# Modelling the 3D-Genome: The development of network theory approaches to characterise and predict active enhancers

**Maninder M. S. Heer**

Biosciences Institute

Newcastle University

Doctor of Philosophy

September 2021

# Abstract

Gene regulation is an important mechanism that ensures the correct functioning of a cell and is generally orchestrated by gene regulatory elements such as transcriptional enhancers. Identification of these genomic regions are important in understanding a wide range of phenomena such as evolution, homeostasis and disease. During gene regulation, signals pertaining to transcriptional activation are transferred across the chromatin regulatory network from enhancers to genes in the form of transcription factors and cofactors that in turn, recruit transcriptional machinery such as RNA Polymerase II to increase the rate of gene transcription. Conceptually, we describe this as a flow of information from enhancers to genes, mediated by the chromatin conformation. We exploit this relationship in order to decode the regulatory landscape of genes and identify active enhancers.

This thesis outlines the difficulties associated with identifying pathogenic mutations in the non-coding genome due to a lack of robust enhancer annotations. We use network theory to annotate these regions and develop a new method, 3D-SearchE, that serves to predict the location of novel putative active enhancers. 3D-SearchE achieves this by reverse engineering the flow of information between enhancers and genes to calculate an **imputed activity score (IAS)** at intergenic loci. We show that intergenic loci with a high IAS are also present for other enhancer associated features including the histone marks H3K27ac, H3K4me1 and H3K4me2, P300, CAGE-seq, Starr-seq, eQTLs and RNA Polymerase II. 3D-SearchE successfully leverages and summarises the relationship between the 3D organisation of chromatin and global gene expression and represents a novel enhancer associated feature that can be used to predict active enhancers.

# Publications

---

Maninder Heer*, Luca Giudice*, Rosalba Giugno, Daniel Rico - ***3D-SearchE: Leveraging gene expression and chromatin architecture to predict intergenic enhancers.*** Unpublished, in preparation.
* Equal Contribution

Maninder Heer*, Preeti Singh*, Anastasia Resteu, Aneta Mikulasova, Laetitia Largeaud, Stephanie Dufrechou, Nais Prade, Rachel Dickinson, Jacinta Bustamante, Venetia Bigley, Eric Delabesse, Daniel Rico, Marlene Pasquet, Matthew Collin - ***GATA2 deficiency phenotype associated with tandem duplication of GATA2 and over-expression of GATA2-AS.*** Blood Advances.
* Equal contribution

# Acknowledgements

I would like to thank my supervisors Dr. Daniel Rico and Professor Sophie Hambleton for their guidance and support throughout my PhD. Daniel, your positivity and enthusiasm always kept me going and encouraged me to explore my interests. I'm looking forward to having a beer with you after submitting!

I would also like to thank my collaborators that have worked with me on the projects that are in this thesis. I literally couldn't have done it without everyone's input and hard work over the last 4 years.

Thanks to everyone in the lab, past and present. Coming into work wouldn't have been anywhere near as enjoyable without Marco's in depth explanations about the e number or sharing my Covid office with Dan.

A huge thanks to all the friends that have kept me going with trips and visits and everything else in between. You've made the last 4 years immeasurably easier #2M, #AllDay #PubClub

Beccy, you're my favourite #2 bioinformatician ever. You supplied me with endless snacks while I wrote this thesis and even though you won the bet and submitted your thesis before me I can't wait for not doing a PhD with you in *insert the country*. Fankz.

Last but definitely not least, I want to say a big thanks to my family. Mum and Dad none of this would have been possible without amazing parents like you two. Thanks for getting me this far! Jaspur (a.k.a.Stinky) I expect that you address me as Dr. Fathead from now on.

# Contents

# List of figures

# List of tables

# Chapter 1 - Introduction

## 1.1 Finding function in junk: The non-coding genome

Four letters, A, T, C and G represent the fundamental information required for life. Humans contain in excess of three billion of these letters, termed bases, the equivalent of 3.3 Giga-bytes of data within our DNA. Classically, genes have been regarded as the primary source of genetic information. Genes are composed of protein coding regions known as exons and non-coding regions known as introns. The central dogma of genetics implies a linear relationship between DNA, RNA and protein in which information encoded within the exons is translated into proteins via the triplet code (Crick 1958). When this flow of information is disrupted or altered it can often manifest in aberrant cell function and in extreme cases disease. At the DNA level, variation in the exonic base sequence can directly affect the amino acid sequence and therefore, protein function. This can occur by a single nucleotide polymorphisms (SNP) resulting in synonymous or nonsynonymous substitutions or by copy number variation (CNV) that include duplications and deletions. In the post genomic era comprehensive databases such as GnomAD (Lek et al. 2016), formerly known as ExAC and The 100,000 Genomes Project (Caulfield et al. 2017) have been curated detailing such alterations of the genetic code. Therefore, our ability to study the effects of SNPs and CNVs and dissect their relevance in the context of human disease has become increasingly feasible.

While the post-genomic era has had a huge impact in our understanding of many diseases, it has largely been confined to the study of exonic sequences. However, exons constitute roughly just two to three percent of DNA (via circa 22,000 genes) in humans and mice. While exons are needed to synthesise proteins, their expression must be carefully coordinated in order to give rise to the 200+ cells that constitute every human being. This is achieved in part by the remaining ~98% of ‘non-coding’ DNA. The non-coding regions can be categorised, broadly, into two domains. Introns that intersperse exons and intergenic regions that delineate genes. There has been much debate about the proportion of non-coding DNA that is functional and can influence gene expression and the proportion that is non-functional and redundant. During the mid 20th century the term “junk DNA” was initially coined to describe the apparent lack of function of the non-coding genome. It was first posited by Ehret and De Haller in which they remark that “it does not follow that all genetic DNA is competent genetic material (viz. “Junk DNA”)” (Ehret and De Haller 1963). The idea that most of the non-coding genome is functionless was subsequently formalised by Susumu Ohno in his article ‘So much “junk” DNA in our genome’ (Ohno 1972). Ohno’s reasoning for junk DNA was based on previous observations that larger genomes do not scale with increasing complexity of the organism, known as the C-value paradox (Thomas 1971). This idea argues that if this is not the case, onions, lungfish and salamanders, all of whom have larger genomes, can be considered our genetic superiors. These models suggest the majority of junk DNA within our genomes can be attributed to the presence of pseudogenes and transposable elements. However, certainly not all DNA beyond exons is junk.

It is within the non-coding genome that many functional elements are thought to reside, including cis- and trans-regulatory elements. In eukaryotes, cis-regulatory elements include promoters, enhancers and silencers. Trans-regulatory elements include transcription factors, DNA editing proteins, post-transcriptional mRNA processing and mRNA binding. Enhancers are of particular interest given their ability to directly upregulate transcriptional activity. However, their composition and function is poorly understood. Estimates for the number of enhancers across the genome run into the hundreds of thousands. This again raises the age old question of just how much of DNA can be considered functional and how much should be regarded as junk. The post genomic era has ushered in a new wave of evidence that suggests an increasingly important functional role for much of the non-coding genome. As recently as 2012, The Encyclopedia of DNA Elements (ENCODE) was published by The ENCODE Project Consortium with claims that they had assigned biochemical functionality to 80% of the human genome (ENCODE Project Consortium 2012).

These claims are in stark contrast to those posited by Ohno and his contemporaries. These differences are partly due to different definitions of the term ‘functional’ (Graur et al. 2013). For example, from an evolutionary context the biochemically active regions defined by ENCODE cover a much larger region of the genome when compared to evolutionary conserved regions which raises doubts about their functionality (Kellis et al. 2014). Furthermore, ENCODEs data implies that transcription, histone modifications, open chromatin, transcription factor binding and DNA methylation indicate function which is disputed (Graur, Zheng, and Azevedo 2015; Kellis et al. 2014). On the contrary, expression quantitative loci (eQTL) studies have shown how genetic variation across the genome can result in significant and quantifiable changes in gene expression (F. Zhang and Lupski 2015). Single nucleotide and copy-number variants have been shown to alter gene expression patterns through alterations in the structural organisation of chromatin (Sadowski et al. 2019). Additionally, enhancers have been found with a distinct lack of sequence conservation (Harmston et al. 2017). While there is also increasing evidence that subtle genome wide influences play a role in complex traits termed the ‘omnigenic’ model (Boyle, Li, and Pritchard 2017). The ambiguity around the term functional, coupled with a lack of knowledge about regulatory sequences and the mechanisms via which they act make estimations about the amount of functional and junk DNA purely academic.

To address these issues, understanding the relationship between DNA sequences such as genes and enhancers in the context of chromatin organisation will be important. We know that the ability of enhancers to regulate the transcriptional activity of genes is aided or impeded by the cell-type specific three-dimensional organisational structure of the chromatin. This is what allows the 200+ individual cell-types to arise from the single set of instructions encoded within the DNA. It has also been shown how variants that disrupt the topological organisation of the chromatin can lead to aberrant gene expression. This would indicate that genes, enhancers and the organisation of chromatin are critical for the normal expression of genes. The relationship between the organisation of chromatin, which we will refer to herein as chromatin architecture, and gene expression is poorly understood. Akin to the chicken and the egg: It is apparent that the chromatin architecture plays a role in the regulation of gene expression while gene expression can drive changes in the chromatin architecture. It is through largely unknown complex feedback mechanisms that manage the coordinated localisation of enhancers to promoters through changes in the chromatin

architecture. Thus, ENCODEs assertion that 80% of the genome is functionally relevant may be correct at the level of chromatin rather than the DNA sequence. It is entirely plausible that 80% of chromatin is functionally active in contorting the DNA into structures that facilitate the expression of genes, thereby acting as a regulatory entity in and of itself. Therefore, research that aims to characterise enhancers in the context of chromatin architecture is of utmost importance.

## 1.2 Enhancers are diverse in composition and function

Enhancers are defined as cis-acting DNA sequences that harbour motifs corresponding to transcription factors (TF's) and cofactors that increase the transcriptional activity of one or more genes above basal levels of transcription afforded by the core promoter. The first example of a regulatory sequence was described by Grosschedl and Birnsteil et al. nearly 40 years ago. They were able to demonstrate a 15- to 20-fold reduction in transcriptional activity of H2A gene activity by deleting a distal sequence of interest from the sea urchin genome (Grosschedl and Birnstiel 1980). The first description of an enhancer soon followed, after work by Julian Banerji in the lab of Walter Schaffner identified a 72bp sequence in the simian virus 40 genome (Banerji, Olson, and Schaffner 1983). Not long after, the first cellular enhancer was discovered within an intron of the mouse human immunoglobulin heavy chain gene (Banerji, Olson, and Schaffner 1983). Importantly, it was presented as a cell type specific enhancer, and thus opened up a new breeding ground for studies into cell-type specific regulation. Enhancers appear to be the main drivers behind the differentially expressed sets of genes between cell-types. It is these varying combinations that result in the diverse array of phenotypes we observe. This has been elegantly demonstrated in the *Drosophila pox neuro* gene. Fifteen enhancers were identified as key regulators of the *pox neuro* gene's expression where they were shown to influence the divergence of developmental phenotypes depending on the presence of one or more of the enhancer sequences (Boll and Noll 2002).

We now know that enhancers are essential to regulate the transcriptional state of almost all genes, housekeeping genes being the exception. Enhancers are able to regulate their cognate genes through direct interactions and have been shown experimentally to occur through forced chromatin looping (Carter et al. 2002; Deng et al. 2012). More recent studies have supported this concept by showing the co-localisation of *Shh* with the ZRS enhancer through imaging and 5C (**see 1.4 Enhancers in context: Chromatin architecture**) (Williamson et al. 2016). While the looping of chromatin has been shown to mediate enhancer-promoter contacts, the stability of these loops have been questioned. It has been shown using reporter gene constructs that two genes can be co-regulated by a single enhancer, leading to co-ordinated bursts of transcription suggesting that enhancer-promoter contacts are not stable (Fukaya, Lim, and Levine 2016). Other work has shown using live imaging that the short transcriptional bursting of the *Sox2* gene is independent of direct enhancer contacts (Alexander et al. 2019). These studies suggest that enhancer-promoter contacts are more dynamic than has been suggested previously (Ghavi-Helm et al. 2014). These dynamic contacts are intricately linked with the topology of chromatin that is organised through a complex interplay of enhancers and proteins such as transcription factors (TFs) and RNA pol II to name a few. This is discussed in detail in **1.4 Enhancers in context: Chromatin architecture**.

Defining enhancers and understanding their spatio-temporal localisation with promoters is important in the context of understanding a wide variety of genetic phenomena such as evolution, homeostasis and disease. For example, the translocation of the GATA2 superhancer to EVI1 results in a downregulation of the GATA2 gene with the concomitant upregulation of the EVI1 gene manifesting as acute myeloid lymphoma (Gröschel et al. 2014). Despite the success in identifying the regulatory mechanisms of some genes, attempts to consolidate enhancers, genetic variation and gene expression are often fraught with problems. **The difficulty in understanding this relationship is derivative of the current issues surrounding enhancer identification and subsequent enhancer-gene pairing**. The enhancer characteristics underpinning these difficulties can be broadly summarised by the following points:

1) Enhancers are promiscuous. A single enhancer can regulate multiple genes (Benabdallah et al. 2019), and conversely, a gene may be regulated by multiple enhancers.

2) Enhancers are dynamic in location and activity. Enhancers can be located upto and in some cases beyond one megabase away from its cognate gene (Schoenfelder and Fraser 2019). In addition to this, enhancers can interact with promoters in mechanistic fashion beyond the prototypical enhancer/ promoter models described previously. For example, the spatio-temporal activity of enhancers can be extremely dynamic and may not always be captured by chromosome conformation assays that reproduce a snapshot style picture of the chromatin contact landscape (Benabdallah et al. 2019). Additionally, enhancers may not act through direct interactions with the genes they regulate (Benabdallah et al. 2019; W. Song, Sharan, and Ovcharenko 2019)

3) Enhancer sequence composition is not ubiquitous. There is no known conserved enhancer sequence that universally defines all enhancers (Harmston 2020).

4) Redundancy. Perturbing one enhancer can have a minimal effect on the expression of a gene that may rely on several enhancers (Osterwalder et al. 2018). Enhancers are therefore extremely diverse in both their composition and their mode of action. To identify enhancers there are several approaches which tend to be low throughput and very accurate or high throughput with the tradeoff of lower accuracy.

## 1.3 Identifying enhancers

### 1.3.1 Enhancer screening

Wet-lab based approaches such as large scale reporter assays such as massively parallel reporter assays (MPRA) have traditionally been used as a brute force attempt to identify enhancers genome wide. However, these assays suffer from issues with sensitivity and specificity (Inoue and Ahituv 2015). More targeted approaches such as with CRISPR-Cas9 technology have proved extremely successful in identifying functional enhancers for specific genes (Korkmaz et al. 2016). Efforts to scale up CRISPR based screens have been achieved (Fulco et al. 2019). Still, the method remains prohibitively expensive and relatively

low throughput. In order to reliably identify enhancers at scale a combination of features that are commonly associated with enhancers can be used for genome wide identification of enhancers.

### 1.3.2 Sequence conservation

Enhancers can be identified by a number of features. One such feature that has been exploited is sequence conservation. Through comparative genomics, methods have been developed to identify highly conserved regions across mammalian species which were, in turn, shown to be enriched in enhancer elements (Bejerano et al. 2004; Pennacchio et al. 2006; Harmston et al. 2017). However, the conservation of sequence is not limited to enhancers and therefore such methods result in many false positives. Indeed, it is also possible to identify enhancers that are not conserved in sequence but are still able to drive gene expression in a similar fashion to their orthologues (Harmston 2020).

### 1.3.3 RNA Polymerase II and enhancer transcription

RNA polymerase II is the essential machinery required to carry out transcription of genes into mRNA. RNA Pol II is able to coordinate transcription through modification of its carboxy-terminal domain (CTD), a constituent of its largest subunit (Cho et al. 1997). Changes in the phosphorylation state of the CTD can modulate the transcriptional activity of RNA Pol II, together these modifications constitute the 'CTD code' (Harlen and Churchman 2017). The Phosphorylation of serine 5 and 7 are typically associated with promoter regions dependent on transcription factor II H (TFIIH). Serine 7 phosphorylation can also be observed at promoter distal regions. Serine 2 phosphorylation is another modification that can be found during transcriptional elongation (Komarnitsky, Cho, and Buratowski 2000; Ni et al. 2011). Both the occupancy of RNA Pol II and the subsequent transcription at non-coding loci can be indicative of enhancer activity. The physicochemical properties of the CTD code have also been shown to influence the formation of phase separation and other similar domains (**see 1.4.3 Phase separation and transcriptional factories**) (Harlen and Churchman 2017).

For a more direct assessment of enhancer activity the transcriptional activity of non-coding sequences can be determined by quantifying the amount of RNA using methods such as RNA-seq. Like genes, some enhancers are transcribed into enhancer RNAs (eRNA) via RNA polymerase II (Kim et al. 2010). Several methods have therefore been developed to measure the transcriptional activity of putative enhancer sequences. Cap analysis of gene expression (CAGE-seq) has been used in the FANTOM5 project to compile a list of putative enhancers across a range of cell-types (Noguchi et al. 2017; Andersson et al. 2014). This method identifies regions of bi-directional transcription which is a feature associated with enhancers. However, like most enhancer features, bi-directional transcription may not be exclusive to enhancers, nor are they ubiquitous across all enhancers (Young et al. 2017). Alternatively, self-transcribing active regulatory region sequencing (STARR-seq) can be used to identify enhancers through their transcriptional activity (Arnold et al. 2013). This approach is more direct and aims to identify sequences of DNA that are able to transcribe themselves when placed downstream of a minimal promoter. The amount of transcribed sequence is then mapped back on the reference genome where the read depth is indicative of the enhancer activity.

### 1.3.4 Chromatin modifications

DNA is often sequestered with proteins into chromatin which helps to organise the chromatin into three-dimensional structures. There is increasing evidence that the proteins that form the chromatin can directly affect the expression of genes through the modification of its histone tail that then influences the recruitment of transcriptional machinery. Additionally, these proteins may have a more passive role in gene regulation such as by modifying the chromatin architecture that makes the DNA more accessible to transcriptional machinery and mediating chromatin interactions between enhancers and promoters. Chromatin modifications can also modulate the physicochemical properties of the chromatin and is associated with the three-dimensional organisation of chromatin (**see 1.3: Enhancers in context: Chromatin architecture**). For example, homotypic attraction describes the process in which chromatin regions with similar properties tend to attract one another. This results in localised distinct compartments within the nucleoplasm. There are several models that describe the consequences of the physicochemical properties and are discussed in further detail later on (**see 1.3.3: Phase separation and transcription factories**). These processes are important to understand in the context of gene regulation as it has been shown through genomic rearrangements and transposition how active genes can be silenced through the abnormal juxtaposition with heterochromatic regions in a process termed position effect variegation (Baker 1968; Karpen 1994). This would suggest that the topology of the chromatin architecture plays a role in gene regulation.

