. Br. J. Haematol. https://doi.org/10.1111/bjh.13596

Muerdter, F., Boryń, Ł.M., Arnold, C.D., 2015. STARR-seq - Principles and applications. Genomics. https://doi.org/10.1016/j.ygeno.2015.06.001

Muerdter, F., Boryn, Ł.M., Woodfin, A.R., Neumayr, C., Rath, M., Zabidi, M.A., Pagani, M., Haberle, V., Kazmar, T., Catarino, R.R., Schernhuber, K., Arnold, C.D., Stark, A., 2018. Resolving systematic errors in widely used enhancer activity assays in human cells. Nat. Methods. https://doi.org/10.1038/nmeth.4534

Munshi, N.C., Hideshima, T., Carrasco, D., Shammas, M., Auclair, D., Davies, F., Mitsiades, N., Mitsiades, C., Kim, R.S., Li, C., Rajkumar, S.V., Fonseca, R., Bergsagel, L., Chauhan, D., Anderson, K.C., 2004. Identification of genes modulated in multiple myeloma using genetically identical twin samples. Blood. https://doi.org/10.1182/blood-2003-02-0402

Nagy, M., Chapuis, B., Matthes, T., 2002. Expression of transcription factors PU.1, Spi-B, Blimp-1, BSAP and oct-2 in normal human plasma cells and in multiple myeloma cells. Br. J. Haematol. https://doi.org/10.1046/j.1365-2141.2002.03271.x

Nahar, R., Ramezani-Rad, P., Mossner, M., Duy, C., Cerchietti, L., Geng, H., Dovat, S., Jumaa, H., Ye, B.H., Melnick, A., Müschen, M., 2011. Pre-B cell receptor-mediated activation of BCL6 induces pre-B cell quiescence through transcriptional repression of MYC. Blood. https://doi.org/10.1182/blood-2011-01-331181

Nara, M., Teshima, K., Watanabe, A., Ito, M., Iwamoto, K., Kitabayashi, A., Kume, M., Hatano, Y., Takahashi, N., Iida, S., Sawada, K., Tagawa, H., 2013. Bortezomib Reduces the Tumorigenicity of Multiple Myeloma via Downregulation of Upregulated Targets in Clonogenic Side Population Cells. PLoS One. https://doi.org/10.1371/journal.pone.0056954

National Center for Biotechnology Information, 2019. Gene symbols [WWW Document]. Hum. gene Symb. URL ftp://ftp.ncbi.nlm.nih.gov/gene/DATA/GENE_INFO/Mammalia/Homo_sapiens.gene_info.gz

Nau, M.M., Brooks, B.J., Battey, J., Sausville, E., Gazdar, A.F., Kirsch, I.R., McBride, O.W., Bertness, V., Hollis, G.F., Minna, J.D., 1985. L-myc, a new myc-related gene amplified and expressed in human small cell lung cancer. Nature. https://doi.org/10.1038/318069a0

Neph, S., Vierstra, J., Stergachis, A.B., Reynolds, A.P., Haugen, E., Vernot, B., Thurman, R.E., John, S., Sandstrom, R., Johnson, A.K., Maurano, M.T., Humbert, R., Rynes, E., Wang, H., Vong, S., Lee, K., Bates, D., Diegel, M., Roach, V., Dunn, D., Neri, J., Schafer, A., Hansen, R.S., Kutyavin, T., Giste, E., Weaver, M., Canfield, T., Sabo, P., Zhang, M., Balasundaram, G., Byron, R., MacCoss, M.J., Akey, J.M., Bender, M.A., Groudine, M., Kaul, R., Stamatoyannopoulos, J.A., 2012. An expansive human regulatory lexicon encoded in transcription factor footprints. Nature. https://doi.org/10.1038/nature11212

Neri, P., Ren, L., Azab, A.K., Brentnall, M., Gratton, K., Klimowicz, A.C., Lin, C., Duggan, P., Tassone, P., Mansoor, A., Stewart, D.A., Boise, L.H., Ghobrial, I.M., Bahlis, N.J., 2011.

Integrin β7-mediated regulation of multiple myeloma cell adhesion, migration, and invasion. Blood. https://doi.org/10.1182/blood-2010-06-292243

Nicolas, D., Phillips, N.E., Naef, F., 2017. What shapes eukaryotic transcriptional bursting? Mol. Biosyst. 13, 1280–1290. https://doi.org/10.1039/c7mb00154a

Nizovtseva, E. V, Todolli, S., Olson, W.K., Studitsky, V.M., 2017. Towards quantitative analysis of gene regulation by enhancers 9, 1–13.

Nora, E.P., Lajoie, B.R., Schulz, E.G., Giorgetti, L., Okamoto, I., Servant, N., Piolot, T., van Berkum, N.L., Meisig, J., Sedat, J., Gribnau, J., Barillot, E., Blüthgen, N., Dekker, J., Heard, E., 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. Nature 485, 381–385. https://doi.org/10.1038/nature11049

Okonechnikov, K., Erkek, S., Korbel, J.O., Pfister, S.M., Chavez, L., 2019. InTAD: Chromosome conformation guided analysis of enhancer target genes. BMC Bioinformatics. https://doi.org/10.1186/s12859-019-2655-2

Olsen, O.E., Sankar, M., Elsaadi, S., Hella, H., Buene, G., Darvekar, S.R., Misund, K., Katagiri, T., Knaus, P., Holien, T., 2018. BMPR2 inhibits activin and BMP signaling via wild-type ALK2. J. Cell Sci. https://doi.org/10.1242/jcs213512

Olsen, O.E., Wader, K.F., Misund, K., Våtsveen, T.K., Rø, T.B., Mylin, A.K., Turesson, I., Størdal, B.F., Moen, S.H., Standal, T., Waage, A., Sundan, A., Holien, T., 2014. Bone morphogenetic protein-9 suppresses growth of myeloma cells by signaling through ALK2 but is inhibited by endoglin. Blood Cancer J. https://doi.org/10.1038/bcj.2014.16

Osborne, C.S., Chakalova, L., Brown, K.E., Carter, D., Horton, A., Debrand, E., Goyenechea, B., Mitchell, J.A., Lopes, S., Reik, W., Fraser, P., 2004. Active genes dynamically colocalize to shared sites of ongoing transcription 36, 1065–1071. https://doi.org/10.1038/ng1423

Oshima, T., Abe, M., Asano, J., Hara, T., Kitazoe, K., Sekimoto, E., Tanaka, Y., Shibata, H., Hashimoto, T., Ozaki, S., Kido, S., Inoue, D., Matsumoto, T., 2005. Myeloma cells suppress bone formation by secreting a soluble Wnt inhibitor, sFRP-2. Blood. https://doi.org/10.1182/blood-2004-12-4940

Pagenkopf, A., Dhakal, B., Hope, C.L., Papadas, A., Johnson, M.G., Nagel, B., Kurudza, E., Partha, S., Ruffolo, B., Leith, C., Miyamoto, S., Hematti, P., Hari, P., Callander, N.S., Asimakopoulos, F., 2017. Versican (VCAN) Proteolysis Predicts T-Cell Infiltration in Myeloma Bone Marrow Post- Autologous Stem Cell Transplant (ASCT). Blood 130, 1756

LP – 1756.

Pages, H., Carlson, M., Falcon, S., Li, N., 2010. AnnotationDbi: Annotation Database Interface. R Packag. version 1.4.

Paiva, B., Puig, N., Cedena, M.T., De Jong, B.G., Ruiz, Y., Rapado, I., Martinez-Lopez, J., Cordon, L., Alignani, D., Delgado, J.A., Van Zelm, M.C., Van Dongen, J.J.M., Pascual, M., Agirre, X., Prosper, F., Martín-Subero, J.I., Vidriales, M.B., Gutierrez, N.C., Hernandez, M.T., Oriol, A., Echeveste, M.A., Gonzalez, Y., Johnson, S.K., Epstein, J., Barlogie, B., Morgan, G.J., Orfao, A., Blade, J., Mateos, M. V., Lahuerta, J.J., San-Miguel, J.F., 2017. Differentiation stage of myeloma plasma cells: Biological and clinical significance. Leukemia. https://doi.org/10.1038/leu.2016.211

Pajtler, K.W., Wei, Y., Okonechnikov, K., Silva, P.B.G., Vouri, M., Zhang, L., Brabetz, S., Sieber, L., Gulley, M., Mauermann, M., Wedig, T., Mack, N., Imamura Kawasawa, Y., Sharma, T., Zuckermann, M., Andreiuolo, F., Holland, E., Maass, K., Körkel-Qu, H., Liu, H.K., Sahm, F., Capper, D., Bunt, J., Richards, L.J., Jones, D.T.W., Korshunov, A., Chavez, L., Lichter, P., Hoshino, M., Pfister, S.M., Kool, M., Li, W., Kawauchi, D., 2019. YAP1 subgroup supratentorial ependymoma requires TEAD and nuclear factor I-mediated transcriptional programmes for tumorigenesis. Nat. Commun. https://doi.org/10.1038/s41467-019-11884-5

Palmer, M., Belch, A., Hanson, J., Brox, L., 1988. Dose intensity analysis of melphalan and prednisone in multiple myeloma. J. Natl. Cancer Inst. https://doi.org/10.1093/jnci/80.6.414

Papantonis, A., Cook, P.R., 2013. Transcription Factories : Genome Organization and Gene Regulation. https://doi.org/10.1021/cr300513p

Parada, L.A., Roix, J.J., Misteli, T., 2003. An uncertainty principle in chromosome positioning. Trends Cell Biol. 13, 393–396. https://doi.org/10.1016/S0962-8924(03)00149-1

Park, B.M., Kim, E.J., Nam, H.J., Zhang, D., Bae, C.H., Kang, M., Kim, H., Lee, W., Bogen, B., Lim, S.K., 2017. Cyclized Oligopeptide Targeting LRP5/6-DKK1 Interaction Reduces the Growth of Tumor Burden in a Multiple Myeloma Mouse Model. Yonsei Med. J. 58, 505–513. https://doi.org/10.3349/ymj.2017.58.3.505

Patro, R., Duggal, G., Love, M.I., Irizarry, R.A., Kingsford, C., 2017. Salmon provides fast and bias-aware quantification of transcript expression. Nat. Methods.

https://doi.org/10.1038/nmeth.4197

Paulus, A., Chitta, K., Akhtar, S., Personett, D., Miller, K.C., Thompson, K.J., Carr, J., Kumar, S., Roy, V., Ansell, S.M., Mikhael, J.R., Dispenzieri, A., Reeder, C.B., Rivera, C.E., Foran, J., Chanan-Khan, A., 2014. AT-101 downregulates BCL2 and MCL1 and potentiates the cytotoxic effects of lenalidomide and dexamethasone in preclinical models of multiple myeloma and Waldenström macroglobulinaemia. Br. J. Haematol. https://doi.org/10.1111/bjh.12633

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., Reinders, M., Wessels, L., van Steensel, B., 2010. Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Mol. Cell 38, 603–613. https://doi.org/10.1016/j.molcel.2010.03.016

Perrot, A., Lauwers-Cances, V., Tournay, E., Hulin, C., Chretien, M.-L., Royer, B., Dib, M., Decaux, O., Jaccard, A., Belhadj, K., Brechignac, S., Fontan, J., Voillat, L., Demarquette, H., Collet, P., Rodon, P., Sohn, C., Lifermann, F., Orsini-Piocelle, F., Richez, V., Mohty, M., Macro, M., Minvielle, S., Moreau, P., Leleu, X., Facon, T., Attal, M., Avet-Loiseau, H., Corre, J., 2019. Development and Validation of a Cytogenetic Prognostic Index Predicting Survival in Multiple Myeloma. J. Clin. Oncol. https://doi.org/10.1200/JCO.18.00776

Pertea, M., Pertea, G.M., Antonescu, C.M., Chang, T.C., Mendell, J.T., Salzberg, S.L., 2015. StringTie enables improved reconstruction of a transcriptome from RNA-seq reads. Nat. Biotechnol. https://doi.org/10.1038/nbt.3122

Perumal, D., Leshchenko, V. V., Kuo, P.Y., Jiang, Z., Divakar, S.K.A., Jay Cho, H., Chari, A., Brody, J., Reddy, M.V.R., Zhang, W., Reddy, E.P., Jagannath, S., Parekh, S., 2016. Dual targeting of CDK4 and ARK5 using a novel kinase inhibitor ON123300 exerts potent anticancer activity against multiple myeloma. Cancer Res. https://doi.org/10.1158/0008-5472.CAN-15-2934

Phillips, J.E., Corces, V.G., 2009. CTCF: Master Weaver of the Genome. Cell 137, 1194–1211. https://doi.org/10.1016/j.cell.2009.06.001

Piñero, J., Ramírez-Anguita, J.M., Saüch-Pitarch, J., Ronzano, F., Centeno, E., Sanz, F., Furlong, L.I., 2020. The DisGeNET knowledge platform for disease genomics: 2019 update. Nucleic Acids Res. https://doi.org/10.1093/nar/gkz1021

Popovic, R., Martinez-Garcia, E., Giannopoulou, E.G., Zhang, Quanwei, Zhang, Qingyang,

Ezponda, T., Shah, M.Y., Zheng, Y., Will, C.M., Small, E.C., Hua, Y., Bulic, M., Jiang, Y., Carrara, M., Calogero, R.A., Kath, W.L., Kelleher, N.L., Wang, J.P., Elemento, O., Licht, J.D., 2014. Histone Methyltransferase MMSET/NSD2 Alters EZH2 Binding and Reprograms the Myeloma Epigenome through Global and Focal Changes in H3K36 and H3K27 Methylation. PLoS Genet. https://doi.org/10.1371/journal.pgen.1004566