With the advent of next generation sequencing, the study of these proteins have become increasingly accessible. Coupled with chromatin immunoprecipitation (ChIP), ChIP-seq as it is known, has enabled comprehensive genome wide mapping of proteins including transcription factors and coactivators as well as histone modifications. ChIP-seq works by sequencing DNA that binds to an antibody specific to a transcription factor or histone modification allowing for the study of the following genomic features.

#### 1.3.4.1 Coactivators

The P300-CBP coactivator family are two closely related proteins that have been shown to increase the transcription of their target genes. Both P300 and CBP act via three main mechanisms to increase the rate of transcription. 1) Modulation of chromatin via histone acetyltransferase (HAT) activity that relaxes chromatin and increases the accessibility of DNA to transcriptional machinery (Raisner et al. 2018). 2) Recruitment of transcriptional machinery such as RNA polymerase II (von Mikecz et al. 2000). 3) Acting as a transcriptional coactivator through binding activators such as transcription factors (J. Chen and Li 2011). P300 has been shown to exist largely at DNaseI hypersensitive sites (DHSs) while up to 75% of P300 has been shown to bind in regions distal to promoters. Indeed, P300 has been shown to localise in regions with histone modifications that are indicative of enhancers (Heintzman et al. 2007). P300 is therefore often used to identify enhancer sequences (Z. Wang et al. 2009; Visel et al. 2009).

#### 1.3.4.2 Transcription factor binding

Enhancers can also be identified through their binding of transcription factors (TFs) which typically bind short degenerate sequences called ‘motifs’. TFs work by binding to enhancers in a combinatorial fashion with multiple homotypic or heterotypic binding events. It is largely

through the combinatorial binding of the 1600 known transcription factors that enhancers can orchestrate the complex spatial and temporal expression of genes (Lambert et al. 2018). In a similar fashion to coactivators, TF binding at enhancers work to regulate gene expression through various mechanisms: A) mediating RNA Pol II recruitment or exclusion. B) modifying histone tails to make transcription more amenable through histone acetyltransferases or more difficult through histone deacetylase activity (HDAC) (Narlikar, Fan, and Kingston 2002). C) the recruitment of coactivators such as P300. The action of TFs generally appears to coordinate the expression of genes through localised changes in chromatin structure and the surrounding microenvironment (**see 1.3.4 Transcription factories**), while they are also directly involved in transcription.

#### 1.3.4.3 Histone modifications

The formation of chromatin occurs when DNA wraps around octamers made up of four core histones, H2a, H2b, H3 and H4, to form nucleosomes. Here, 147 base pairs of DNA are wrapped in 1.65 turns around each histone protein with 50bp of free DNA between them. Each of these four core histones contain a globular C-terminal domain and an unstructured N-terminal domain that are subject to a diverse set of post translational modifications such as methylation, phosphorylation, ubiquitination and sumoylation by histone modifying enzymes.

Histone modifications are one of the most commonly used features to identify enhancers. There are two main histone modifications that are used: H3K4me1 which is correlated with enhancers in the active, inactive and primed states (Heintzman et al. 2007); and H3K27ac which is correlated with enhancers in the active state when found with H3K4me1 (Creyghton et al. 2010). There is also increasing evidence that promoters can also have enhancer activity and as such have been termed ePromoters (Dao et al. 2017). Although useful, histone modifications like all other enhancer features are not a perfect indicator of enhancer function. In fact, the functional role of histone modifications and their relationship with enhancers are not fully understood. For example, some studies have shown that both an increase and decrease in global levels of H3K4me1 result in negligible effects on gene expression. Some speculate that histone marks such as H3K4me1 may be a bi-product of the cognate methyltransferase M113/4 whose function is the key driver behind changes in gene expression. The acetylation of H3K27 is carried out by acetyltransferases such as P300/CBP. Again, its function is not fully understood but acetylation appears to neutralise the positive charge on the lysine which means the positive charge no longer attracts the negative charge of DNA and relaxes the chromatin. This residue can also be antagonistically tri-methylated during polycomb repressive complex 2 (PRC2) mediated silencing. H3K27ac has therefore also been suggested to act as a bookmark by preventing PRC2 mediated methylation (Pengelly et al. 2013).

### 1.3.5 Chromatin accessibility

Changes to the chromatin generally act to make chromatin less accessible in the case of heterochromatin or more accessible in the case of euchromatin where genes in the latter are more likely to be transcribed. Within euchromatin exist DNase hypersensitivity sites (DHSs) that are known to contain transcriptionally active regions of DNA such as genes and enhancers. Experimental techniques have therefore been developed to exploit this phenomenon such as DNase accessibility assays. DNase accessibility assays utilise the

enzyme DNase I to digest DHSs (Dorschner et al. 2004). An alternative approach is ATAC-seq. This approach offers a faster and more sensitive approach by inserting sequencing adapters at open regions of the chromatin using an adapted Tn5 transposase (Buenrostro et al. 2015). These regions can be overlapped with various other more direct features of enhancers to aid in their identification.

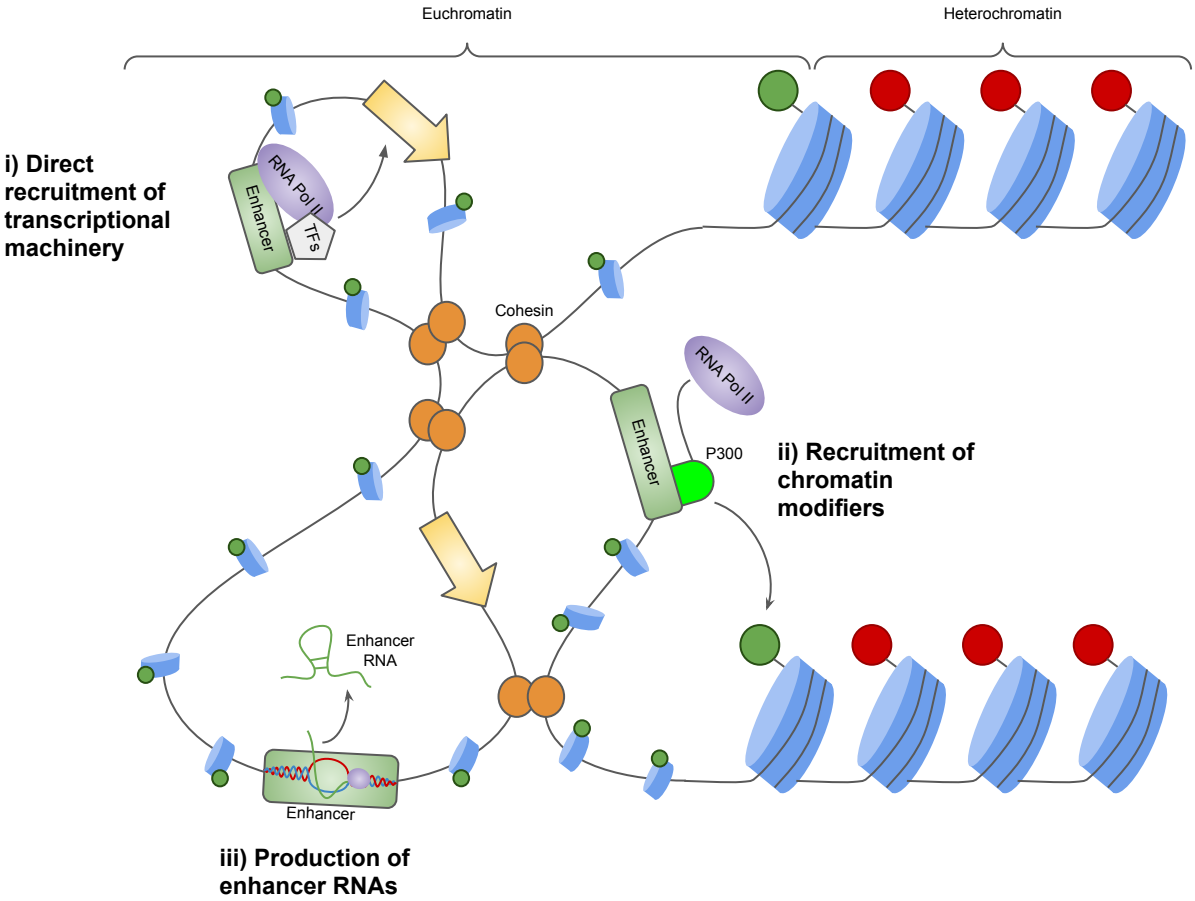



***Figure 1.1. Enhancer features and mode of action.*** i) Enhancers can recruit transcriptional machinery including RNA Polymerase II and transcription factors. ii) Enhancers can recruit chromatin modifiers such as P300 to make chromatin more accessible through histone acetyltransferase activity. iii) Enhancers are transcribed during activation of a gene to produce enhancer RNA's.

## 1.4 Enhancers in context: Chromatin architecture

The enhancer features described previously can be used to identify enhancer sequences. However, reconciling them with their cognate enhancers is more difficult. Enhancers are known to be able to regulate gene expression irrespective of the linear distance and orientation of the target gene. The reason that distal enhancers are able to interact with genes that can be up to, and in some cases beyond 1Mb away, is a consequence of the flexibility of the DNA polymer and their physical interaction with their target promoters (Dekker et al. 2002a; Bickmore 2013). To understand how and why enhancers work we must first consider the nature of the relationship between distal enhancers and their target promoters. This relationship is best explained by the organisation of chromatin into higher order structures such as A/B compartments (Lieberman-Aiden et al. 2009a), nuclear lamina associated domains (LADs) (Lieberman-Aiden et al. 2009b) and topologically associated domains (TADs) (Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012a); which package the linear DNA polymer into increasingly more complex secondary and tertiary structures. These structures fold the DNA in such a way that the relevant enhancers, TFs and transcriptional machinery can access genes through chromatin looping and localised changes in nucleoplasm (Fudenberg et al. 2016; Hnisz et al. 2017). The structural organisation of the chromatin results in a multilayered topology of chromatin organisation that dictates a unique set of transcriptional instructions for each of the ~200 different known cell types. It has also been shown that the phenomenon of chromatin directed transcription occurs between individual cells of the same type where single cell genomics has shown variation in the transcriptomes of cell sub-populations (Trapnell 2015).

In previous years, attempts to probe the underlying basis of the chromatin architecture have been restrained by experimental limitations. It is now an intensely active field of research largely owing to the development of new protocols such as formaldehyde cross linking studies including 3C (Dekker et al. 2002b) and its derivatives 4C (Zhao et al. 2006), 5C (Schoenfelder et al. 2018), Hi-C (Belton et al. 2012) and more specialised enrichment conformation captures such as promoter capture Hi-C (PCHi-C) (Schoenfelder et al. 2018). These methods are collectively referred to as 3C based methods. Fundamentally, they all work by cross-linking, digesting and sequencing interacting DNA fragments. The differences between the methods arise from the “view-point” where 3C identifies interactions between a single pair of genomic loci, 4C the interactions between a single loci with many other loci, 5C that identifies many interacting loci in defined regions and finally, Hi-C that aims to identify interacting pairs genome wide. Additionally, the study of chromatin architecture has been improved by technologies such as fluorescent *in situ* hybridisation (FISH) (Gall and Pardue 1969) have been coupled with superresolution microscopy (Sigal, Zhou, and Zhuang 2018). This has allowed for the study of chromatin organisation at an unprecedented level of detail and has furthered our understanding of how gene expression and chromatin organisation influence one another.

### 1.4.1 A and B compartments

At the highest level the chromatin is segregated into two distinct types of chromatin, termed A and B compartments. A compartments typically consist of gene rich chromatin and active histone marks and are located towards the centre of the nucleus. B compartments on the

other hand tend to be gene poor, with chromatin markers consistent with transcriptional silencing and tend to be located at the nuclear periphery (Lieberman-Aiden et al. 2009b).

### 1.4.2 Topologically associating domains (TADs) and chromosomal looping

Within A/B domains chromatin is further organised into structures at the 10Kb to Mb resolution, known as topologically associating domains (TADs) (Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012b). TADs are the most striking feature observed in chromosome conformation capture studies and display a relatively higher internal frequency of contacts. Their function is largely unknown with several explanations as to their existence. TADs are typically demarcated by boundaries consisting of convergent CTCF binding motifs (Rao et al. 2014). According to the loop extrusion model, the prevailing theory on TAD formation, these act as stalling sites for DNA loops that are traversed through cohesin rings (Fudenberg et al. 2016). Indeed, depletion of CTCF has been shown to disrupt TAD structure but interestingly, does not extend further to A/B compartmentalisation (Rao et al. 2014). There is also increasing evidence that the general organisation of chromatin plays a more subtle role in the regulation of gene expression (Sadowski et al. 2019).

TADs and similar domains are observed across species from mammals (Dixon et al. 2012; Nora et al. 2012; Sexton et al. 2012b) to *Drosophila (Sexton et al. 2012b)* where TADs were first described. As well as in *C.elegans* (Crane et al. 2015) and *A.thaliana (Congmao Wang et al. 2015)*, while TAD like domains have also been described in both yeast and bacteria (Mizuguchi et al. 2014; Le et al. 2013). TADs are also shown to be conserved between species and may serve as stable heritable blocks of the genome (Harmston et al. 2017; Krefting, Andrade-Navarro, and Ibn-Salem 2018). However, TADs at the structural level show a diversity between the species described, and the mechanisms in which they are formed and maintained vary too. Despite this, one feature of TADs that remains unwavering between these species is the association between their structure and gene expression, particularly in the context of cell fate determination (Gonzalez-Sandoval and Gasser 2016). Additionally, TAD boundaries correlate strongly with replication domains and may serve as a regulatory unit of replication timing (Pope et al. 2014). Several studies have also implicated the perturbation of the TAD architecture by boundary element disruption with disease (Ibn-Salem et al. 2014; Spielmann and Mundlos 2013; Anne-Laure Valton 2016). However there is an argument to be had on whether the loss or alteration of TAD structure itself is the driving factor and or if it is due to the specific perturbation of the boundaries (Despang et al. 2019).

### 1.4.3 Phase separation and transcription factories

The organisation of chromatin can also be influenced by the physicochemical properties of chromatin and associated proteins. Liquid-liquid phase separation (LLPS) is thought to be one of the processes that drive these organisational changes (Hnisz et al. 2017). LLPS describes membraneless compartmentalisation within the nucleus similar to those such as nucleoli, cajal bodies and nuclear speckles. These processes are akin to droplets of oil suspended in water. It is thought that elements such as enhancers and genes, proteins such as RNA Pol II and TFs and processes such as transcription producing coding- and non-coding RNA all contribute to the formation of LLPS regions.

Transcription factories are another model that describe discrete nucleoplasmic regions. Rather than describing a more general model of compartmentalisation, transcription factories are specific to foci of transcription. Indeed, phase separation has been suggested to be the underlying mechanism for the formation of transcription factories (Palikyras and Papantonis 2019).

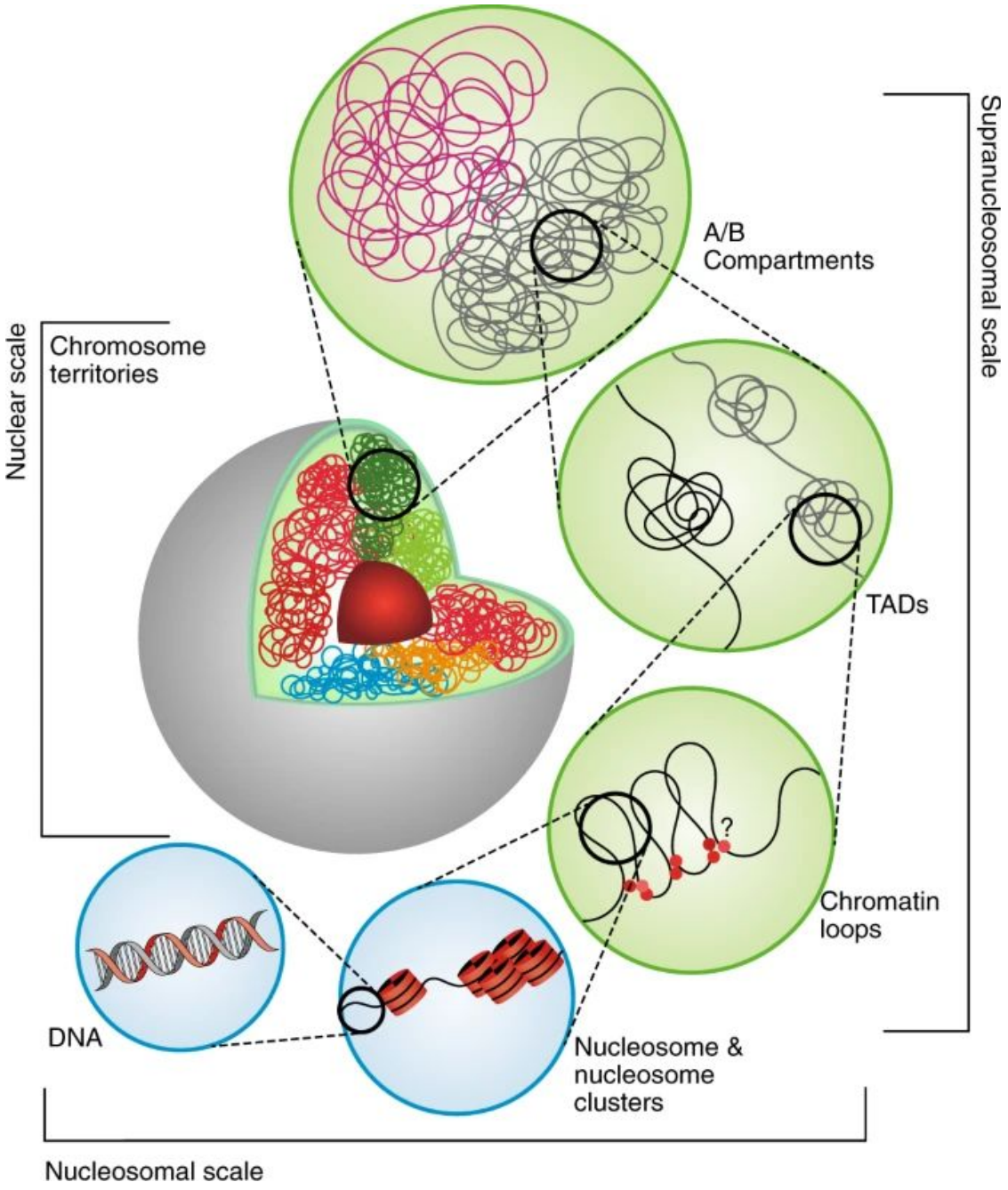



***Figure 1.2. The hierarchical structure of chromatin.*** Chromatin is organised into A/B compartments, topologically associating domains, chromatin loop and nucleosome clusters. Adapted from Dogan and Liu 2018.

## 1.5 Predicting enhancer position and target genes

Enhancers are incredibly diverse in their activity across different cell types, in addition to this, they contain little to no sequence or structural specificity that makes them easy to identify. This combined with the fact that they can often be located up to 1Mb away, presents a significant challenge of identifying enhancers and their target genes. Experimental methods for identifying enhancers are limited by their low throughput or cost or a combination of the two. Computational methods seek to address these issues by taking advantage of known enhancer features in order to predict genome wide *de novo* enhancer sequences. Indeed, there are many pioneering predictive algorithms that have sought to leverage this information in order to predict the location of enhancers and their cognate genes. Many of these algorithms utilise the features that are commonly associated with enhancers to train and validate their predictions. The use of *in silico* predictions provides a useful tool to supplement experimental data that can be costly. Here we will briefly discuss enhancer prediction methods, the benefits and drawbacks of these methods and outline the key requirements needed to improve existing models.

There are a plethora of computational methods to annotate enhancers genome wide. At the time of writing there are currently 749 publications that can be found using the specific search terms “enhancer prediction” and 131,000 publications are found where the search terms “enhancer” and “prediction” are used. This highlights not only the sheer number of methods being developed, but also their widespread adoption in enhancer biology research. Early enhancer prediction models investigated the DNA sequence as a predictor for enhancers. These often use k-mer based methods (oligomers of length k) coupled with a support vector machine (SVM) classifier to predict enhancer sequences (Lee, Karchin, and Beer 2011). With the advent of new enhancer associated features, methods integrating other sources of epigenomic data have been developed. RFECS for example utilises a random-forest based classifier that is trained on chromatin states and aims to predict enhancers defined by both P300 binding sites and DNase-I hypersensitive sites (Rajagopal et al. 2013). Other approaches, such as DEEP, present a hybrid SVM/ artificial neural network (ANN) model trained on multiple sources of enhancer data published from ENCODE, FANTOM5 and VISTA (Kleftogiannis, Kalnis, and Bajic 2015). One major caveat of methods that rely on linear genomic data to predict enhancers is that identifying their cognate gene. Assigning enhancer-promoter pairs is often dealt with by either window based linkage or nearest gene linkage. The window based linkage approach uses a defined region around a gene to assign an enhancer (Schmidt et al. 2017; Ouyang, Zhou, and Wong 2009). The nearest gene linkage approach identifies the nearest gene to an enhancer to establish a pair (Lee, Karchin, and Beer 2011). However, our knowledge of how enhancers can act over long distances as well as the complex structure of chromatin undermines the validity of these approaches. To solve this problem, 3C data that maps the 3D interactions of chromatin, can be used as a more robust measure of EP pairs (Moore et al. 2020).

Despite the advancements in the prediction of enhancers many problems persist. Much of these problems are rooted in the quality of the data used to generate and validate the predictions. A thematic issue in enhancer biology is, as previously discussed, the small number of validated enhancers in very few cell types due to the current limitations in the scalability of CRISPR based screening. This has two main problems: one is that the small number of validated enhancers means that this data is rarely useful to validate a models

predictions nor does its size lend itself to powerful statistical analysis; secondly, when enhancer features, such as the histone modifications used by ENCODE, are used in lieu of directly validated enhancers they should be considered ‘putative enhancers’ as they are not 100% accurate. The use of enhancer features as validation sets such as those provided by ENCODE, FANTOM5 and other means can lead to ‘validation creep’ in which non-validated putative enhancers are considered true enhancers. Using putative enhancers in this way can lead to predictions progressively moving further from the true set of enhancers (Halfon 2019).

The fundamental lack of knowledge of enhancers and how they work can also lead to bias in the interpretation of results. For example, the use of 3C data in validating EP pairs is typically based on the interaction between two chromatin fragments that contain a gene and a predicted enhancer. However, evidence suggests that enhancers can also act to regulate a gene without persistent direct interactions (Benabdallah et al. 2019), or through indirect interactions such as with enhancer chains (W. Song, Sharan, and Ovcharenko 2019). The current use of 3C data means that EP pairs are missed because the 3C technology is not yet sensitive enough to reliably and consistently detect transient EP contacts. Secondly, the validation using data representing pairs of interacting chromatin neglects our evolving understanding of how enhancers regulate gene expression.

To address these challenges higher throughput CRISPR screens have been developed to produce a more comprehensive list of validated enhancers although these are limited to a few cell-types (Fulco et al. 2019). Until robust databases of validated enhancers are available computational methods must be clear in that they are putative to prevent ‘validation creep’ (Halfon 2019). As such, it is becoming increasingly obvious that models must incorporate more enhancer features in order to predict and validate putative enhancers. Therefore, to aid the prediction and validation of enhancers new methods should be developed to identify novel features of enhancers. For example, 3C methods offer the opportunity to characterise and understand the spatial features of enhancers within the chromatin architecture and provide a potential new enhancer feature. This can be achieved in part by addressing the representation and investigation of EP contacts by considering the EP pairs beyond the classical model of direct interaction.

## 1.6 Summary

In order to link non coding variation with changes in gene expression we not only need to identify enhancers, but must also assign a gene or genes to the enhancer. This will greatly improve our understanding of how cells manifest with a wide range of phenotypes such as disease. In recent years there has been a great deal of success in identifying mutations in specific enhancers that affect specific genes that lead to specific phenotypes. These have relied upon low-throughput assays that can accurately define enhancer activity. While more high-throughput methods have been developed, they are neither cost efficient or as reliable. This has resulted in a scramble to computationally predict enhancer sequences genome wide using known enhancer features. The problem, however, is that we do not yet know enough about the makeup of enhancers to produce robust predictions. Nor can we faithfully predict the genes they regulate. It is also becoming increasingly apparent that while individual elements are essential for appropriate gene function, they often work in concert rather than as individuals. The expression of genes, the role of regulatory enhancers and the

conformation of chromatin in three-dimensional space are intimately linked and exhibit a recursive influence on one another. To unpick these dynamic relationships and improve enhancer predictions new integrative approaches must be developed to study gene regulation genome wide in a more holistic manner.

## 1.7 Aims and objectives of this thesis

The folding of chromatin into hierarchical structures ultimately acts to localise genes with the prerequisites for transcription such as RNA pol II and the appropriate chromatin modifications. The folding of chromatin into said structures also appears to be under the influence of RNA pol II and chromatin modifications. As to whether gene expression dictates the chromatin architecture or vice-versa is poorly understood. In either case, **it is clear that the organisation of chromatin plays an important role in mediating the communication between genes and enhancers**.

The aim of this thesis is to present a new body of work that outlines how the chromatin architecture and gene expression data can be used to find enhancers. To do so we represented 3C data in the form of networks. This allowed us to to use network theory approaches to investigate the topological characteristics of various loci within the chromatin architecture. We then aimed to reconcile these topological characteristics with enhancer loci. We then sought to improve the characterisation of enhancers by developing a framework to integrate gene expression data into the network. Ultimately, we hypothesised that the integration of gene expression data and chromatin conformation (3C) data could be used to further classify enhancer loci.

# Chapter 2: Targeted sequencing of the GATA2 locus of patients with GATA2-like deficiency reveals a novel structural alteration

---

## 2.1 Introduction

Hematopoiesis is the driving process in the differentiation and cell lineage commitment for hematopoietic stem cells (HSCs) (Ng and Alexander 2017). The proliferation and differentiation of HSCs to their terminal state are tightly regulated by epigenetic modifiers, regulatory elements and a carefully prescribed cocktail of transcription factors (Muench and Grimes 2015). GATA2 is expressed in early hematopoietic progenitors and is required for stem cell longevity and lympho-myeloid differentiation (Laurenti et al. 2013). The GATA2 gene is located in chromosome 3 and encodes a transcription factor that plays an essential role in regulating hematopoiesis by facilitating the development and proliferation of specific hematopoietic and endocrine cell lineages (Orkin 2000). During normal function both alleles, maternal and paternal, are transcribed. Perturbation of this delicate equilibrium can have significant and malignant downstream effects.

Single nucleotide variants (SNVs) in the GATA2 gene have been reported to cause amino acid substitutions, frameshift/deletions, loss of intronic enhancer function or aberrant splicing events (Spinner et al. 2014; Dickinson et al. 2014; Wlodarski et al. 2016; Donadieu et al. 2018; Kozyra et al. 2020). This can lead to the loss of either allele and the resultant haploinsufficiency leads to phenotypes including, but not limited to, monocytopenia, neutropenia, and mycobacterial infection, dendritic cell, B and natural killer (NK) lymphoid deficiencies, as well as familial myelodysplastic syndrome (MDS)/ acute myeloid leukemia (AML) and Emberger syndrome (Spinner et al. 2014). Typical GATA2 deficiency phenotypes can also present with the exclusion of any coding mutations. Although less common, the alteration of gene dosage by structural variation (SV) has been reported. This has been shown to occur through either germinal losses of the locus or the somatic inversion of its cognate -110 super enhancer by inv(3) (q21;q26) or t(3;3)(21;q26) (Gröschel et al. 2014; Vinh et al. 2018).

Here, we present a patient (KERNA) with classical features of GATA2 haploinsufficiency including severe monocytopenia, B and NK lymphopenia and myelodysplasia but lacking in any known synonymous mutations in the GATA2 gene. KERNA was screened alongside 11 other patients for SVs that could explain the manifestation of GATA2 haploinsufficient phenotypes in the absence of a coding mutation. In KERNA we identified a novel *de novo* tandem duplication of a region (187Kb) comprising the entire GATA2 gene as well as its -110 enhancer region and a deletion of 25kb of the 5’ end of RPN1.

# 2.2 Materials and Methods

Material for KERNA was obtained along with 11 other patients with informed consent through the French GATA2-like Project and assigned to Newcastle Biobank by material transfer agreement (Newcastle and North Tyneside 1 Research Ethics Committee Reference 17/NE/0361) for 12 patients (**Table 2.1**). Extended capture targeted amplicon sequencing was carried out on a 550,059bp region (chr3:127,849,971-128,400,030) that contains the GATA2 gene (chr3:128,198,265-128,212,030) flanked by EEFSEC and RPN1, it's -110 super enhancer and 8 other neighbouring genes (**Figure 1.1**). This work was carried out by the lab of Prof. Matthew Collin.

I received the raw sequencing data as raw Fasta files. These were then processed and aligned the sequences to GRCh37/hg19 assembly using BWA (v.0.7.6.a) (Li and Durbin 2009) and reads were later indexed using Samtools (v.1.3.1) (Li et al. 2009). Because amplicon sequencing was used, duplicates were not removed. Two patients, A2163 and A2744 were previously identified as having non-synonymous mutations within the exons of the GATA2 gene (**Table 2.1**). To ensure that none of the remaining 10 patients contained non-synonymous SNVs the GATA2 gene and its distal enhancer were first screened for SNVs using GATK3 following the recommended best practices for germline variant calling (Van der Auwera et al. 2013).

We first carried out a comprehensive search of methods to detect structural variations in amplicon-based sequencing data. Of those tested, to detect structural variations we opted to use ONCOCNV (v.6.8) (Boeva et al. 2014); a copy number calling algorithm based on read-depth to identify a CNV in KERNA and identify any copy number variations in the rest of the cohort. Although a method designed for tumor vs control CNV detection it is fundamentally designed to tackle the challenge of detection copy number changes in amplicon NGS data. Of the 12 patients, two (KERNA and MURET) were sequenced alongside family members which were used as the controls to generate a baseline model for the CNV analysis. The baseline model is constructed from a principal component analysis (PCA) analysis where each principal component accounts for the levels of variation between the samples. The mother of KERNA (DANMU) was excluded from the control group following the identification of a population CNV and previously suspected GATA2-like phenotypes. Common population copy number variable regions were filtered using the population structural variant reference database provided in the paper by Sudmant et al, (Sudmant et al. 2015). Results of the CNV analysis were then validated and the breakpoints were pinpointed by manual inspection using the Integrative Genomics Viewer (IGV) (Robinson et al. 2011).

| Sample Name | File Name | Sample Details |
|---|---|---|
| DEMSO | WTCHG_406087_015 | MonoMac with PAP; died after BMT |
| MURET | WTCHG_406087_002 | No information |
| DANMU | WTCHG_406087_001 | Mother of MURET; died from aplasia |
| MURSE | WTCHG_406087_010 | Father of MURET; no symptoms |
| MURAL | WTCHG_406087_020 | Sister of MURET; no symptoms |
| BALMA | WTCHG_406087_006 | MonoMac with PAP, peripheric thrombocytopenia and mycobacteria in infancy (father thrombopenic too) |
| KERNA | WTCHG_406087_012 | MonoMac with mycobacteria, and tri8 in BM, BMT in 2016 (april) |
| HELLA | WTCHG_406087_019 | Mother of KERNA; no symptoms |
| KERME | WTCHG_406087_005 | Father of KERNA; no symptoms |
| NAGFL | WTCHG_406087_013 | No information |
| 15MB004476 | WTCHG_406087_023 | MDS, tri8, warts, infections; c.229+13_229+14insGCCins203_229+13; p.? |
| A6981 | WTCHG_354665_002 | MDS and lymphoedema, 59 years old, FL=8986, GATA2 AEI |
| A2604 | WTCHG_354665_004 | Colitis and premalignant lesions in colon, Stage III CIN and VIN HPV, 28 years old, FL=11400, GATA2 AEI |
| A2931 | WTCHG_354665_007 | Developmental delay, MDS with small megakaryocytes, 11 years, FL=7078, GATA2 AEI |
| A7200 | WTCHG_354665_008 | no clinical info, 13 years, FL 3439, no GATA2 AEI |
| A2163 | WTCHG_354665_019 | known GATA2 mutation (3:128200113G>A, c.1192C>T, R398W) |
| A2744 | WTCHG_354665_021 | known GATA2 mutation (3:128204841_128204842insC, c.599_600insG, G200fs) |

***Table 2.1. Targeted sequencing data*** for 12 patients and 5 family members obtained through the French GATA2-like Project and assigned to Newcastle Biobank by material transfer agreement (Newcastle and North Tyneside 1 Research Ethics Committee Reference 17/NE/0361) for 12 patients

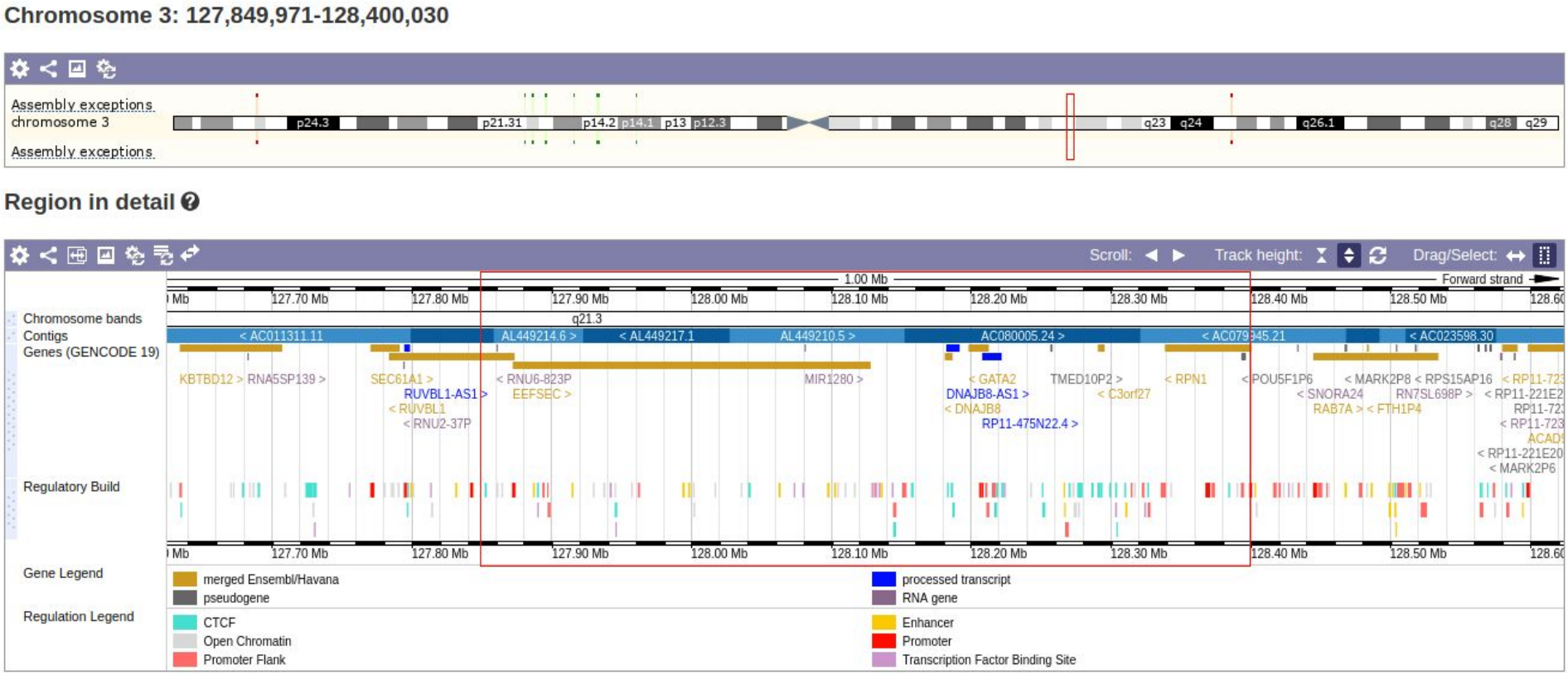



***Figure 2.1. Targeted sequencing region (red box).*** *A 550,059bp region (chr3:127,849,971-128,400,030) that contains the GATA2 gene (chr3:128,198,265-128,212,030) flanked by EEFSEC and RPN1, it's -110 super enhancer and 8 other neighboring genes*

# 2.3 Results

## 2.3.1 Targeted sequencing GATA2 locus of patients with GATA2-like deficiency reveals a new structural alteration

We analysed a cohort of 12 patients presenting with a GATA2 deficient phenotype in which 10 were concurrent with the absence of non-synonymous mutations in the coding sequence of the gene or non-coding mutations in its distal enhancer.