Prideaux, S.M., Conway O'Brien, E., Chevassut, T.J., 2014. The genetic architecture of multiple myeloma. Adv. Hematol. 2014. https://doi.org/10.1155/2014/864058

Qi, L.S., Larson, M.H., Gilbert, L.A., Doudna, J.A., Weissman, J.S., Arkin, A.P., Lim, W.A., 2013. Repurposing CRISPR as an RNA-γuided platform for sequence-specific control of gene expression. Cell. https://doi.org/10.1016/j.cell.2013.02.022

Qiang, Y.W., Barlogie, B., Rudikoff, S., Shaughnessy, J.D., 2008. Dkk1-induced inhibition of Wnt signaling in osteoblast differentiation is an underlying mechanism of bone loss in multiple myeloma. Bone. https://doi.org/10.1016/j.bone.2007.12.006

Qu, Y., Siggens, L., Cordeddu, L., Gaidzik, V.I., Karlsson, K., Bullinger, L., Döhner, K., Ekwall, K., Lehmann, S., Lennartsson, A., 2017. Cancer-specific changes in DNA methylation reveal aberrant silencing and activation of enhancers in leukemia. Blood. https://doi.org/10.1182/blood-2016-07-726877

Quinlan, A.R., Hall, I.M., 2010. BEDTools: A flexible suite of utilities for comparing genomic features. Bioinformatics. https://doi.org/10.1093/bioinformatics/btq033

Rada-Iglesias, A., Grosveld, F.G., Papantonis, A., 2018. Forces driving the three-dimensional folding of eukaryotic genomes. Mol. Syst. Biol. https://doi.org/10.15252/msb.20188214

Radomska, H.S., Shen, C.P., Kadesch, T., Eckhardt, L.A., 1994. Constitutively expressed Oct-2 prevents immunoglobulin gene silencing in myeloma × T cell hybrids. Immunity. https://doi.org/10.1016/1074-7613(94)90034-5

Rao, L., Veirman, K. De, Giannico, D., Saltarella, I., Desantis, V., Frassanito, M.A., Solimando, A.G., Ribatti, D., Prete, M., Harstrick, A., Fiedler, U., Raeve, H. De, Racanelli, V., Vanderkerken, K., Vacca, A., 2018. Targeting angiogenesis in multiple myeloma by the VEGF and HGF blocking DARPin® protein MP0250: A preclinical study. Oncotarget. https://doi.org/10.18632/oncotarget.24351

Rao, S.S.P., Huntley, M.H., Durand, N.C., Stamenova, E.K., Bochkov, I.D., Robinson, J.T., Sanborn, A.L., Machol, I., Omer, A.D., Lander, E.S., Aiden, E.L., 2014. A 3D map of the

human genome at kilobase resolution reveals principles of chromatin looping. Cell. https://doi.org/10.1016/j.cell.2014.11.021

Rashid, N.U., Giresi, P.G., Ibrahim, J.G., Sun, W., Lieb, J.D., 2011. ZINBA integrates local covariates with DNA-seq data to identify broad and narrow regions of enrichment, even within amplified genomic regions. Genome Biol. https://doi.org/10.1186/gb-2011-12-7-r67

Rasmussen, K.D., Jia, G., Johansen, J. V., Pedersen, M.T., Rapin, N., Bagger, F.O., Porse, B.T., Bernard, O.A., Christensen, J., Helin, K., 2015. Loss of TET2 in hematopoietic cells leads to DNA hypermethylation of active enhancers and induction of leukemogenesis. Genes Dev. https://doi.org/10.1101/gad.260174.115

Ravi, P., Kumar, S.K., Cerhan, J.R., Maurer, M.J., Dingli, D., Ansell, S.M., Rajkumar, S.V., 2018. Defining cure in multiple myeloma: A comparative study of outcomes of young individuals with myeloma and curable hematologic malignancies. Blood Cancer J. https://doi.org/10.1038/s41408-018-0065-8

Reagan, M.R., Liaw, L., Rosen, C.J., Ghobrial, I.M., 2015. Dynamic interplay between bone and multiple myeloma: Emerging roles of the osteoblast. Bone. https://doi.org/10.1016/j.bone.2015.02.021

Rendeiro, A.F., Schmidl, C., Strefford, J.C., Walewska, R., Davis, Z., Farlik, M., Oscier, D., Bock, C., 2016. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. Nat. Commun. 7. https://doi.org/10.1038/ncomms11938

Ribatti, D., Vacca, A., 2018. New Insights in Anti-Angiogenesis in Multiple Myeloma. Int. J. Mol. Sci. https://doi.org/10.3390/ijms19072031

Rickels, R., Herz, H.M., Sze, C.C., Cao, K., Morgan, M.A., Collings, C.K., Gause, M., Takahashi, Y.H., Wang, L., Rendleman, E.J., Marshall, S.A., Krueger, A., Bartom, E.T., Piunti, A., Smith, E.R., Abshiru, N.A., Kelleher, N.L., Dorsett, D., Shilatifard, A., 2017. Histone H3K4 monomethylation catalyzed by Trr and mammalian COMPASS-like proteins at enhancers is dispensable for development and viability. Nat. Genet. https://doi.org/10.1038/ng.3965

Ritchie, M.E., Phipson, B., Wu, D., Hu, Y., Law, C.W., Shi, W., Smyth, G.K., 2015. limma powers differential expression analyses for RNA-sequencing and microarray studies. Nucleic Acids

Res. https://doi.org/10.1093/nar/gkv007

Ronchetti, D., Agnelli, L., Pietrelli, A., Todoerti, K., Manzoni, M., Taiana, E., Neri, A., 2018. A compendium of long non-coding RNAs transcriptional fingerprint in multiple myeloma. Sci. Rep. https://doi.org/10.1038/s41598-018-24701-8

Ronchetti, D., Todoerti, K., Tuana, G., Agnelli, L., Mosca, L., Lionetti, M., Fabris, S., Colapietro, P., Miozzo, M., Ferrarini, M., Tassone, P., Neri, A., 2012. The expression pattern of small nucleolar and small Cajal body-specific RNAs characterizes distinct molecular subtypes of multiple myeloma. Blood Cancer J. https://doi.org/10.1038/bcj.2012.41

Roth, D.B., 2000. From lymphocytes to sharks: V(D)J recombinase moves to the germline. Genome Biol. 1, REVIEWS1014. https://doi.org/10.1186/gb-2000-1-2-reviews1014

Roy, A.L., Sen, R., Roeder, R.G., 2012. Enhancer-promoter communication and transcriptional regulation of Igh 32, 532–539. https://doi.org/10.1016/j.it.2011.06.012.Enhancer-promoter

Sainsbury, S., Bernecky, C., Cramer, P., 2015. Structural basis of transcription initiation by RNA polymerase II. Nat. Rev. Mol. Cell Biol. 16, 129–143. https://doi.org/10.1038/nrm3952

Salameh, A., Fan, X., Choi, B.-K., Zhang, S., Zhang, N., An, Z., 2017. HER3 and *LINC00052* interplay promotes tumor growth in breast cancer. Oncotarget. https://doi.org/10.18632/oncotarget.14313

Sanborn, A.L., Rao, S.S.P., Huang, S.-C., Durand, N.C., Huntley, M.H., Jewett, A.I., Bochkov, I.D., Chinnappan, D., Cutkosky, A., Li, J., Geeting, K.P., Gnirke, A., Melnikov, A., McKenna, D., Stamenova, E.K., Lander, E.S., Aiden, E.L., 2015. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. Proc. Natl. Acad. Sci. 112, 201518552. https://doi.org/10.1073/pnas.1518552112

Sanyal, A., Lajoie, B.R., Jain, G., Dekker, J., 2012. The long-range interaction landscape of gene promoters. Nature 489, 109–113. https://doi.org/10.1038/nature11279

Sarasquete, M.E., Martínez-López, J., Chillón, M.C., Alcoceba, M., Corchete, L.A., Paiva, B., Puig, N., Sebastián, E., Jiménez, C., Mateos, M.V., Oriol, A., Rosiñol, L., Palomera, L., Teruel, A.I., González, Y., Lahuerta, J.J., Bladé, J., Gutiérrez, N.C., Fernández-Redondo, E., González, M., San Miguel, J.F., García-Sanz, R., 2013. Evaluating gene expression profiling by quantitative polymerase chain reaction to develop a clinically feasible test for outcome prediction in multiple myeloma. Br. J. Haematol.

https://doi.org/10.1111/bjh.12519

Satterthwaite, A.B., 2018. Bruton's tyrosine kinase, a component of B cell signaling pathways, has multiple roles in the pathogenesis of lupus. Front. Immunol. https://doi.org/10.3389/fimmu.2017.01986

Schatz, D.G., Swanson, P.C., 2011. V(D)J Recombination: Mechanisms of Initiation. Annu. Rev. Genet. 45, 167–202. https://doi.org/10.1146/annurev-genet-110410-132552

Schaukowitch, K., Joo, J.Y., Liu, X., Watts, J.K., Martinez, C., Kim, T.K., 2014. Enhancer RNA facilitates NELF release from immediate early genes. Mol. Cell. https://doi.org/10.1016/j.molcel.2014.08.023

Schwab, M., Alitalo, K., Klempnauer, K.H., Varmus, H.E., Bishop, J.M., Gilbert, F., Brodeur, G., Goldstein, M., Trent, J., 1983. Amplified DNA with limited homology to myc cellular oncogene is shared by human neuroblastoma cell lines and a neuroblastoma tumour. Nature. https://doi.org/10.1038/305245a0

Scott, M.S., Ono, M., 2011. From snoRNA to miRNA: Dual function regulatory non-coding RNAs. Biochimie. https://doi.org/10.1016/j.biochi.2011.05.026

Sengupta, K., Camps, J., Mathews, P., Barenboim-Stapleton, L., Nguyen, Q.T., Difilippantonio, M.J., Ried, T., 2008. Position of human chromosomes is conserved in mouse nuclei indicating a species-independent mechanism for maintaining genome organization. Chromosoma 117, 499–509. https://doi.org/10.1007/s00412-008-0171-7

Sengupta, S., George, R.E., Chen, J.E., Glover, G.H., 2017. Super-Enhancer-Driven Transcriptional Dependencies in Cancer. Trends in cancer 3, 269–281. https://doi.org/10.1016/j.trecan.2017.03.006

Sevcikova, S., Paszekova, H., Besse, L., Sedlarikova, L., Kubaczkova, V., Almasi, M., Pour, L., Hajek, R., 2015. Extramedullary relapse of multiple myeloma defined as the highest risk group based on deregulated gene expression data. Biomed. Pap. https://doi.org/10.5507/bp.2015.014

Shaffer, A.L., Emre, N.C.T., Lamy, L., Ngo, V.N., Wright, G., Xiao, W., Powell, J., Dave, S., Yu, X., Zhao, H., Zeng, Y., Chen, B., Epstein, J., Staudt, L.M., 2008. IRF4 addiction in multiple myeloma. Nature. https://doi.org/10.1038/nature07064

Shah, V., Sherborne, A.L., Walker, B.A., Johnson, D.C., Boyle, E.M., Ellis, S., Begum, D.B., Proszek, P.Z., Jones, J.R., Pawlyn, C., Savola, S., Jenner, M.W., Drayson, M.T., Owen, R.G.,

Houlston, R.S., Cairns, D.A., Gregory, W.M., Cook, G., Davies, F.E., Jackson, G.H., Morgan, G.J., Kaiser, M.F., 2018. Prediction of outcome in newly diagnosed myeloma: A meta-analysis of the molecular profiles of 1905 trial patients. Leukemia. https://doi.org/10.1038/leu.2017.179

Shan, Y., Ying, R., Jia, Z., Kong, W., Wu, Y., Zheng, S., Jin, H., 2017. LINC00052 Promotes Gastric Cancer Cell Proliferation and Metastasis via Activating the Wnt/β-Catenin Signaling Pathway. Oncol. Res. Featur. Preclin. Clin. Cancer Ther. https://doi.org/10.3727/096504017x14897896412027

Shaughnessy, J.D., Zhan, F., Burington, B.E., Huang, Y., Colla, S., Hanamura, I., Stewart, J.P., Kordsmeier, B., Randolph, C., Williams, D.R., Xiao, Y., Xu, H., Epstein, J., Anaissie, E., Krishna, S.G., Cottler-Fox, M., Hollmig, K., Mohiuddin, A., Pineda-Roman, M., Tricot, G., Van Rhee, F., Sawyer, J., Alsayed, Y., Walker, R., Zangari, M., Crowley, J., Barlogie, B., 2007. A validated gene expression model of high-risk multiple myeloma is defined by deregulated expression of genes mapping to chromosome 1. Blood. https://doi.org/10.1182/blood-2006-07-038430

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., Ren, B., 2012. A map of the cis-regulatory sequences in the mouse genome. Nature 488, 116–120. https://doi.org/10.1038/nature11243

Sherafatian, M., 2018. Tree-based machine learning algorithms identified minimal set of miRNA biomarkers for breast cancer diagnosis and molecular subtyping. Gene. https://doi.org/10.1016/j.gene.2018.07.057

Shi, C., Sun, L., Song, Y., 2019. FEZF1-AS1: a novel vital oncogenic lncRNA in multiple human malignancies. Biosci. Rep. 39, BSR20191202. https://doi.org/10.1042/BSR20191202

Shi, J., Whyte, W.A., Zepeda-Mendoza, C.J., Milazzo, J.P., Shen, C., Roe, J.S., Minder, J.L., Mercan, F., Wang, E., Eckersley-Maslin, M.A., Campbell, A.E., Kawaoka, S., Shareef, S., Zhu, Z., Kendall, J., Muhar, M., Haslinger, C., Yu, M., Roeder, R.G., Wigler, M.H., Blobel, G.A., Zuber, J., Spector, D.L., Young, R.A., Vakoc, C.R., 2013. Role of SWI/SNF in acute leukemia maintenance and enhancer-mediated Myc regulation. Genes Dev. https://doi.org/10.1101/gad.232710.113