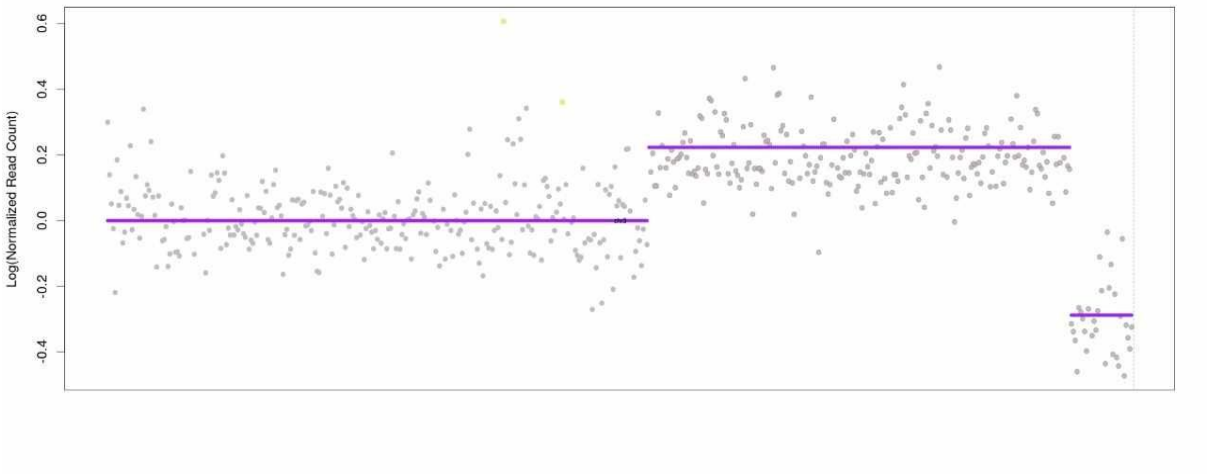
In KERNA we identified a *de novo* tandem duplication of a region (187Kb) comprising the entire GATA2 gene as well as its -110 enhancer region and a deletion of 25kb of the 5' end of RPN1 (**Figure 2.1**). All 12 patients were screened for CNVs. Following read depth analysis, patients KERNA and NAGFL were found to contain CNV's (**Figure 2.2**). This was complemented by work done in parallel suggesting a CNV in KERNA in the capture region and was determined by MLPA targeting the GATA2 gene in which a 1.5x copy-number increase was observed between KERNA (**Figure 2.3a**) and the mother (**Figure 2.3b**). The CNV deletion event was determined to be inherited from the mother. However, this was determined to be a population common CNV region (found in 1.37% of the population) when cross matching against The Database of Genomic Variants (MacDonald et al. 2014). The deletion event was shown to remove an alternative transcription start site for RPN1 at 128,400Kb while it is also associated with CTCF and H3K27ac binding peaks suggesting potential structural changes in the chromatin conformation. The CNV identified in the patient NAGFL was also within the population variable region and suggests that this region is more susceptible to SVs (**Figure 2.2b**).

The exact breakpoint coordinates of the duplication were identified using the IGV genome browser by identifying soft-clipped reads (mismatched reads at the 5' and 3' ends) (GRCh37/hg19 chr3:128190688-128404309) (+/- 1bp) (**Figure 2.4a**). The nature of the duplication was established by analysing the read mapping orientation. Under normal circumstances for paired-end sequencing data the read orientation will always be divergent for 5' to 3' and 3' to 5' reads. When a proportion of the reads (from the duplicated region) are inverted with respect to the reference genome this results in converging read orientations and suggests a tandem duplication (**Figure 2.4b**). Analysis of SNPs (rs2335237 (exon 1) and rs1573858 (exon 2)) were shown to be heterozygous in KERNA and homozygous in the mother with a 2:1 ratio suggesting the duplication was inherited from the mother.

This was later confirmed by the lab of Prof Collin with QPCR analysis of GATA2 expression in PBMC and fibroblasts comparing $\Delta\Delta$Ct (GATA2-GAPDH) in two controls and the patient using three primer pairs for GATA2 exons 1-2, exons 3-4, and the GATA2 antisense transcript (AS1) (**Figure 2.5a**). Strikingly, the expression of GATA2-AS1 was shown to increase in both PBMC and fibroblasts by Q-PCR and RNA-seq (**Figure 2.5b**). GATA2-AS1 is a long non-coding RNA (lncRNA) that has been shown previously to inhibit GATA2 expression through interaction with the GATA1 protein at the GATA2 promoter (L. Zhang et al. 2019). Following duplication the second copy of GATA2 is placed closer to its distal enhancer and may activate the transcription of GATA2-AS1 more freely and in turn, inhibit the expression of all three copies of GATA2 (**Figure 2.5c**).

There are no current reported cases of a germinal copy-number gain of the gene or its regulatory regions associated with a typical GATA2 haploinsufficiency phenotype. The results from KERNA  suggest a novel mechanism by which a GATA2 deficient phenotype can manifest through a *de novo* germline duplication event of  loci containing the GATA2 gene. A manuscript for these findings has been submitted to Blood Advances.

A




B

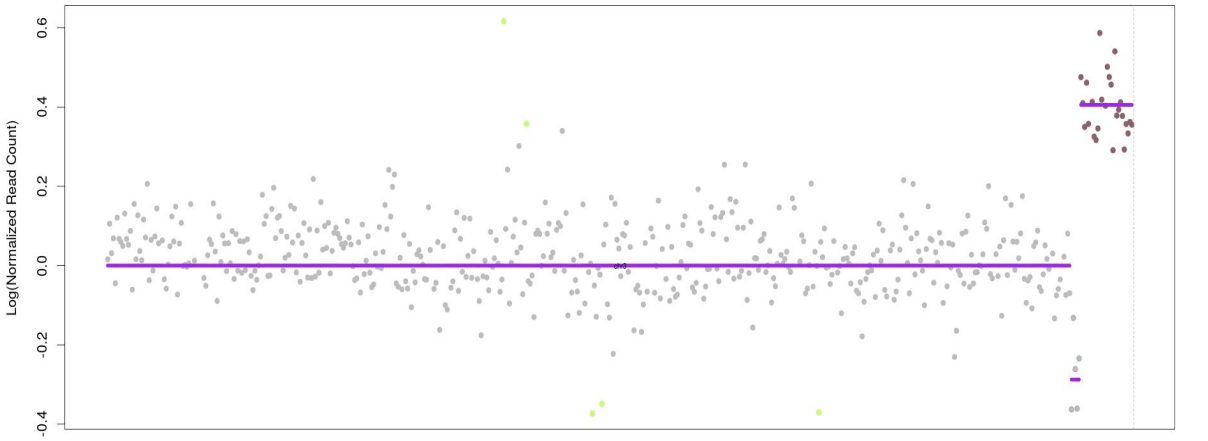
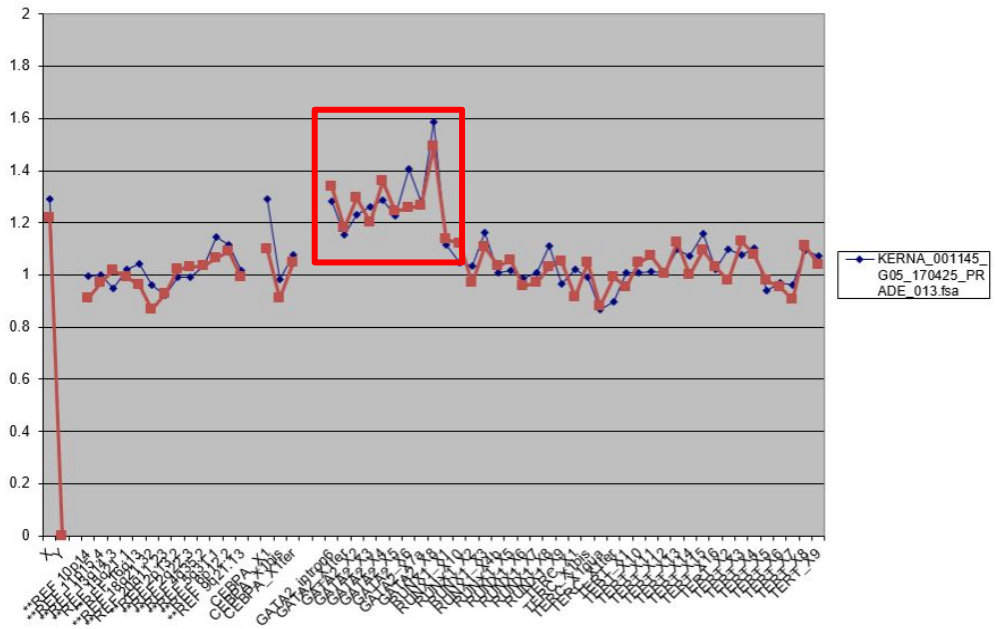



***Figure 2.2. Read-depth analysis*** of KERNA (a) and NAGFL (b). Two patients, KERNA and NAGFL were found to contain duplication and deletion events. Read-depth analysis was carried out using ONCOCNV (v.6.8). MURSE and MURAL used as controls to construct the baseline model.

A




B

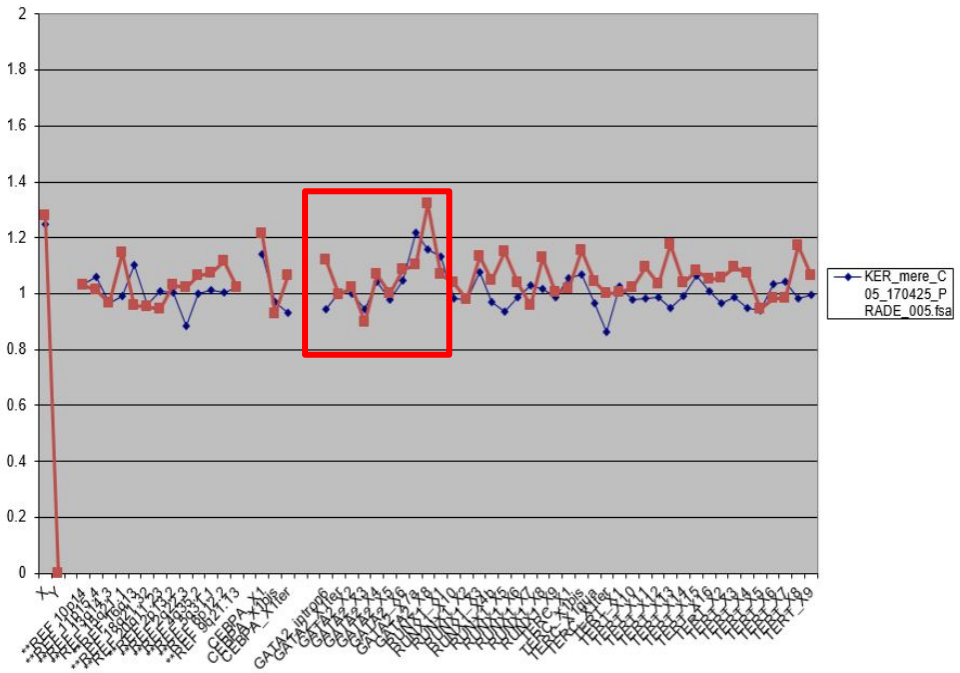



***Figure 2.3. MLPA data for KERNA and DANMU*** a 1.5x copy-number increase of GATA2 was observed between KERNA (**a**) and the mother (**b**)

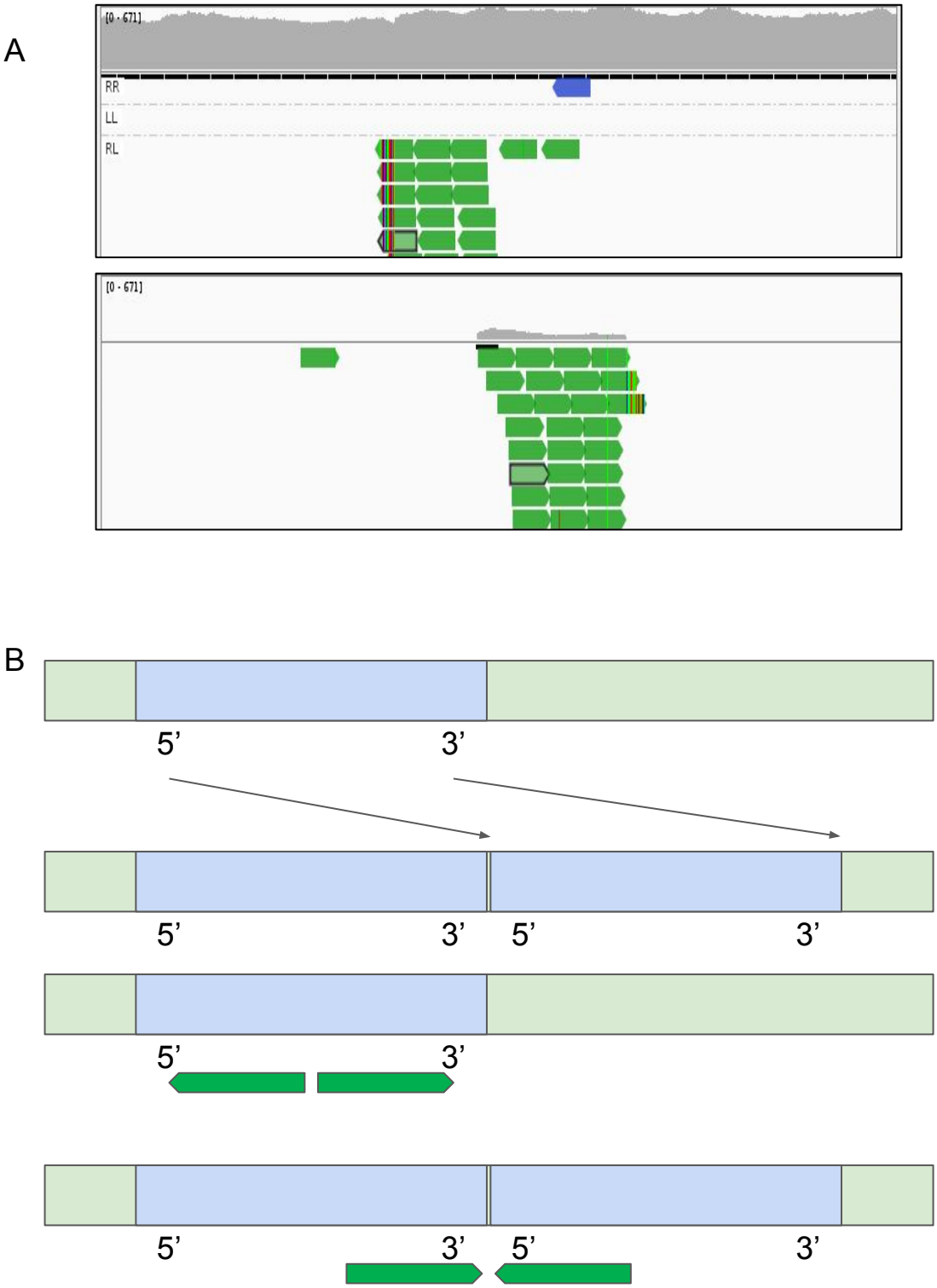



***Figure 2.4.*** **Identification of breakpoints and duplication type. A)** Exact breakpoints were identified using IGV genome browser by identifying soft-clipped reads (GRCh37/hg19 chr3:128190688-128404309) (+/- 1bp). **B)** Normal read orientation always be divergent for 5’ to 3’ and 3’ to 5’ reads. When a proportion of the reads (from the duplicated region) are inverted with respect to the reference genome this results in converging read orientations and indicates a tandem duplication

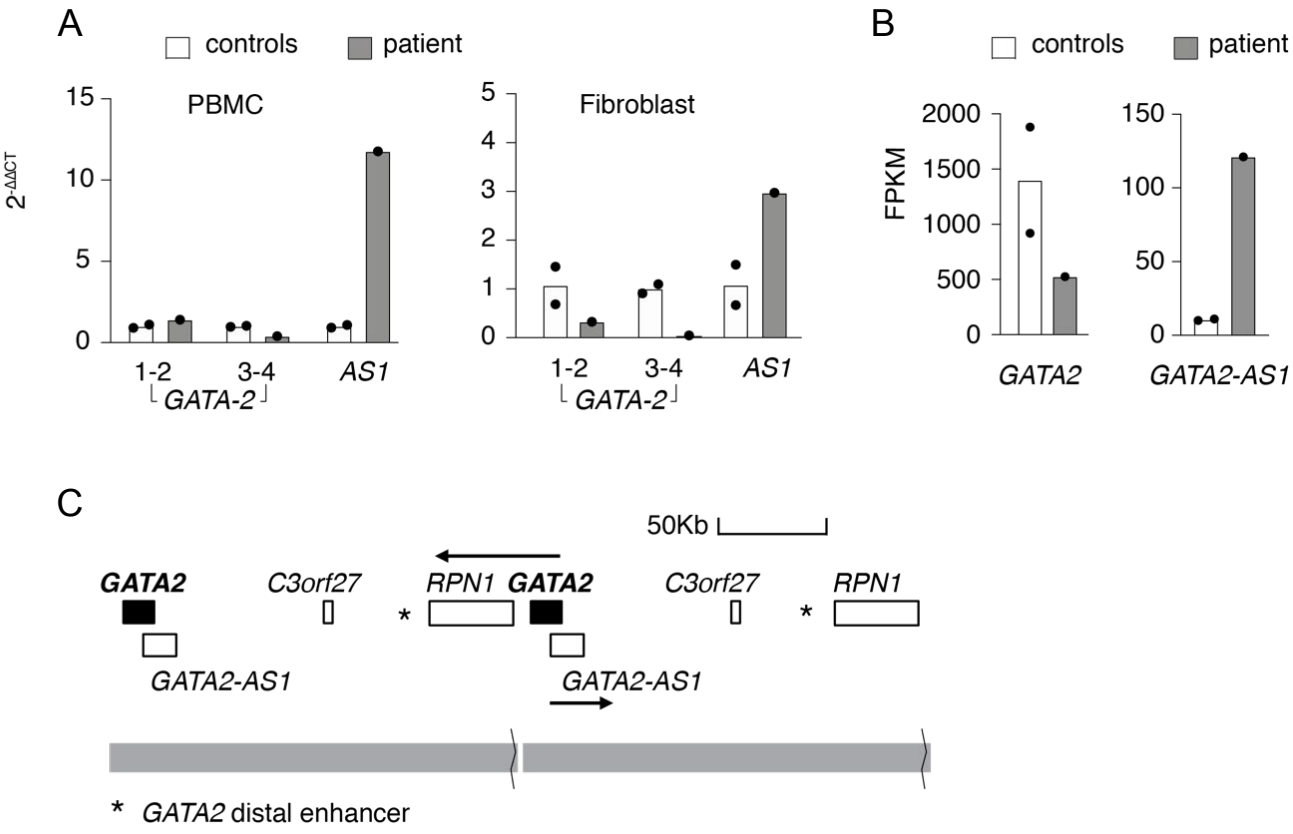



***Figure 2.5. Characterisation of a GATA2 tandem duplication***. A) QPCR analysis of GATA2 expression in PBMC and fibroblast comparing ΔΔCt (GATA2-GAPDH) in two controls (white) and the patient (gray) using three primer pairs for GATA2 exons 1-2, exons 3-4, and the GATA2 anti-sense transcript (AS1). B) The expression of GATA2 and GATA2-AS1 determined by RNA-seq in PBMC for the patient (grey) and controls (white). C) Structure of the tandem duplication with the position of the GATA2 distal enhancer (asterisk)

# 2.4 Discussion

Immune disorders such as GATA2 haploinsufficiency describe the condition of deficient, overactive or even malignant immune systems (Hsu, McReynolds, and Holland 2015). The broad range of phenotypes are driven by the complex nature of the genetic network that underpins our immune systems. Current strategies are well placed to detect direct mutations in coding genes that drive specific phenotypes. This is exemplified by the non-synonymous mutations in the GATA2 gene of our patients A2163 and A2744. Up until recently, most diagnostic tests for pathogenic mutations have focused on protein coding genes as a result of the previously prohibitive costs of WGS compared to WES. Targeted capture of a locus can cast a wider net and include some of the non-coding regions that flank a gene of interest but, as we show it is still limited. Even with the expansion to WGS difficulties still remain in interpreting the functional effects of variants beyond exons. We found no pathogenic SNVs in any of the exons captured in the remaining 10 patients of the cohort. There may be a subset of these phenotypes that can be explained by other mechanisms. Synonymous mutations that affect the stability of RNA, for example, have been previously reported (Kozyra et al. 2020). However, non-coding SNVs and SVs beyond the capture region used in this study may also contribute towards GATA2 haploinsufficiency.