Shlyueva, D., Stampfel, G., Stark, A., 2014. Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. 15, 272–86. https://doi.org/10.1038/nrg3682

Siggers, T., Duyzend, M.H., Reddy, J., Khan, S., Bulyk, M.L., 2011. Non-DNA-binding cofactors enhance DNA-binding specificity of a transcriptional regulatory complex. Mol. Syst. Biol. 7, 1–14. https://doi.org/10.1038/msb.2011.89

Smith, D., Mann, D., Yong, K., 2016. Cyclin D type does not influence cell cycle response to DNA damage caused by ionizing radiation in multiple myeloma tumours. Br. J. Haematol. https://doi.org/10.1111/bjh.13982

Smith, E.M., Lajoie, B.R., Jain, G., Dekker, J., 2016. Invariant TAD Boundaries Constrain Cell-Type-Specific Looping Interactions between Promoters and Distal Elements around the CFTR Locus. Am. J. Hum. Genet. 98, 185–201. https://doi.org/10.1016/j.ajhg.2015.12.002

Soares, L.M., He, P.C., Chun, Y., Suh, H., Kim, T.S., Buratowski, S., 2017. Determinants of Histone H3K4 Methylation Patterns. Mol. Cell. https://doi.org/10.1016/j.molcel.2017.10.013

Spitz, F., 2016. Gene regulation at a distance: From remote enhancers to 3D regulatory ensembles. Semin. Cell Dev. Biol. https://doi.org/10.1016/j.semcdb.2016.06.017

Spitz, F., Furlong, E.E.M., 2012. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. 13, 613–626. https://doi.org/10.1038/nrg3207

Starks, R.R., Biswas, A., Jain, A., Tuteja, G., 2019. Combined analysis of dissimilar promoter accessibility and gene expression profiles identifies tissue-specific genes and actively repressed networks. Epigenetics and Chromatin 12, 1–16. https://doi.org/10.1186/s13072-019-0260-2

Stergachis, A.B., Neph, S., Reynolds, A., Humbert, R., Miller, B., Paige, S.L., Vernot, B., Cheng, J.B., Thurman, R.E., Sandstrom, R., Haugen, E., Heimfeld, S., Murry, C.E., Akey, J.M., Stamatoyannopoulos, J.A., 2013. Developmental fate and cellular maturity encoded in human regulatory DNA landscapes. Cell 154, 888–903. https://doi.org/10.1016/j.cell.2013.07.020

Stewart, J.P., Thompson, A., Santra, M., Barlogie, B., Lappin, T.R.J., Shaughnessy, J., 2004. Correlation of TACC3, FGFR3, MMSET and p21 expression with the t(4;14)(p16.3;q32) in multiple myeloma. Br. J. Haematol. https://doi.org/10.1111/j.1365-2141.2004.04996.x

Storti, P., Agnelli, L., dalla Palma, B., Todoerti, K., Marchica, V., Accardi, F., Sammarelli, G., Deluca, F., Toscani, D., Costa, F., Vicario, E., Todaro, G., Martella, E., Neri, A., Giuliani, N., 2019. A retained transcriptomic profile characterizes CD138+ cells in the short time

progression from smoldering to active multiple myeloma. Haematologica. https://doi.org/10.3324/haematol.2018.209999

Stovner, E.B., Sætrom, P., 2019. epic2 efficiently finds diffuse domains in ChIP-seq data. Bioinformatics. https://doi.org/10.1093/bioinformatics/btz232

Strickfaden, H., Zunhammer, A., van Koningsbruggen, S., Köhler, D., Cremer, T., 2010. 4D chromatin dynamics in cycling cells: Theodor Boveri's hypotheses revisited. Nucleus 1, 284–97. https://doi.org/10.4161/nucl.1.3.11969

Strømme, O., Psonka-Antonczyk, K.M., Stokke, B.T., Sundan, A., Arum, C.J., Brede, G., 2019. Myeloma-derived extracellular vesicles mediate HGF/c-Met signaling in osteoblast-like cells. Exp. Cell Res. https://doi.org/10.1016/j.yexcr.2019.07.003

Stubbs, M.C., Burn, T.C., Sparks, R., Maduskuie, T., Diamond, S., Rupar, M., Wen, X., Volgina, A., Zolotarjova, N., Waeltz, P., Favata, M., Jalluri, R., Liu, H., Liu, X.M., Li, J., Collins, R., Falahatpisheh, N., Polam, P., DiMatteo, D., Feldman, P., Dostalik, V., Thekkat, P., Gardiner, C., He, X., Li, Y., Covington, M., Wynn, R., Ruggeri, B., Yeleswaram, S., Xue, C.B., Yao, W., Combs, A.P., Huber, R., Hollis, G., Scherle, P., Liu, P.C.C., 2019. The novel bromodomain and extraterminal domain inhibitor INCB054329 induces vulnerabilities in myeloma cells that inform rational combination strategies. Clin. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-18-0098

Sudbery, I., 2019a. Sudlab github pipeline_apa [WWW Document]. URL https://github.com/sudlab/pipeline_apa

Sudbery, I., 2019b. Sudlab github pipeline_denovo_motifs [WWW Document]. URL https://github.com/sudlab/pipeline_denovo_motifs

Sun, Y., Miao, N., Sun, T., 2019. Detect accessible chromatin using ATAC-sequencing, from principle to applications. Hereditas. https://doi.org/10.1186/s41065-019-0105-9

Taberlay, P.C., Statham, A.L., Kelly, T.K., Clark, S.J., Jones, P.A., 2014. Reconfiguration of nucleosome-depleted regions at distal regulatory elements accompanies DNA methylation of enhancers and insulators in cancer. Genome Res. 24, 1421–1432. https://doi.org/10.1101/gr.163485.113

Tai, Y.T., Podar, K., Mitsiades, N., Lin, B., Mitsiades, C., Gupta, D., Akiyama, M., Catley, L., Hideshima, T., Munshi, N.C., Treon, S.P., Anderson, K.C., 2003. CD40 induces human multiple myeloma cell migration via phosphatidylinositol 3-kinase/AKT/NF-κB signaling.

Blood. https://doi.org/10.1182/blood-2002-09-2813

Takaoka, A., Tamura, T., Taniguchi, T., 2008. Interferon regulatory factor family of transcription factors and regulation of oncogenesis. Cancer Sci. https://doi.org/10.1111/j.1349-7006.2007.00720.x

Tan-wong, S.M., Wijayatilake, H.D., Proudfoot, N.J., 2009. Gene loops function to maintain transcriptional memory through interaction with the nuclear pore complex 2610–2624. https://doi.org/10.1101/gad.1823209.Freely

Tang, H., Liu, P., Yang, L., Xie, Xinhua, Ye, F., Wu, M., Liu, X., Chen, B., Zhang, L., Xie, Xiaoming, 2014. miR-185 Suppresses Tumor Proliferation by Directly Targeting E2F6 and DNMT1 and Indirectly Upregulating BRCA1 in Triple-Negative Breast Cancer. Mol. Cancer Ther. https://doi.org/10.1158/1535-7163.mct-14-0243

Taub, R., Kirsch, I., Morton, C., Lenoir, G., Swan, D., Tronick, S., Aaronson, S., Leder, P., 1982. Translocation of the c-myc gene into the immunoglobulin heavy chain locus in human Burkitt lymphoma and murine plasmacytoma cells. Proc. Natl. Acad. Sci. U. S. A.

Team_BC, Maintainer_BP, 2019. TxDb.Hsapiens.UCSC.hg38.knownGene: Annotation package for TxDb object(s).

Teoh, P.J., Chung, T.-H., Chng, P.Y.Z., Toh, S.H.M., Chng, W.J., 2019. IL6R-STAT3-ADAR1 (P150) interplay promotes oncogenicity in 1q21(amp) multiple myeloma. Haematologica. https://doi.org/10.3324/haematol.2019.221176

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Matthew, T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., Garg, K., Sandstrom, R., Bates, D., Canfield, T.K., Diegel, M., Dunn, D., Ebersol, A.K., Frum, T., Giste, E., Johnson, A.K., Johnson, E.M., Kutyavin, T., Lajoie, B., Lee, K.K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, D., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., Lee, K.K., London, D., Lotakis, D., Neph, S., Neri, F., Nguyen, D., Reynolds, A.P., Roach, V., Safi, A., Sanchez, M.E., Sanyal, A., Shafer, A., Simon, J.M., Song, L., Vong, S., Weaver, M., Zhang, Z., Lenhard, B., Tewari, M., Dorschner, M.O., Hansen, R.S., Navas, P.A., Iyer, V.R., Lieb, J.D., Sunyaev, S.R., 2012. The accessible chromatin landscape of the human genome 489, 75–82. https://doi.org/10.1038/nature11232.The

Tiedemann, R.E., Mao, X., Shi, C.X., Yuan, X.Z., Palmer, S.E., Sebag, M., Marler, R., Chesi, M., Fonseca, R., Bergsagel, P.L., Schimmer, A.D., Stewart, A.K., 2008. Identification of kinetin riboside as a repressor of CCND1 and CCND2 with preclinical antimyeloma activity. J. Clin. Invest. https://doi.org/10.1172/JCI34149

Tooze, R.M., 2013. A replicative self-renewal model for long-lived plasma cells: Questioning irreversible cell cycle exit. Front. Immunol. https://doi.org/10.3389/fimmu.2013.00460

Trimarchi, T., Bilal, E., Ntziachristos, P., Fabbri, G., Dalla-Favera, R., Tsirigos, A., Aifantis, I., 2014. Genome-wide mapping and characterization of notch-regulated long noncoding RNAs in acute leukemia. Cell. https://doi.org/10.1016/j.cell.2014.05.049

Türkmen, S., Binder, A., Gerlach, A., Niehage, S., Theodora Melissari, M., Inandiklioglu, N., Dörken, B., Burmeister, T., 2014. High prevalence of immunoglobulin light chain gene aberrations as revealed by FISH in multiple myeloma and MGUS. Genes Chromosom. Cancer. https://doi.org/10.1002/gcc.22175

UCSC, 2019. Lift Genome Annotations [WWW Document]. URL http://genome.ucsc.edu/cgi-bin/hgLiftOver

Udi, J., Schüler, J., Wider, D., Ihorst, G., Catusse, J., Waldschmidt, J., Schnerch, D., Follo, M., Wäsch, R., Engelhardt, M., 2013. Potent in vitro and in vivo activity of sorafenib in multiple myeloma: Induction of cell death, CD138-downregulation and inhibition of migration through actin depolymerization. Br. J. Haematol. https://doi.org/10.1111/bjh.12226

Ulianov, S. V., Tachibana-Konwalski, K., Razin, S. V., 2017. Single-cell Hi-C bridges microscopy and genome-wide sequencing approaches to study 3D chromatin organization. BioEssays 39, 1–8. https://doi.org/10.1002/bies.201700104

Ullah, T.R., 2019. The role of CXCR4 in multiple myeloma: Cells' journey from bone marrow to beyond. J. Bone Oncol. https://doi.org/10.1016/j.jbo.2019.100253

Vale, A., 2010. Clinical Consequences of Defects of B cell Development. J. Allergy Clin. Immunol. 125, 778–787. https://doi.org/10.1016/j.jaci.2010.02.018.Clinical

Vallet, S., Raje, N., Ishitsuka, K., Hideshima, T., Podar, K., Chhetri, S., Pozzi, S., Breitkreutz, I., Kiziltepe, T., Yasui, H., Ocio, E.M., Shiraishi, N., Jin, J., Okawa, Y., Ikeda, H., Mukherjee, S., Vaghela, N., Cirstea, D., Ladetto, M., Boccadoro, M., Anderson, K.C., 2007. MLN3897, a novel CCR1 inhibitor, impairs osteoclastogenesis and inhibits the interaction of multiple

myeloma cells and osteoclasts. Blood. https://doi.org/10.1182/blood-2007-05-093294

Valton, A., Dekker, J., 2016. ScienceDirect TAD disruption as oncogenic driver. Curr. Opin. Genet. Dev. 36, 34–40. https://doi.org/10.1016/j.gde.2016.03.008

van de Donk, N.W.C.J., Mutis, T., Poddighe, P.J., Lokhorst, H.M., Zweegman, S., 2016. Diagnosis, risk stratification and management of monoclonal gammopathy of undetermined significance and smoldering multiple myeloma. Int. J. Lab. Hematol. https://doi.org/10.1111/ijlh.12504

Van Duijvenboden, K., De Boer, B.A., Capon, N., Ruijter, J.M., Christoffels, V.M., 2015. EMERGE: A flexible modelling framework to predict genomic regulatory elements from genomic signatures. Nucleic Acids Res. https://doi.org/10.1093/nar/gkv1144

Venken, K.J.T., Bellen, H.J., 2014. Chemical mutagens, transposons, and transgenes to interrogate gene function in Drosophila melanogaster. Methods. https://doi.org/10.1016/j.ymeth.2014.02.025

Vietri Rudan, M., Barrington, C., Henderson, S., Ernst, C., Odom, D.T., Tanay, A., Hadjur, S., 2015. Comparative Hi-C Reveals that CTCF Underlies Evolution of Chromosomal Domain Architecture. Cell Rep. 10, 1297–1309. https://doi.org/10.1016/j.celrep.2015.02.004

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J. a, Plajzer-frick, I., Shoukry, M., Wright, C., Chen, F., Afzal, V., Ren, B., Rubin, E.M., Pennacchio, L. a, Division, G., Berkeley, L., 2010. ChIP-seq accurately predicts tissue-specific activity of enhancers 457, 854–858. https://doi.org/10.1038/nature07730.ChIP-seq

Walter, J., Schermelleh, L., Cremer, M., Tashiro, S., Cremer, T., 2003. Chromosome order in HeLa cells changes during mitosis and early G1, but is stably maintained during subsequent interphase stages. J. Cell Biol. 160, 685–697. https://doi.org/10.1083/jcb.200211103

Wang, H., Chen, Y., Han, J., Meng, Q., Xi, Q., Wu, G., Zhang, B., 2016. DCAF4L2 promotes colorectal cancer invasion and metastasis via mediating degradation of NFκb negative regulator PPM1B. Am. J. Transl. Res.