Mutations in the protein coding genes (2% of the genome) currently only account for roughly 25 to 50% of pathogenic variants in inherited mendelian diseases (Beaulieu et al. 2014), (Yang et al. 2013). It remains that up to ~50% of pathogenic mutations occur somewhere in the remaining 98% of the non-coding regions (NCRs) of the genome. In fact, there is also evidence to suggest that even subtle, genome wide perturbations can also drive complex transcriptional patterns (Boyle, Li, and Pritchard 2017). These complex transcriptional patterns and variants across multiple loci have been implicated in complex diseases such as Alzheimer's (Jiang et al. 2017).

In the last decade the cost of WGS has dropped dramatically and has facilitated more widespread adoption in clinical diagnostics. For example the 100,000 Genomes Project aims to establish WGS diagnostics within the UK's National Health System (NHS) which will allow researchers to identify drivers of disease across the whole genome (Turnbull et al. 2018). To do so effectively, these approaches rely on the annotation of the non-coding genome using linear features such as histone modifications, chromatin accessibility and expression quantitative trait loci (eQTL) as well as non-linear features such as 3D chromatin interactions **(see Chapter 1)**. These are then used to annotate variants as pathogenic if they occur in enhancers or TAD boundaries for example, as perturbations in these regions often result in gene expression changes. We expect that many variants are falsely labeled as non-pathogenic due to incomplete characterisation of the non-coding genome. These problems can be addressed by higher quality data relating to linear features and the discovery of new ones. There is also a wealth of 3C data that describes the conformation of chromatin and the interactions between various loci and evidence to suggest that the chromatin conformation is involved in the regulation of gene expression. This work would benefit from case and control 5C study of this region to understand if and how this region is rewired via the duplication event. Currently, there is a limited understanding of how these interactions dictate the expression patterns of genes. To address these problems we sought to predict enhancer loci by analysing the global patterns encoded within the chromatin

architecture. The results of which will facilitate a better understanding of the relationship between enhancers and gene expression that is mediated by the chromatin organisation and ultimately, improve the identification of non-coding variants that contribute towards disease.

# Chapter 3: The identification of enhancer nodes with network centrality measures

---

## 3.1 Introduction

Disease states occur as a result of disrupting the regulation of key genes by mutations at sensitive loci. These loci can contain enhancer sequences, which when perturbed directly affects the expression of a gene. Alternatively, more complex and indirect mechanisms can lead to aberrant gene expression. One example of this is the disruption of the 3D organisation of chromatin by the ablation of TADs. These changes can affect the localisation of enhancers to promoters, resulting in the misexpression of genes (Anne-Laure Valton 2016). Identifying these loci are important when determining the functionality of non-coding variants across the genome. The lack of functional variants found in 9 of the 12 GATA2 patients discussed in **Chapter 2** highlights the importance of this. One solution to identify functional non-coding variants is to identify direct physical interactions between the loci of the variant and gene promoters. This is achieved using 3C data and methods have been developed to identify these promoter-enhancer interactions (Ron et al. 2017). This is achieved by identifying loci that interact with promoters as these are often enriched for enhancer elements (Dekker 2008).

Enhancers are typified by their ability to increase the probability of gene expression through the localisation of RNA pol II and transcription factors and modulation of the chromatin structure. As discussed previously (**See chapter 1: Enhancers**), the study of enhancers is limited by the lack of, or as yet unidentified, a ubiquitous and unifying feature such as the ‘cis-regulatory code’ to identify all enhancers (Yáñez-Cuna, Kvon, and Stark 2013). Therefore, enhancer associated features such as transcription factors, co-activators, RNA pol II, histone marks and DNase accessibility have been used in lieu of a definitive feature. There are numerous publications that have sought to utilise enhancer associated features in order to drive computational predictions of putative enhancers (Chengqi Wang, Zhang, and Zhang 2013) and the genes they regulate (Hariprakash and Ferrari 2019). Despite the apparent success of many of these algorithms, they are hamstrung by our limited understanding of what an enhancer is and how it works. Our predictions are only as good as the data that they are built on. Enhancer features, as useful as they have proved, do not provide absolute and direct evidence for enhancer activity (**See Chapter 1: Histone modifications**). These features can also be observed in other genomic elements. For example, there is a growing ambiguity between promoters and enhancers as evidence suggests they are more alike than previously reported, such elements are termed Epromoters (Medina-Rivera et al. 2018; Andersson and Sandelin 2019). So, while there is progress in the field of enhancer prediction, without further defining what an enhancer is, the specificity of such algorithms will always be limited. Therefore, identifying additional features of enhancers is of utmost importance in order to both predict enhancers and further our understanding of how this class of regulatory element drives cell-type specific gene expression.

One particular genomic feature that has come to light in the last decade and contributes towards gene regulation is the **three dimensional folding of chromatin**. It is known to localise cis regulatory elements (CREs) such as enhancers to gene promoters at large distances across the linear genome to drive cell-type specific gene expression (Schoenfelder and Fraser 2019). This is observed at the global level where enhancers are dynamically organised to work in concert and orchestrate the correct spatio-temporal expression of genes. This organisation is achieved through the folding of the chromatin fibre into hierarchical compartments, domains and enhancer promoter (E-P) loops. Chromatin conformation capture (3C) methods (**See Chapter 1.4**) have emerged as the leading technology to map the structural organisation of chromatin and thus, elucidate the mechanisms that drive cell-type specific gene expression.

However, traditional pairwise representations of 3C data are susceptible to missing both transient and indirect enhancer promoter interactions while some enhancers remain distal to the gene they regulate (W. Song, Sharan, and Ovcharenko 2019). For example transcriptional bursting has been demonstrated enabling the periodic transcription of genes with stronger enhancer sequences shown to result in more frequent bursts (Fukaya, Lim, and Levine 2016). Enhancer activity has also been observed with a reduction in enhancer-promoter proximity following the activation of a gene (Alexander et al. 2019). There is also evidence that enhancers may establish and maintain regulatory circuits by localising and stabilising the contact between a gene and the primary enhancer. In this scenario secondary and tertiary enhancers may interact with the gene less frequently than the primary enhancer, or not at all (W. Song, Sharan, and Ovcharenko 2019). These types of indirect interactions are susceptible to be missed when 3C data is interrogated in a pairwise manner. As a consequence, information such as indirect contacts and cooperative regulatory subnetworks are lost. In addition to this, there is evidence that subtle changes across the genome can influence gene regulatory networks (Boyle, Li, and Pritchard 2017) while accumulation of these changes have been implicated with disease phenotypes (Westra et al. 2013).

To address the analytical problem of identifying indirect associations in networks, computational network analysis methods such as clustering and module finding algorithms have been developed for a wide range of applications (Brohée and van Helden 2006). The unifying aim of these approaches is to identify the modular structures within the networks based on the properties of the nodes and/or edges. However, these methods have been found to be less effective than the simpler guilt-by-association methods when tested across a PPI network in *S.cerevisiae* (J. Song and Singh 2009). However, network theory has re-emerged in recent years as computational resources have become more accessible and performance has increased. Network theory approaches may offer a useful adjunct in identifying intergenic enhancers by modelling chromatin interactions as networks. In these approaches 3C contacts are transformed into networks where 3C restriction fragments are represented as nodes and their interactions as edges. Such networks have been termed chromatin interaction networks (ChINs) (Sandhu et al. 2012).

Representing 3C data as ChINs can alleviate some of the issues discussed by explicitly defining indirect relationships and, as a consequence of this, capture the global relationships between loci across the genome. The use of ChINs has been used to describe all manner of

biological associations, from proteins to histone marks (Juan et al. 2016; Lundberg et al. 2016). Network measures can then be calculated to describe the topological characteristics of particular nodes and identify those that are associated with particular biological features. This approach has been used previously to interrogate the architecture of 3D-genomes in simple organisms such as yeast (Hoang and Bekiranov 2013; Kruse, Sewitz, and Babu 2013) to more complex organisms such as mice (Babaei et al. 2015; Juan et al. 2016)). And, of course, in humans; for example, Sandhu et al constructed a ChIN from RNA polymerase-II mediated interactions describing large scale organisation into chromatin communities with specific function (Sandhu et al. 2012). Pancaldi et al. were able to use ChINs and assortativity measures to demonstrate how specific features of the chromatin are enriched in promoter-promoter contacts vs promoter-non-promoter contacts (Pancaldi et al. 2016). Finally, Thibbodeu and colleagues showed that broad domains and super enhancers in ChINs maintained unique and distinct topological properties that could be used to distinguish broad domains from promoters and super-enhancers from normal enhancers (Thibodeau et al. 2017). The representation and interrogation of 3C data as ChINs can be a powerful medium by which to understand the relationships between the 3D genome and gene regulation by enhancers. These specific patterns of 3D chromatin conformation add yet another layer of complexity to the genome and represent another feature from which enhancers can be identified.

# 3.2 Aims

In this chapter we aimed to understand how both cell-type and 3C methods contribute to variability in the ChIN composition and topology in order to A) understand the effects of using different cell-type and capture type 3C data on the ChIN architecture. B) investigate the use of network centrality measures to identify enhancer nodes. C) Assess the limitations of current enhancer features and improve the labelling of enhancer nodes.

# 3.3 Materials and Methods

## 3.3.1 3C datasets

We identified six datasets for analysis across a range of cell types and 3C viewpoints (**See Methods**). Three 3C data sets derived from mESCs were used and include a single cell Hi-C (mcHi-C) for multiple cells that were combined (Nagano et al. 2017). A Promoter capture Hi-C (PCHi-C) originally generated by Schoenfelder et al. (Schoenfelder et al. 2015) and reprocessed using CHiCAGO by Pancaldi et al (Pancaldi et al. 2016) as well as a DNaseI hypersensitive site capture Hi-C (DNaseI-CHi-C) from Joshi et al (Joshi et al. 2015). Raw data from Nagano et al was aligned and reprocessed previously by Inmaculada Hernandez (Rico Lab) to the mm9 reference genome and contacts were normalised and loops were called using the CHiCAGO algorithm with default settings (Cairns et al. 2016). We then generated a multi-cell Hi-C dataset (mcHi-C) where contacts that existed in at least two cells were kept. PCHi-C data was also downloaded for 17 primary immune cells from the publication of Javierre et al (Javierre et al. 2016). We then filtered this dataset for significant interactions for monocytes, neutrophils and CD4+ T-cells where a score equal or higher than 5 was used as a selection threshold following the recommendation from the CHiCAGO

documentation. These 6 datasets we also selected on the basis that there were relevant datasets to annotate putative enhancers.

## 3.3.2 Enhancer features

### Chromatin States

For mESCs a 20 state chromatin state model was downloaded from the publication of Juan et al (Juan et al. 2016). The state model was generated using ChromHMM (Ernst and Kellis 2012) and determined by a combination of 13 histone marks (H3K20me3, H3K9me3, H2Aub1, H3K27me3, H2AZ, H3K4me3, H3K4me2, H3K9ac, HeK27ac, H3K4me1, H3K36me2, H3K36me3 and H3K79me2), 3 cytosine modifications (5hmC, 5fc and 5mc) and the CTCF insulator protein (**Table 3.1**).

For the primary immune cells (PICs) an 11 state chromatin state model was downloaded from the publication of Carillo-de-Santa-Pau et al (Carrillo-de-Santa-Pau et al. 2017). The state model was also generated using ChromHMM (Ernst and Kellis 2012) and determined by a combination of 6 histone marks (H3K4me1, H3K27ac, H3K4me3, H3K27me3, H3K9me3 and H3K36me3). The chromatin states were then collapsed into 5 functional states transcription (E1 and E2), heterochromatin (E3, E4, E5 and E6), repressed promoter (E7), enhancer (E8, E9 and E10) and promoter (E11) (**Table 3.2**). Note that here we also considered the 10th state of promoters to represent Epromoters thus forming one of the enhancer annotations.

3C fragments were annotated with the chromatin states of 200bp bins using the findOverlaps function from Bioconductors GenomicRanges package (v. 1.38.0). The percentage coverage of each chromatin state for each fragment was then calculated.

| State | Associated Function | Histone Marks |
|---|---|---|
| State 1 (E1) | Transcription | H3K4me1 & H3K79me3 & H3K36me3 |
| State 2 (E2) | Transcription | H3K4me1 & 5hmC & H3K9ac & H3K4me2/3 H3K36me3 |
| State 3 (E3) | Transcription | H3K4me1 & 5hmC & H3K36me3 |
| State 4 (E4) | Transcription | H3K36me3 |
| State 5 (E5) | Transcription | H3K36me3 |
| State 6 (E6) | Heterochromatin | H3K9me3 & H4K20me3 |
| State 7 (E7) | Low Signal | N/A |
| State 8 (E8) | Heterochromatin | 5fC |
| State 9 (E9) | Low Signal | N/A |
| State 10 (E10) | Heterochromatin | 5hmC & 5mC |
| State 11 (E11) | Enhancer | H3K4me1 & 5hmC |
| State 12 (E12) | Enhancer | H3K4me1 &5hmC &H3K4me2/3 |
| State 13 (E13) | Enhancer | H3K4me1 |
| State 14 (E14) | Enhancer | H3K4me1 & H3K27ac & H3K9ac & H3K4me2/3 |
| State 15 (E15) | Active Promoter | H3K4me1/2/3 & H3K27ac & H3K9ac & H3K79me2 & H3K36me3 |
| State 16 (E16) | Active Promoter | H3K4me2/3 & H3K27ac & H3K9ac & H3K79me2 |
| State 17 (E17) | Active Promoter | H3K4me2/3 |
| State 18 (E18) | Bivalent Promoter | H3K4me1/2/3 & H3K9ac &5hmC & H3K27me3 |
| State 19 (E19) | Repressed Promoter | H3K27me3 |
| State 20 (E20) | Insulator | CTCF |

***Table 3.1. Summary table of chromatin states for mouse embryonic stem cells***

| State | Associated Function | Histone Marks |
|---|---|---|
| State 1 (E1) | Weak Transcription | H3K36me3 |
| State 2 (E2) | Transcription | H3K36me3 |
| State 3 (E3) | Heterochromatin | H3K9me3 |
| State 4 (E4) | Low signal | N/A |
| State 5 (E5) | Heterochromatin | H3K27me3 & H3K9me3 |
| State 6 (E6) | Heterochromatin | H3K27me3 |
| State 7 (E7) | Repressed Promoter | H3K4me1 & H3K4me3 & H3K27me3 |
| State 8 (E8) | Enhancer | H3K4me1 |
| State 9 (E9) | Enhancer | H3K4me1 & H3K27ac |
| State 10 (E10) | E-Promoter | H3K4me1 & H3K27ac & H3K4me3 |
| State 11 (E11) | Promoter | H3K27ac & H3K4me3 |

***Table 3.2. Summary table of chromatin states for primary immune cells***

### RNA polymerase II

Three RNA polymerase II variants (s2p, s5p and s7p) were downloaded from Juan et al (Juan et al. 2016).

### P300

Processed P300 ChIP-seq data was downloaded from Juan et al (Juan et al. 2016).

### FANTOM5 CAGE-seq

Human and mouse permissive enhancers as identified by CAGE-seq were downloaded from the FANTOM5 data repository (Noguchi et al. 2017; Andersson et al. 2014).

### Starr-Seq

Starr-seq data for mESCs were downloaded from Peng et al (GSE143544) as a .tsv file. (Peng et al. 2020). Rows that contained an activity higher than 0 for columns padj_SL_rep1 and padj_SL_rep2 were filtered to include enhancer activity detected in metastable ESC's.

### Expression quantitative trait loci (eQTL)

Expression QTL data was downloaded from (L. Chen et al. 2016). eQTL's with a p.value <0.05 were filtered for and used for analysis.

## 3.3.3 Annotation of 3C fragments

3C fragments were downloaded and formatted into an 8 column data frame consisting of the genomic coordinates (chromosome, start, end) for each pair of ligated fragments and an additional two columns with an ID number for each unique fragment. Additional columns including the enhancer features were then appended for each annotation.

Coordinates for the primary transcript of protein coding genes from genome build MGSCv37 (corresponding to UCSC version mm9) were downloaded from Ensembl version 75 using biomart. Gene coordinates from build GRCh37.p13 were downloaded from Ensembl version 75. Protein coding genes according to the Ensembl biotype column were filtered for and the X, Y and mitochondrial chromosomes were removed. Genes were then assigned to 3C fragments based on a minimum 1bp overlap using the findOverlaps function from Bioconductors GenomicRanges package.