Wang, L., Yao, Z.Q., Moorman, J.P., Xu, Y., Ning, S., 2014. Gene expression profiling identifies IRF4-Associated molecular signatures in hematological malignancies. PLoS One. https://doi.org/10.1371/journal.pone.0106788

Wang, Z., Gerstein, M., Snyder, M., 2009. RNA-Seq: A revolutionary tool for transcriptomics.

Nat. Rev. Genet. https://doi.org/10.1038/nrg2484

Wei Wang and Shimin Hu, 2015. Neoplastic plasma cells mimic mature neutrophils in plasma cell myeloma with t(11;14)(q13;q32). Blood. https://doi.org/10.1182/blood-2014-12-617795

Weirauch, M.T., Yang, A., Albu, M., Cote, A.G., Montenegro-Montero, A., Drewe, P., Najafabadi, H.S., Lambert, S.A., Mann, I., Cook, K., Zheng, H., Goity, A., van Bakel, H., Lozano, J.-C., Galli, M., Lewsey, M.G., Huang, E., Mukherjee, T., Chen, X., Reece-Hoyes, J.S., Govindarajan, S., Shaulsky, G., Walhout, A.J.M., Bouget, F.-Y., Ratsch, G., Larrondo, L.F., Ecker, J.R., Hughes, T.R., 2014. Determination and Inference of Eukaryotic Transcription Factor Sequence Specificity. Cell. https://doi.org/10.1016/j.cell.2014.08.009

Weiss, B.M., Abadie, J., Verma, P., Howard, R.S., Kuehl, W.M., 2009. A monoclonal gammopathy precedes multiple myeloma in most patients. Clin. Trialsand Obs. 113, 5418–5422. https://doi.org/10.1182/blood-2008-12-195008.An

Wendt, K.S., Yoshida, K., Itoh, T., Bando, M., Koch, B., Schirghuber, E., Tsutsumi, S., Nagae, G., Ishihara, K., Mishiro, T., Yahata, K., Imamoto, F., Aburatani, H., Nakao, M., Imamoto, N., Maeshima, K., Shirahige, K., 2008. Cohesin mediates transcriptional insulation by CCCTC-binding factor 451. https://doi.org/10.1038/nature06634

West, A.G., Gaszner, M., Felsenfeld, G., 2002. Insulators: Many functions, many mechanisms. Genes Dev. 16, 271–288. https://doi.org/10.1101/gad.954702

Whyte, W.A., Orlando, D.A., Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., Young, R.A., 2013. Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell. https://doi.org/10.1016/j.cell.2013.03.035

Wiench, M., John, S., Baek, S., Johnson, T.A., Sung, M.H., Escobar, T., Simmons, C.A., Pearce, K.H., Biddie, S.C., Sabo, P.J., Thurman, R.E., Stamatoyannopoulos, J.A., Hager, G.L., 2011. DNA methylation status predicts cell type-specific enhancer activity. EMBO J. https://doi.org/10.1038/emboj.2011.210

Wissink, E.M., Vihervaara, A., Tippens, N.D., Lis, J.T., 2019. Nascent RNA analyses: tracking transcription and its regulation. Nat. Rev. Genet. https://doi.org/10.1038/s41576-019-0159-6

Wu, S., Ou, T., Xing, N., Lu, J., Wan, S., Wang, C., Zhang, X., Yang, F., Huang, Y., Cai, Z., 2019. Whole-genome sequencing identifies ADGRG6 enhancer mutations and FRS2 duplications

as angiogenesis-related drivers in bladder cancer. Nat. Commun. https://doi.org/10.1038/s41467-019-08576-5

Wu, S.P., Pfeiffer, R.M., Ahn, I.E., Mailankody, S., Sonneveld, P., Duin, M. Van, Munshi, N.C., Walker, B.A., Morgan, G., Landgren, O., 2016. Impact of genes highly correlated with MMSET myeloma on the survival of non-MMSET myeloma patients. Clin. Cancer Res. https://doi.org/10.1158/1078-0432.CCR-15-2366

Xie, Z., Chng, W.J., 2014. MMSET: Role and therapeutic opportunities in multiple myeloma. Biomed Res. Int. https://doi.org/10.1155/2014/636514

Xu, Y., Chen, B., George, S.K., Liu, B., 2015. Downregulation of microRNA-152 contributes to high expression of DKK1 in multiple myeloma. RNA Biol. https://doi.org/10.1080/15476286.2015.1094600

Yaccoby, S., 2010. Advances in the understanding of myeloma bone disease and tumour growth. Br. J. Haematol. https://doi.org/10.1111/j.1365-2141.2010.08141.x

Yamazaki, H., Suzuki, M., Otsuki, A., Shimizu, R., Bresnick, E.H., Engel, J.D., Yamamoto, M., 2014. A remote GATA2 hematopoietic enhancer drives leukemogenesis in inv(3)(q21;q26) by activating EVI1 expression. Cancer Cell. https://doi.org/10.1016/j.ccr.2014.02.008

Yan, J., Ojo, D., Kapoor, A., Lin, X., Pinthus, J.H., Aziz, T., Bismar, T.A., Wei, F., Wong, N., De Melo, J., Cutz, J.C., Major, P., Wood, G., Peng, H., Tang, D., 2016. Neural cell adhesion protein CNTN1 promotes the metastatic progression of prostate cancer. Cancer Res. https://doi.org/10.1158/0008-5472.CAN-15-1898

Yáñez-cuna, J.O., Dinh, H.Q., Kvon, E.Z., Shlyueva, D., 2012. Uncovering cis -regulatory sequence requirements for context specific transcription factor binding.