## 3.3.4 Genomic distance

The linear genomic distance was calculated for each ligation product by calculating the difference in bases between each. The upstream and downstream fragments were established using the orientation column. The end coordinate of the downstream fragment

was then subtracted from the start coordinate of the upstream fragment and each value was assigned to the distance column.

### 3.3.5 Network Generation

ChINs for each of the cell-types were generated by representing the chromatin fragments and their interactions into nodes and edges. Any given chromatin fragment may interact with multiple other chromatin fragments (**Figure 3.1a**). These chromatin interactions are captured in the PCHi-C assay and are represented as pairs of contacts (**Figure 3.1b**). Each contact is determined by the CHiCAGO protocol with a threshold of 5. The higher the threshold, the lower the probability of calling a false contact (often caused by self relegation of the ligation product). Higher thresholds also result in a more sparse network while lower thresholds result in a more dense network. We then transform the contact data for each cell-type into a ChIN. In each of the ChINs the chromatin fragments are represented as nodes and their interactions as edges and will be referred to as nodes and edges hereafter (**Figure 3.1c**). By transforming the data into network type representation the contacts can be investigated beyond the pairwise manner as represented by traditional 3C contact maps. For example if node A interacts with node B and node B interacts with node C we are able to capture the indirect relationship between nodes A and C (**Figure 3.1c**).

Each ChIN was stored as an igraph object where the annotated fragments were transformed from a dataframe into an unweighted and undirected network (sometimes called graphs) $G = (V, E)$using the graph_from_data_frame function from igraph (Csardi G 2006). The 3C fragments are represented as nodes$V$and their interactions as edges $E$ and each node and edge were labeled with their respective annotations. The ChINs were exported to Cytoscape using the RCy3 library for R (Ono et al. 2015) for visualisation.

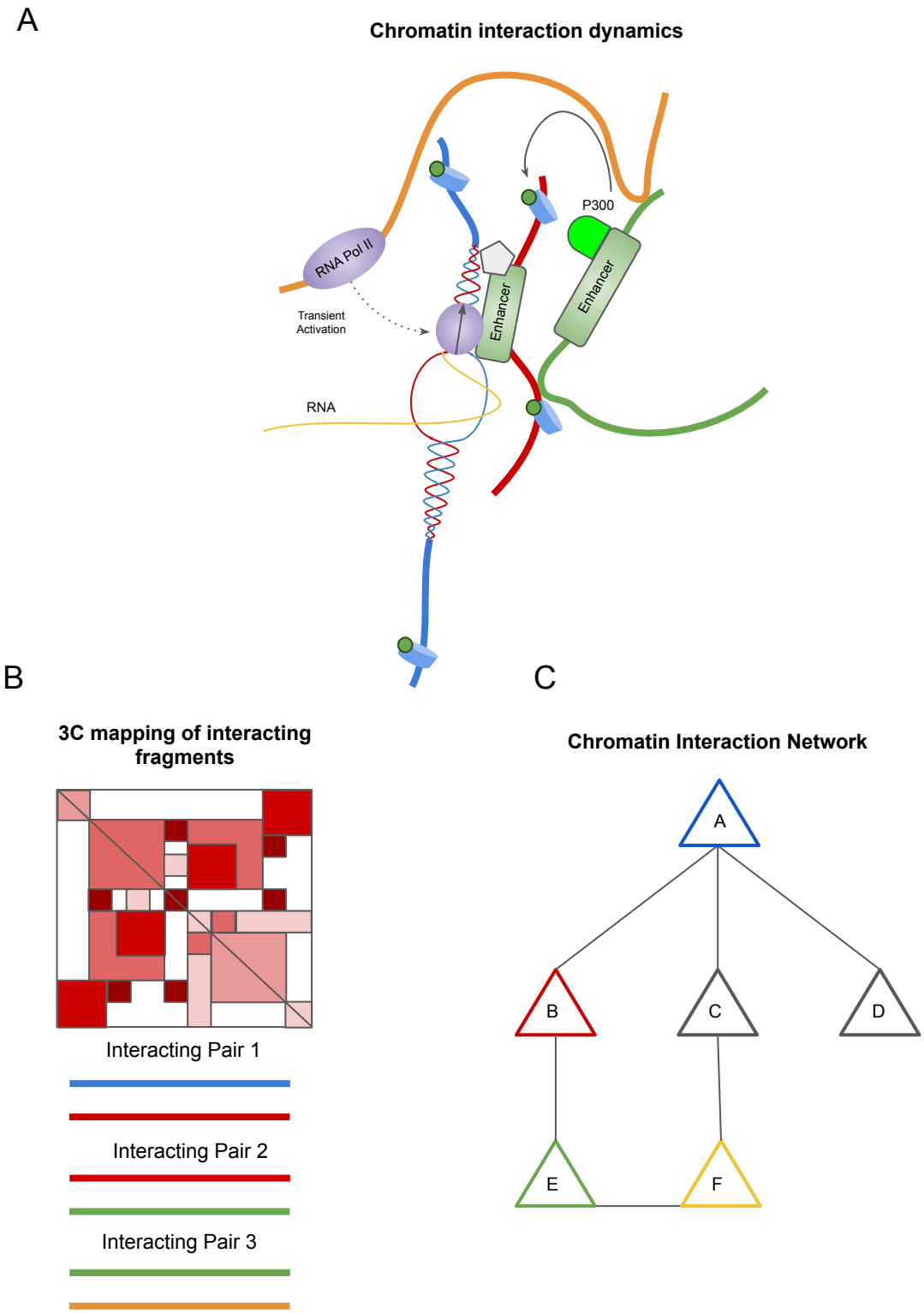



***Figure 3.1. The translation of chromatin contacts into networks*. A)** Enhancers can interact and modulate the expression of a gene by several direct and indirect interactions as shown by the different coloured strands of DNA. **B)** Chromatin interactions, including those with enhancers, can be captured by 3C methods. These are often shown as heatmaps and represent pairwise interactions. The same colours of the interacting pairs correspond with figure A. **C)** Indirect interactions can be examined by representing pairwise data as a network where the colours of the nodes correspond with figures A and B.

## 3.3.6 Network Feature Analysis

### Small world properties

Small world properties were identified for each ChIN based on the Watts Strogatz model by determining if the normalised shortest path $\lambda \approx 1$and clustering coefficient $\gamma > 1$(Watts and Strogatz 1998). The average shortest path $L$ was calculated using the mean_distance function and the clustering coefficient $C$ using the transitivity function for each ChIN. An ensemble of 100 random networks was also generated based on the Erdos-Renyi model (Erdős and Rényi 1964) using the erdos.renyi.game function from igraph to serve as a null-model. Both the average shortest path $Lr$ and the clustering coefficient $Cr$ were calculated for these networks to give an average. $\lambda$ and $\gamma$ were calculated as $L/Lr$ and $C/Cr$, respectively.

### Scale free properties

Scale free networks are defined as networks where the number of nodes with degree k follows a power law k−α. (A. L. Barabasi and Albert 1999). The possession of scale free properties were determined for each CH by calculating the degree distribution using the degree_distribution function in igraph. If the distribution was observed to follow a power law, the CH was determined to possess scale-free-like properties.

### Network measures

All other network measures including the centrality measures were calculated using the appropriate functions provided within the igraph library (version 1.2.5) for R.

## 3.3.7 Precision-recall curves

Precision-recall curves plot, at increasing thresholds, the precision which is the fraction of nodes correctly predicted as enhancers and the recall which is the fraction of the total number of enhancers predicted. Often, a tradeoff occurs between these two metrics whereby increased recall reduces the precision and vice versa. When plotted on the same graph the area under the curve produced by this plot indicates the general classification performance of the model. The precision and recall of enhancer nodes using centrality measures and the corresponding area under the curve (AUPRC) were calculated using the precrec library (version 0.12.7). The baseline values were calculated as the total number of enhancer nodes as a proportion of the total number of nodes in the ChIN.

# 3.4 Results

## 3.4.1 Different flavours of 3C and contact calling algorithms contribute to variability in genome coverage, resolution and contact distances

Each of the three capture types provide a different viewpoint of the true ChIN. We first investigated the properties of the genomic fragments and the contacts between them. The PCHi-C CH captures promoter regions defined by genome annotations of the TSS, these regions are known as baits. The chromatin interacting with the baits are known as other-ends (OE's) and can include chromatin with a wide range of elements including enhancers. Because the baits are designed to capture TSS regions it is important to note that these regions are therefore independent of the cell-type and thus are captured irrespective of whether the TSS is present at active or inactive genes. In contrast, the mESC DNaseI-CHi-C ChIN relies on a functional annotation where probes are designed to target DNase I hypersensitive sites (DHSs), while the mESC mcHi-C ChIN captures nearly all chromatin regions and interactions. In the case of the DNaseI-CHi-C and the mcHi-C ChINs the probes capture non-promoter as well as promoter regions including enhancers. Each of the three capture types provide varying perspectives on the true ChIN.

Differences between 3C methods result in the enrichment of different chromatin regions. This can affect the total number of chromatin interactions captured which affects the completeness of the ChINs. Each ChIN contains information about the size of the DNA fragments, in which combinations they interact and at which distances they are able to do so. Each of these properties are a potential source of variability for the properties of the table of contacts and are introduced by the cell-type and the 3C capture type. The *in silico* pipelines used to process the raw data should not act as a variable as all were processed using CHiCAGO. This can result in the enrichment of specific regions of the genome, such as with PCHi-C enrichment at promoters and varying resolutions that affect the size of the genomic fragments and the contact distance. Here we examine the variability that exists between the genome coverage, fragment resolution and contact differences. In the six data sets used we observe a number of differences that can be attributed to the sources of variation previously discussed.

Differences can be seen between the immune cells; the number of fragments and interactions vary between approximately 10,000 fragments and 50,000 interactions. When compared to the changes seen in the mESCs we observe a difference of approximately 200,000 fragments and 5.3 million interactions from the PCHi-C and the mcHi-C data sets. This is a direct result of using an enrichment protocol for promoters vs the global enrichment of Hi-C. The largest differences appear when different capture types are used. This is driven by the enrichment of specific predefined chromatin regions and the technical aspects of each type of experiment. For example, the distribution of fragment sizes, that is the length of the linear DNA in base pairs, remains consistent between the three immune cell types when using the same 3C assay (PCHi-C) (**Figure 3.2a.**) with an average size between 4kb and 5kb where each experiment uses the HindIII endonuclease. In contrast, the DNaseI, PCHi-C and mcHi-C assays across the mESC dataset results in distinct fragment size distributions

(**Figure 3.3a.**). In particular the mcHi-C dataset yields a median fragment size of 1kb compared to DNaseI 3.3kb and PCHi-C 5.2kb. This is largely driven by the shorter average length achieved by the DpnII endonuclease used in the DNaseI-CHi-C dataset as opposed to the HindIII endonuclease used in the PCHi-C dataset. In the case of the mcHi-C dataset the bins are artificially extended from the capture to account for the low read depth which is a common limitation of full Hi-C.

Genome coverage shows the the amount of the genome that is captured by the 3C assay. The total coverage of the mESC genome achieved by the DNaseI capture is ~22% (12% covered by the baits), approximately double that of the PCHi-C with ~11% (4.9% covered by the baits) and is consistent across all 19 autosomes (**Figure 3.4**). There is an increase in depth at PCHi-C (yellow) regions that are largely mirrored by the DNaseI regions (blue) shown in **Figure 3.5**. Here, 28.66% of DNaseI fragments overlap with PCHi-C fragments while 55.77% of PCHi-C fragments overlap with DNaseI fragments. Both the DNaseI and PCHi-C overlap with the mcHi-C dataset in mESC's (grey) by >99.5% where the mcHi-C dataset captures ~95% of the total mESC genome. It should be noted that the mcHi-C dataset coverage is based on the artificial reduction of the resolution due to a low sequencing depth. Full Hi-C methods are forced to artificially reduce the resolution of the fragments although computational algorithms are now developed to estimate the size of the fragments to increase the resolution (Cameron, Dostie, and Blanchette 2020). The areas with reduced or no coverage for any of the datasets is likely due to chromatin inaccessible regions as well as low mappability regions.

The final consideration that must be made is related to the contact distances. We do not observe any significant changes between the contact distances between the three primary immune cell-types. The distribution of interaction distances remains very similar between monocytes, neutrophils and T-cells (**Figure 3.2b**). While the interaction distances are varied between the different flavours of 3C in the mESCs (**Figure 3.3b**). The PCHi-C data set yielded a slightly higher average distance than both the DNase and mcHi-C data sets. This is in part a consequence of promoters establishing long range contacts with regulatory regions (Schoenfelder and Fraser 2019). This is particularly prominent in the mcHi-C data set where intermittent spikes can be seen separated by a distance of 1kb (**Figure 3.3b.**). This is an entirely artificial artefact of the processing pipeline. Each segment of the genome is divided into 1kb bins and no two adjacent bins can be called as an interacting pair. This results in the gapped output on the density plot where the artefact is more pronounced in fragments interacting over shorter distances.

Each of these cases highlight some important considerations when building a ChIN for the analysis of enhancers. The large fragment sizes of mcHi-C are part of the tradeoff for higher coverage. Conversely, the fragments produced by the DNaseI and PCHi-C are increasingly smaller with the caveat of capturing fewer interactions. This is related to the resolution of the fragments for each experiment as higher resolution fragments are more suited to specific small scale interactions such as between promoters and enhancers. Conversely, the study of large chromatin structures are more suited to lower resolution fragments to increase the coverage (**See Chapter 1: Chromatin conformation capture**).

This tradeoff needs to be carefully considered depending on the type of analysis required. Comparing PCHi-C vs DNaseHi-C vs mcHi-C the increased resolution of PCHi-C and

DNaseHi-C may be beneficial in identifying more specific regions of the genome that may harbour enhancers. However, the lower coverage of higher resolution methods mean that fewer regions and interactions are captured and therefore are less representative of the true topology of the chromatin architecture and networks we use to model it. This will ultimately influence the results of any network theory approaches used to identify enhancer nodes. Our limited knowledge of where enhancers are located and how they regulate the expression of genes means that the mcHi-C data set may be beneficial in the context of capturing a wider sample of the genome meaning more enhancers regions are likely to be captured while it would also produce a more representative snapshot of the ChINs topology. However, with the knowledge that PCHi-C and DNaseI capture methods enrich for more transcriptionally active regions of the genome they are likely to capture most enhancer sequences and may be considered superior datasets to use for the identification of enhancers given their higher resolution.

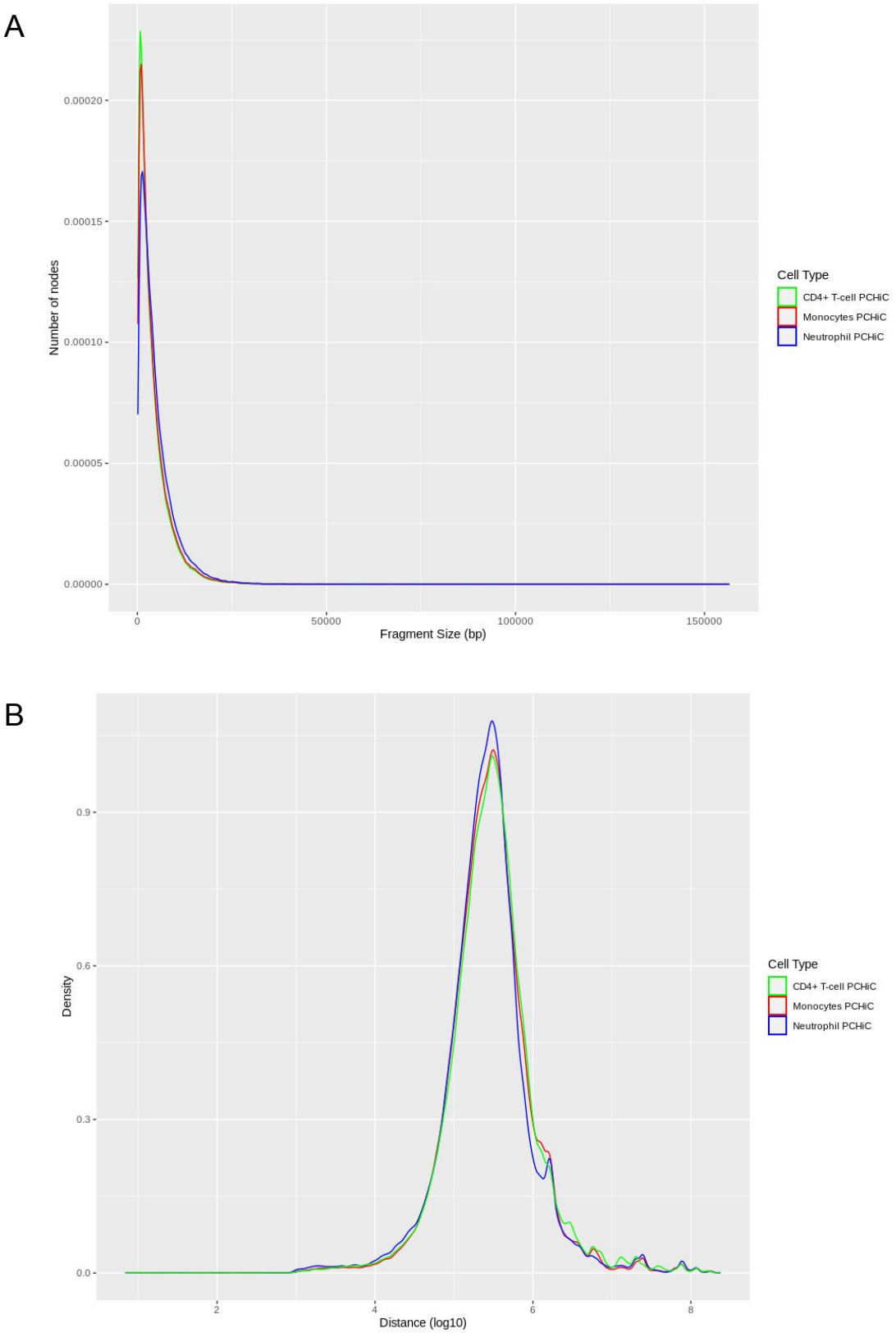



***Figure 3.2. Density plots of 3C fragment size and interaction distance for three primary immune cells.*** A) The fragment size in bp for monocytes (red), neutrophils (blue) and CD4+ T-Cells (green). B) Distribution of interaction distances log10(bp).