Yao, X., Tan, J., Lim, K.J., Koh, J., Ooi, W.F., Li, Z., Huang, D., Xing, M., Chan, Y.S., Qu, J.Z., Tay, S.T., Wijaya, G., Lam, Y.N., Hong, J.H., Lee-Lim, A.P., Guan, P., Ng, M.S.W., He, C.Z., Lin, J.S., Nandi, T., Qamra, A., Xu, C., Myint, S.S., Davies, J.O.J., Goh, J.Y., Loh, G., Tan, B.C., Rozen, S.G., Yu, Q., Tan, I.B.H., Cheng, C.W.S., Li, S., Chang, K.T.E., Tan, P.H., Silver, D.L., Lezhava, A., Steger, G., Hughes, J.R., Teh, B.T., Tan, P., 2017. VHL deficiency drives enhancer activation of oncogenes in clear cell renal cell carcinoma. Cancer Discov. 7, 1284–1305. https://doi.org/10.1158/2159-8290.CD-17-0375

Yasuda, T., Kometani, K., Takahashi, N., Imai, Y., Aiba, Y., Kurosaki, T., 2011. ERKs induce expression of the transcriptional repressor Blimp-1 and subsequent plasma cell

differentiation. Sci. Signal. https://doi.org/10.1126/scisignal.2001592

Yegnasubramanian, S., Wu, Z., Haffner, M.C., Esopi, D., Aryee, M.J., Badrinath, R., He, T.L., Morgan, J.D., Carvalho, B., Zheng, Q., De Marzo, A.M., Irizarry, R.A., Nelson, W.G., 2011. Chromosome-wide mapping of DNA methylation patterns in normal and malignant prostate cells reveals pervasive methylation of gene-associated and conserved intergenic sequences. BMC Genomics 12, 313. https://doi.org/10.1186/1471-2164-12-313

You, J.S., Kelly, T.K., De Carvalho, D.D., Taberlay, P.C., Liang, G., Jones, P.A., 2011. OCT4 establishes and maintains nucleosome-depleted regions that provide additional layers of epigenetic regulation of its target genes. Proc. Natl. Acad. Sci. https://doi.org/10.1073/pnas.1111309108

Young, M.D., Wakefield, M.J., Smyth, G.K., Oshlack, A., 2010. Gene ontology analysis for RNA-seq: accounting for selection bias. Genome Biol. https://doi.org/10.1186/gb-2010-11-2-r14

Young, R.A., Lau, A., Lin, C.Y., Bradner, J.E., Lovén, J., Orlando, D.A., Lee, T.I., Vakoc, C.R., Hoke, H.A., 2013. Selective Inhibition of Tumor Oncogenes by Disruption of Super-Enhancers. Cell 153, 320–334. https://doi.org/10.1016/j.cell.2013.03.036

Yue, F., n.d. Northwestern University Genome Browser [WWW Document]. URL http://promoter.bx.psu.edu/hi-c

Zaret, K.S., Mango, S.E., 2016. Pioneer transcription factors, chromatin dynamics, and cell fate control. Curr. Opin. Genet. Dev. https://doi.org/10.1016/j.gde.2015.12.003

Zhan, F., Hardin, J., Kordsmeier, B., Bumm, K., Zheng, M., Tian, E., Sanderson, R., Yang, Y., Wilson, C., Zangari, M., Anaissie, E., Morris, C., Muwalla, F., Van Rhee, F., Fassas, A., Crowley, J., Tricot, G., Barlogie, B., Shaughnessy, J., 2002. Global gene expression profiling of multiple myeloma, monoclonal gammopathy of undetermined significance, and normal bone marrow plasma cells. Blood. https://doi.org/10.1182/blood.V99.5.1745

Zhan, F., Huang, Y., Colla, S., Stewart, J.P., Hanamura, I., Gupta, S., Epstein, J., Yaccoby, S., Sawyer, J., Burington, B., Anaissie, E., Hollmig, K., Pineda-Roman, M., Tricot, G., Van Rhee, F., Walker, R., Zangari, M., Crowley, J., Barlogie, B., Shaughnessy, J.D., 2006. The molecular classification of multiple myeloma. Blood. https://doi.org/10.1182/blood-2005-11-013458

Zhang, T., Li, B., Feng, Q., Xu, Z., Huang, C., Wu, H., Chen, Z., Hu, L., Gao, L., Liu, P., Yang, G.,

Zhang, H., Lu, K., Li, T., Tao, Y., Wu, X., Shi, J., Zhu, W., 2018. DCZ0801, a novel compound, induces cell apoptosis and cell cycle arrest via MAPK pathway in multiple myeloma. Acta Biochim. Biophys. Sin. (Shanghai). https://doi.org/10.1093/abbs/gmz033

Zhang, X., Choi, P.S., Francis, J.M., Imielinski, M., Watanabe, H., Cherniack, A.D., Meyerson, M., 2015. Identification of focally amplified lineage-specific super-enhancers in human epithelial cancers. Nat. Genet. 1–8. https://doi.org/10.1038/ng.3470

Zhang, Y., Liu, T., Meyer, C.A., Eeckhoute, J., Johnson, D.S., Bernstein, B.E., Nussbaum, C., Myers, R.M., Brown, M., Li, W., Shirley, X.S., 2008. Model-based analysis of ChIP-Seq (MACS). Genome Biol. https://doi.org/10.1186/gb-2008-9-9-r137

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., Dekker, J., 2012a. Chromosomal translocations are guided by the spatial organization of the genome. Cell 148, 908–921. https://doi.org/10.1016/j.cell.2012.02.002.Chromosomal

Zhang, Y., McCord, R.P., Ho, Y.J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., Dekker, J., 2012b. Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell 148, 908–921. https://doi.org/10.1016/j.cell.2012.02.002

Zhao, L., Lee, V.H.F., Ng, M.K., Yan, H., Bijlsma, M.F., 2018. Molecular subtyping of cancer: current status and moving toward clinical applications. Brief. Bioinform. 1–13. https://doi.org/10.1093/bib/bby026

Zheng, S., Papalexi, E., Butler, A., Stephenson, W., Satija, R., 2018. Molecular transitions in early progenitors during human cord blood hematopoiesis. Mol. Syst. Biol. 14, 1–20. https://doi.org/10.15252/msb.20178041

Zhou, M., Zhao, H., Wang, Z., Cheng, L., Yang, L., Shi, H., Yang, H., Sun, J., 2015. Identification and validation of potential prognostic lncRNA biomarkers for predicting survival in patients with multiple myeloma. J. Exp. Clin. Cancer Res. https://doi.org/10.1186/s13046-015-0219-5

Zhou, Y., Barlogie, B., Shaughnessy, J.D., 2009. The molecular characterization and clinical management of multiple myeloma in the post-genome era. Leukemia. https://doi.org/10.1038/leu.2009.160

Zhou, Y., Zhang, Q., Stephens, O., Heuck, C.J., Tian, E., Sawyer, J.R., Cartron-Mizeracki, M.A.,

Qu, P., Keller, J., Epstein, J., Barlogie, B., Shaughnessy, J.D., 2012. Prediction of cytogenetic abnormalities with gene expression profiles. Blood. https://doi.org/10.1182/blood-2011-10-388702

Zhu, Y., Sun, L., Chen, Z., Whitaker, J.W., Wang, T., Wang, W., 2013. Predicting enhancer transcription and activity from chromatin modifications. Nucleic Acids Res. https://doi.org/10.1093/nar/gkt826

Zlotorynski, E., 2018. Developmental enhancers in action. Nat. Publ. Gr. 2018. https://doi.org/10.1038/nrm.2018.15