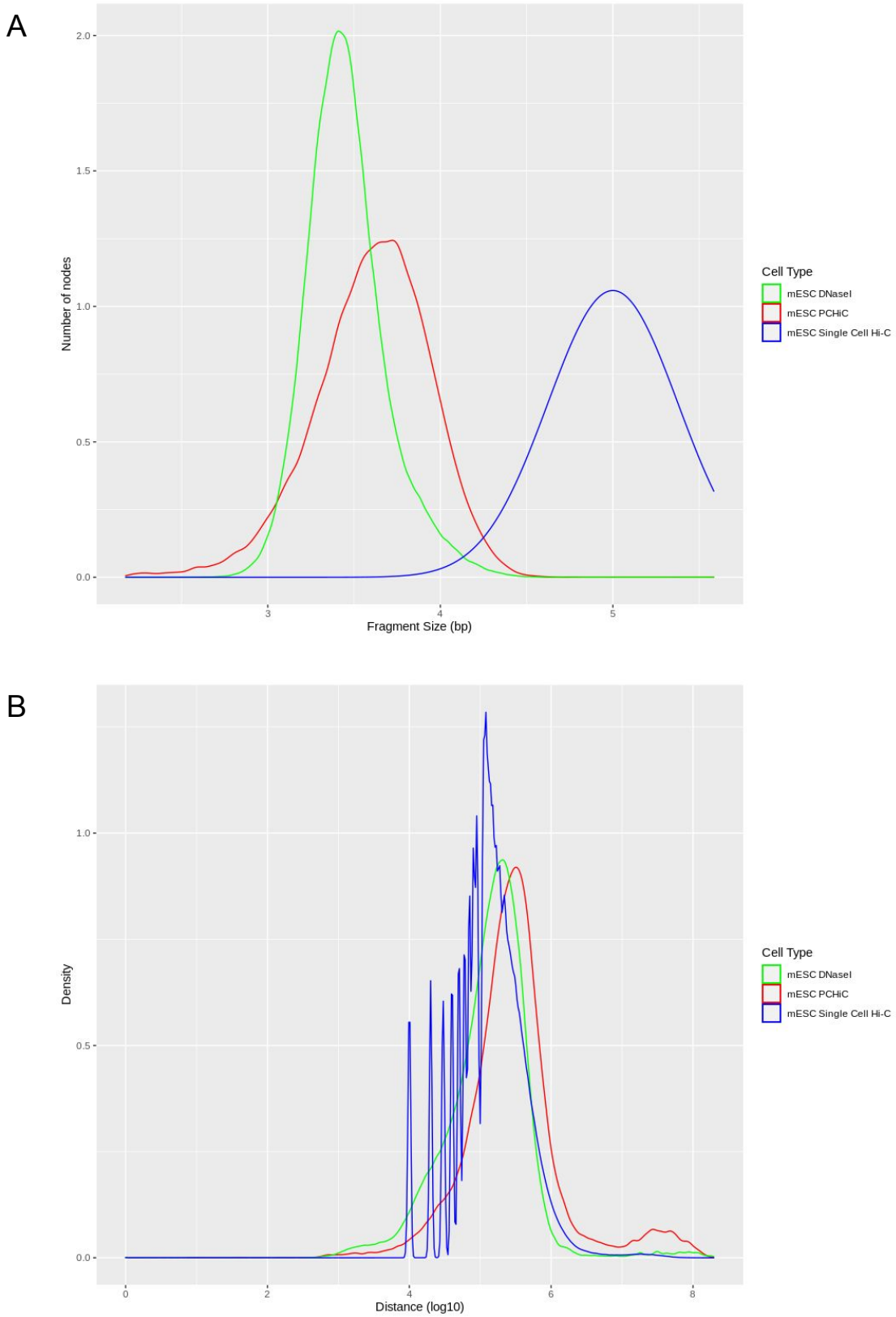



***Figure 3.3. Density plots of 3C fragment size and interaction distance for three capture types in mESCs.*** A) The fragment size in bp for PCHi-C (red), single cell Hi-C (blue) and DNaseI (green). B) Distribution of interaction distances log10(bp).

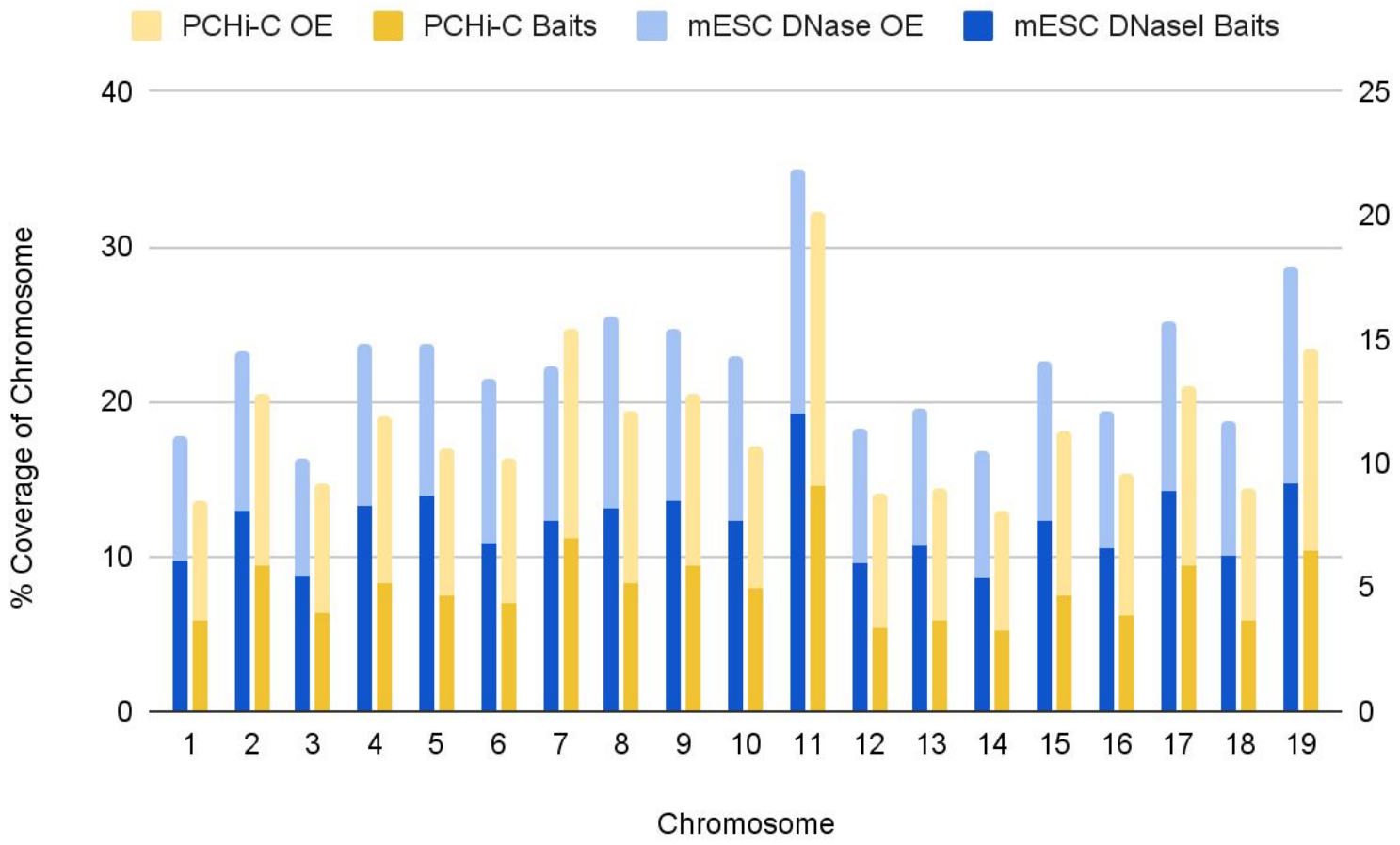



***Figure 3.4. Percentage coverage of each chromosome using DNaseI and promoter capture Hi-C in mESCs***. The percentage of the chromosome captured by the baits (dark yellow and dark blue) and the percentage of the chromosome captured by the interacting other ends (light yellow and light blue).

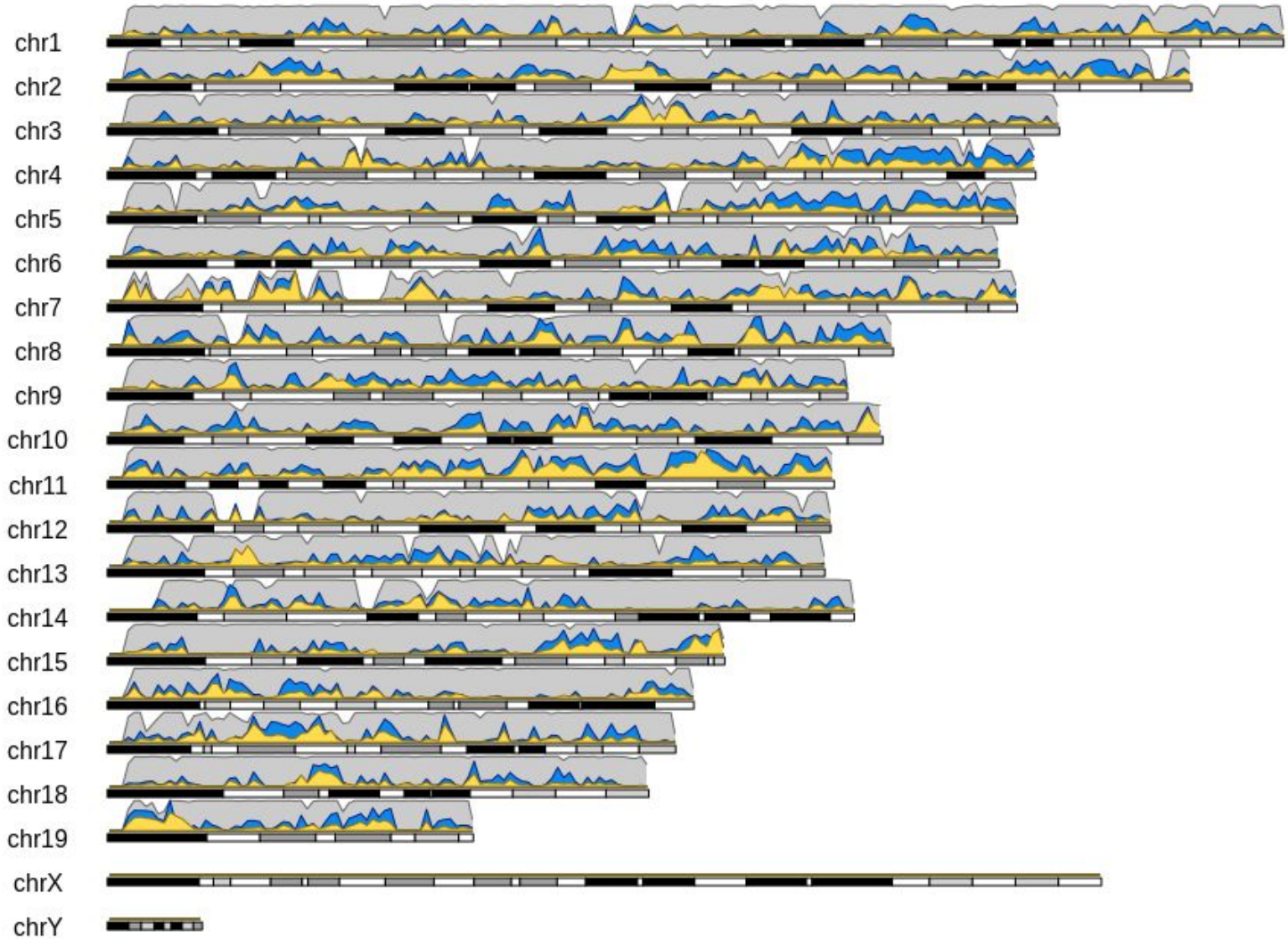



***Figure 3.5. Chromosome ideogram*** plotted with the coverage density across all chromosomes. Densities for promoter capture Hi-C (yellow), DNaseI capture Hi-C (blue) and single cell Hi-C (grey)

## 3.4.2 How do the different capture types affect the network topology?

There different types of 3C capture methods can affect the resolution, the contact ranges and most importantly the coverage of the genome. We then wanted to ask how these variances may affect the topology of our ChINs and if any differences are found, how this may bias different types of network analysis. Here we look at the number of interactions and how these vary between the cell-types of the PIC's and the different capture types in the mESCs. However, the large size of the mESC derived mcHi-C ChIN required a large memory allocation to analyse the features of the network and for this reason we decided against proceeding with the analysis of this particular network. One of the simplest metrics used to understand the topology of a network is the degree centrality. The degree centrality can be used to rank all of the nodes in the network according to the number of edges that each node maintains. Centrality measures such as this can be used to identify nodes of interest including promoters and enhancers if they maintain distinct connectivity patterns within the networks (**see 3.4.5 and 3.4.8**). It is therefore important to understand if and how these properties change between cell-types and 3C methods. We first calculated the average degree for each network. Across the PICs the mean degree deviates by 0.55 edges between the monocytes (3.56) and the CD4+ T-Cells (4.11) with the neutrophils almost identical to the monocytes (3.62). In the case of the mESC ChINs there is a dramatic rise in the mean degree from 2.61 in the PCHi-C ChIN to 9.69 in the DNase ChIN.

Taking a closer look at the node degrees in both the mESC DNaseI and PCHi-C ChINs both networks also contain a high proportion of leaf nodes, those with a degree of one. These can either form isolated graphs made up of two nodes, which are detrimental to the connectivity of the network. Alternatively, they can form the nodes on the edge of the network. We find that leaf nodes are more frequent in the DNaseI ChIN where they make up 34% compared to 19% of the PCHi-C ChIN. Each network also contains nodes with high degree centrality scores. The PCHi-C ChIN contains 9 (0.02%) nodes that exceed a degree of 50, on the other hand the DNaseI ChIN contains 6903 (4.24%). In fact, the median and the mean degree in the DNaseI ChIN are three times higher than those in the PCHi-C ChIN (**Table 3.3**). These results would indicate that the DNaseI capture Hi-C identifies regions of high connectivity not captured by the promoter capture Hi-C. Of the chromatin regions represented in both networks the DNaseI ChIN shares only a ~29% overlap with the PCHi-C ChIN (**Figure 3.5**). It may also suggest that the DNaseI capture Hi-C method is more sensitive than the promoter capture Hi-C method in identifying interactions of the same region. For example, we know that ~55% of the chromatin regions represented in the PCHi-C ChIN are also found in the DNaseI ChIN (**Figure 3.5**). Anecdotally, the node with the highest degree in the DNaseI ChIN has 880 connections. This node is also represented in the PCHi-C ChIN where the node has a degree of 4 and the size of the 3C fragment is 39% smaller. We did not however check the extent to which this occurs.

Perhaps more interesting than the relative changes in degree is the shared pattern of non-uniform distribution of the degree as the median degree in both networks are three times smaller than the 3rd quartile (**Table 3.4**). These results would suggest that there are a small proportion of nodes within the network that are of high degree and the frequency of nodes increase as the degree gets smaller and that this pattern is more pronounced in the

DNaseI ChIN. These properties are suggestive of scale-freeness which is explored in further detail in **3.4.4**.

These results show that the PCHi-C derived networks are less well connected than the DNaseI derived network while the results of **3.4.1** show that the PCHi-C captures less of the total genome. Together these results would indicate that the DNaseI ChIN is less representative of the true mESC ChIN than the DNaseI ChIN. If and how these differences impact the prediction of enhancer nodes is investigated in the following results. .

| | Total number of nodes | Total number of edges | Mean interactions |
|---|---|---|---|
| Monocytes (PCHi-C) | 96,442 | 171,430 | 3.56 |
| Neutrophils (PCHi-C) | 81,242 | 147,195 | 3.62 |
| CD4+ T-Cells (PCHi-C) | 106,848 | 219,439 | 4.11 |
| mESC (DNaseI) | 162,615 | 791,903 | 9.69 |
| mESC (PCHi-C) | 53,920 | 72,221 | 2.61 |
| mESC (scHi-C) | 242,974 | 5,391,054 | - |

***Table 3.3. Summary table of the number of nodes, edges and the average degree centrality*** for the promoter capture Hi-C ChINs of monocytes neutrophils and T-cells and the DNaseI capture Hi-C, promoter capture Hi-C and single cell Hi-C ChINs in mESCs.

| | mESC DNaseI | mESC PCHi-C |
|---|---|---|
| 1st Quartile | 1 | 1 |
| Median | 3 | 1 |
| Mean | 9.69 | 2.61 |
| 3rd Quartile | 9 | 3 |
| Maximum | 880 | 78 |

***Table 3.4. Summary statistics of the degree centrality for the DNaseI capture Hi-C and promoter capture Hi-C ChINs in mESCs***

## 3.4.3 DNaseI capture derived networks are more connected than PCHi-C derived networks

In the previous analysis (**see 3.4.2**) we identified varying proportions of isolated and leaf nodes in the PCHi-C and DNaseI ChINs. We therefore analysed the connectivity of all five networks. The connectivity of the network can be measured in several ways. Since both the PCHi-C and DNase networks from mESCs contain a proportion of isolated and leaf nodes, we first looked at the number of connected components; this is the number of isolated subgraphs, known as components, in the network. These can vary in size from the very large down to two nodes where both are considered leaf nodes. The number of nodes in the largest connected component of the DNaseI ChIN is ~66% while for the mESC DNaseI ChIN almost all of the nodes (~98%) are found in the largest connected component. Comparing the mESC PCHi-C and DNaseI ChIN we can identify reasons as to why this occurs. There is evidence that mammalian genomes such as humans and mice are organised into distinct topological regions including chromosomal regions (**See Chapter 1: Chromatin architecture**) and these regions are connected by long range intrachromosomal interactions. A loss of these interactions may explain the loss of connectivity across the DNaseI ChIN compared to the DNaseI-CHi-C network. These effects can also be observed in the PIC networks.  A visual example of the largest 16 connected components for the monocyte DNaseI ChIN can be seen in **Figure 3.6**. Each colour represents a unique chromosome. We can see that the largest connected component contains a mix of different chromosomes. All of the other 15 subnetworks show distinct chromosome regions supporting the idea that the interchromosomal interactions are integral for a fully connected network. We then compared the number of intrachromosomal interactions between the mESC DNaseI and DNaseI ChINs (**Table 3.5**). Here we see that the DNaseI ChIN contains roughly 4 times the amount of intrachromosomal interactions. The reduction in long range interactions are a product of the specialised nature of the promoter capture Hi-C experiment enriching for promoter contacts which appear not to interact frequently, if at all in a trans-chromosomal manner (Johanson et al. 2018).  This explains why the nodes of the DNaseI ChINs are split across many connected components. Although the differences between the 3C capture types were our primary focus, we also analysed the connectivity of the PIC networks. All of the PCHiC derived networks form a large number of connected components that are disconnected from one another (**Table 3.5**). For the PIC networks, the total number of nodes that form part of the largest connected component is between ~32% and ~52%.

The use of PCHi-C datasets to generate our networks results in a disconnected network. Although the DNaseI capture derived networks are also disconnected, the number of connected components are small and the majority of nodes are in the largest connected component. This would lend the DNaseI derived network to be better suited for network analysis although there are some measures that have been optimised for disconnected networks such as the harmonic centrality in lieu of the closeness centrality (**see 3.4.5 and 3.4.7**). In any case, understanding the connectivity of the network is an important consideration for any future network analysis.

| | Interchromasomal interactions | Intrachromasomal interactions | Connected components | Percentage of nodes in the largest connected component | Density |
|---|---|---|---|---|---|
| Monocytes (PCHi-C) | 170,329 | 1,101 | 1938 | 34.71% | 0.000037 |
| Neutrophils (PCHi-C) | 146,146 | 1,049 | 2136 | 32.26% | 0.000045 |
| CD4+ T-Cells (PCHi-C) | 218,377 | 1,062 | 1375 | 51.63% | 0.000038 |
| mESC (DNaseI) | 782,466 | 9,457 | 776 | 98.12% | 0.000060 |
| mESC (PCHi-C) | 69,847 | 2,374 | 4073 | 66.23% | 0.000048 |

***Table 3.5. Summary statistics for the networks*** of the promoter capture Hi-C ChINs of monocytes, neutrophils and T-cells and the promoter capture Hi-C and DNaseI capture Hi-C ChINs of mESCs

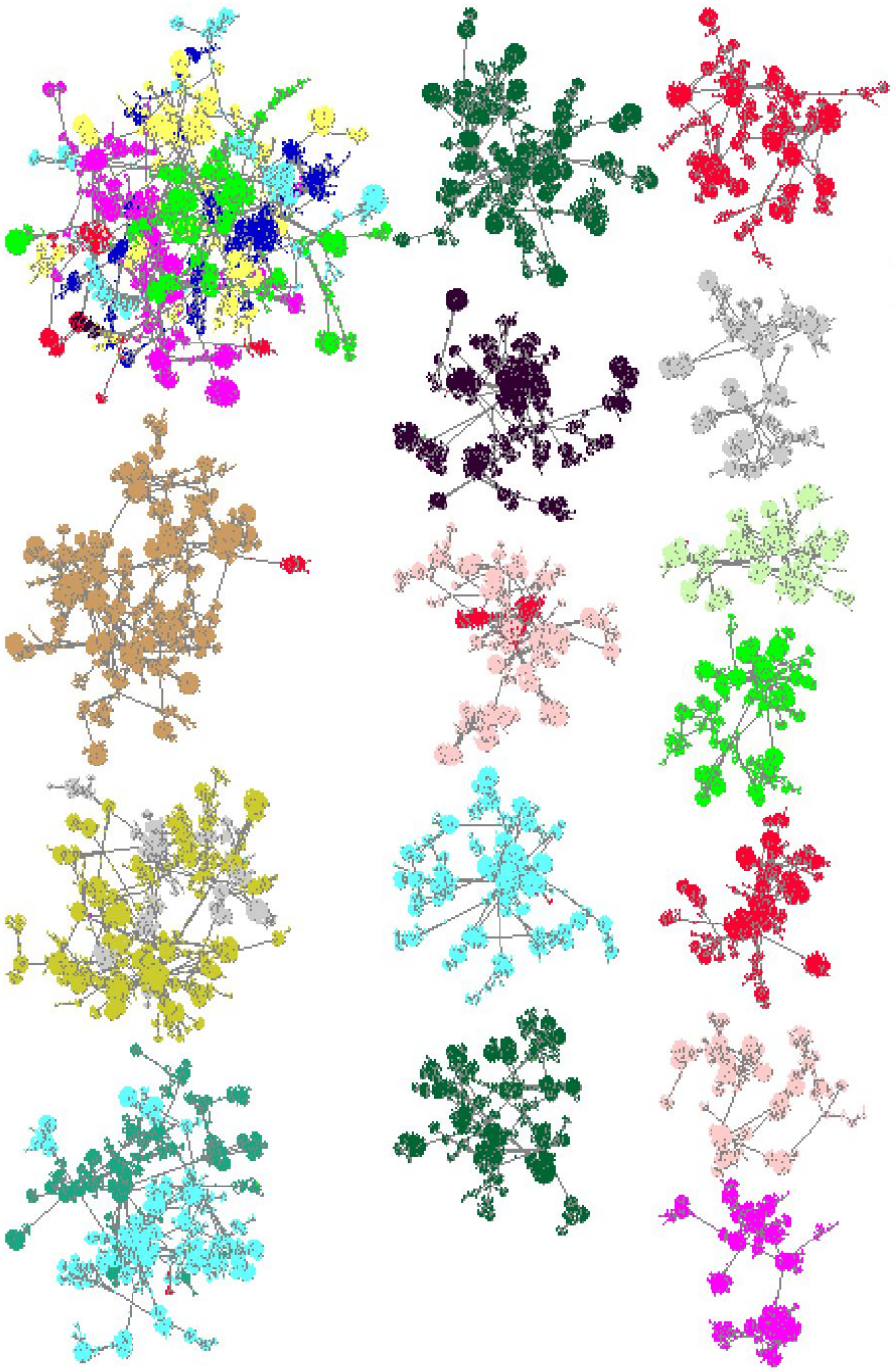


***Figure 3.6. Connected components of the monocyte promoter capture Hi-C ChIN.*** The largest 16 connected components for the monocyte PCHi-C derived ChIN. Each colour represents a different chromosome.

## 3.4.4 All of the networks have small world and scale free properties

We have shown that the mESC DNaseI derived network is better connected than the mESC DNaseI ChIN. To a lesser extent, we also observe small differences in connectivity between the PIC networks. Changes in connectivity can be realised by genuine biological differences as likely shown by the differences observed between the PICs. They can also arise as a result of experimental differences as shown by the differences between the mESC networks. To understand if these changes resulted in fundamental changes to the networks we further investigated and charachterised the global architecture of the networks.

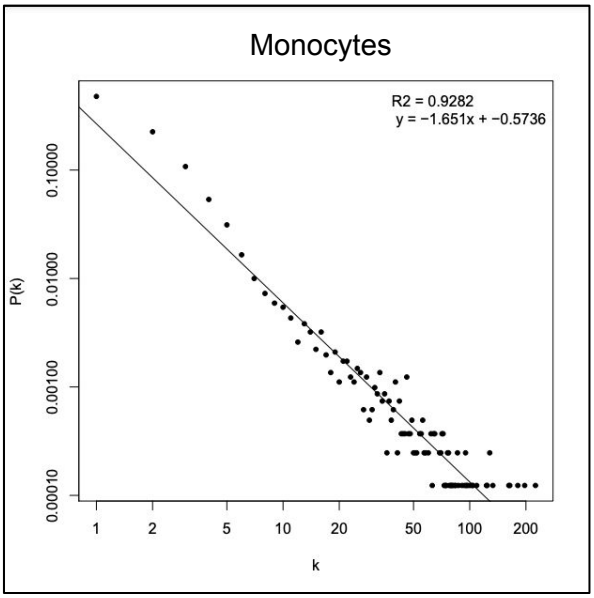
Across all of our networks we identified small-world and scale-free-like properties (**See Methods: Network analysis**). The scale-free properties of each of the five networks are a common feature of complex networks and contain several properties that determine how information is communicated across them. Scale-free networks are defined by the degree distribution following a power law where the probability that a node has the degree $k$ is given by: $P(k) \sim k^{-\gamma}$. Where the exponent $\gamma$ is between 2 and 3, the network is often considered to be scale-free (A.-L. Barabasi and Albert 1999). This is not always the case as shown by the Reactome and BioGRID networks (**Table 3.6**). Because of this ambiguity, we compared the degree distribution of the largest connected components of each ChIN with the degree distributions of networks derived from Reactome and BioGrid (**see methods: 3.3.6 Network Feature Analysis: Scale-free properties**). We show that the five networks follow a power law distribution ($R^2$ = 0.8959 to 0.9412) as both the Reactome ($R^2$ = 0.7096) and BioGRID ($R^2$ = 0.8843) networks do. We also observe, with the exception of the mESC DNaseI and mESC PCHi-C ChINs, that $\gamma$ is lower than 2 (**Table 3.6 & Figure 3.7**). Although there is evidence to suggest that $\gamma$ can sometimes be found to lie either side of this scale (Choromański, Matuszak, and Miękisz 2013) and both metabolic networks such as Reactome and PPI networks such as BioGRID have been shown to maintain scale-free-like properties (Jeong et al. 2000; Barabási and Oltvai 2004). To further characterise the networks, we looked at the percentage of hub nodes. Hubs are nodes that greatly exceed the average degree and are expected in scale-free networks. We observe hubs in each of the five networks,where between 17.5% and 23% of nodes can be classified as hubs when the degree of a node exceeds the average degree of all other nodes in the network. While these networks follow some of the general principles to be defined as scale-free they are by no means definitive. The mESC DNaseI ChIN for example deviates from the power law distribution and has the lowest $R^2$ value of the ChINs but is still higher than the Reactome and BioGRID networks which have previously been described as having scale-free properties. Other tests such as targeted attacks to test network robustness were not carried out to further characterise the scale-free properties (Ercal and Matta 2013).

Scale-free properties are also associated with small world properties. The small-world properties define a network in which the number of ‘steps’ to another node is small. This is charachterised by a high clustering coefficient $\gamma > 1$ and an average shortest path $\lambda \approx 1$. The small-world properties were calculated for each of the networks (**See Methods: Small world properties**). In each of the networks the shortest path was calculated to be approximately 1 and the clustering coefficient more than 1. Small-world networks also contain an abundance of hubs, as previously described, that mediate the short paths.

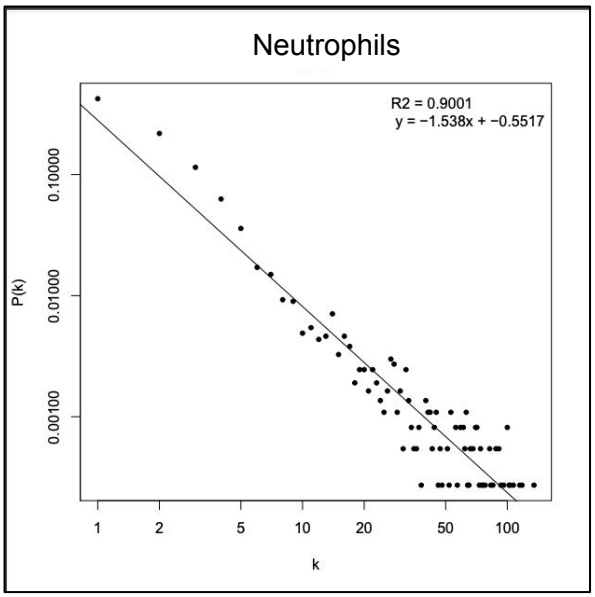
Together these properties facilitate quick and efficient information transfer across the network.

All five networks present properties that are consistent between cell types and more importantly between the PCHi-C and DNaseI capture methods. These metrics show that the broad underlying characteristics of the network are not significantly altered between either cell-types or between capture methods despite the lower connectivity of the DNaseI ChINs.

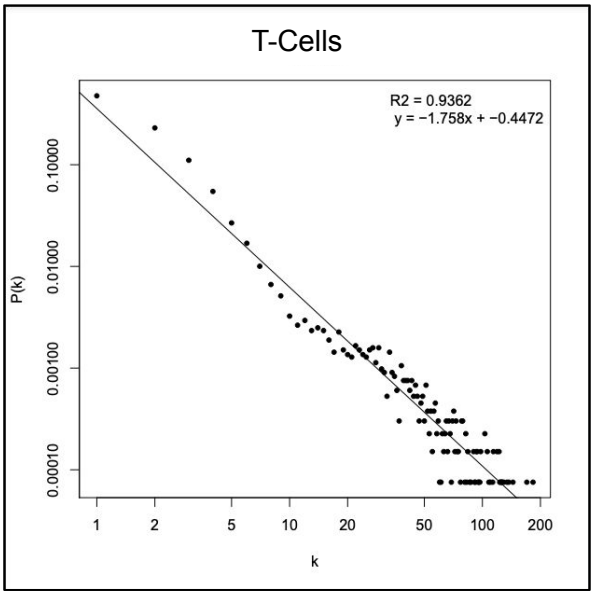
A


Monocytes
R2 = 0.9282
y = −1.651x + −0.5736
P(k)
0.10000
0.01000
0.00100
0.00010
1
2
5
10
20
50
100
200
k

B


Neutrophils
R2 = 0.9001
y = −1.538x + −0.5517
P(k)
0.10000
0.01000
0.00100
1
2
5
10
20
50
100
k

C


T-Cells
R2 = 0.9362
y = −1.758x + −0.4472
P(k)
0.10000
0.01000
0.00100
0.00010
1
2
5
10
20
50
100
200
k

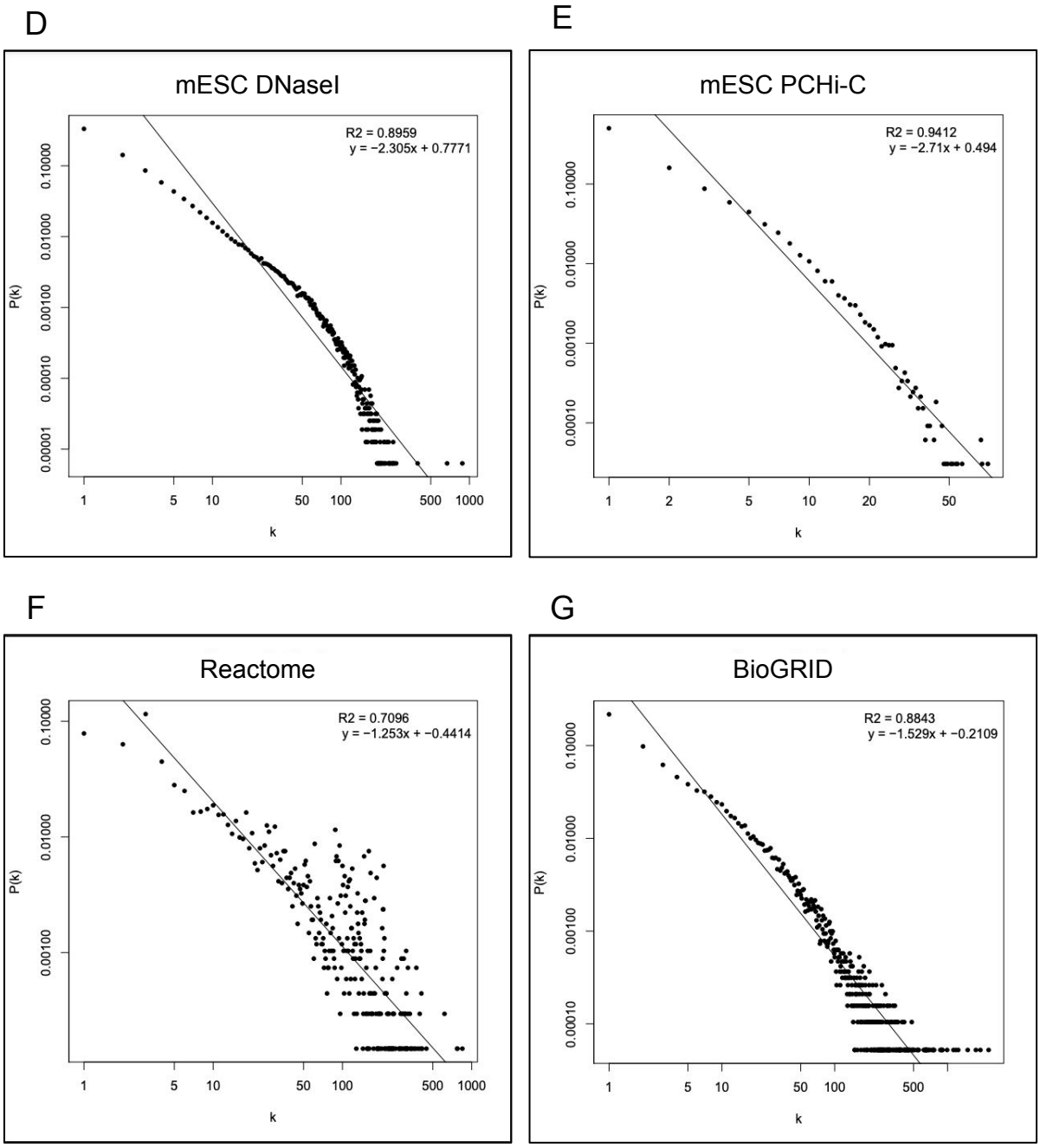



***Figure 3.7. Degree distribution.*** Each graph shows the degree distribution for the three primary immune cell promoter capture Hi-C ChINs of **A)** Monocytes, **B)** Neutrophils, **C)** T-cells and **D)** mESC DNaseI capture Hi-C network, **E)** mESC promoter capture Hi-C network, **F)** Reactome network (provided in splineTimeR), **G)** BioGRID network (provided in splineTimeR). Plots were produced using splineTimeR (version 1.18.0)

| Network | Nodes | Edges | Degree Exponent |
|---|---|---|---|
| Monocytes | 8,104 | 15,780 | 1.6513 |
| Neutrophils | 3,676 | 8,953 | 1.5378 |
| T-Cells | 13,251 | 26,171 | 1.7582 |
| mESC DNaseI | 158,907 | 782,332 | 2.3052 |
| mESC PCHi-C | 32,762 | 51,086 | 2.7102 |
| Reactome | 6,770 | 148,733 | 1.2531 |
| BioGRID | 19,127 | 213,150 | 1.5292 |

***Table 3.6. Summary statistics for the ChINs*** of the promoter capture Hi-C ChINs of monocytes, neutrophils and T-cells and the promoter capture Hi-C and DNaseI capture Hi-C ChINs of mESCs

## 3.4.5 Genic and intergenic nodes maintain distinct topological characteristics

We have so far investigated the large-scale structural features of the networks. We then sought to understand how individual nodes are connected within the networks and whether we could relate distinct connectivity patterns with the biological functions of the nodes. We first looked at whether we could distinguish between nodes that represent genic regions of the chromatin and nodes that represent intergenic regions of the chromatin. To achieve this, we first labeled the nodes with genomic data that defined all of the protein coding genes for each cell type (**See methods: Annotation**). We observe in the case of monocytes, neutrophils and CD4+ T-cells that the proportion of genic and intergenic nodes shows little divergence of ±3% (**Figure 3.8**). We also observe a small increase of 4% from PCHi-C to DNaseI in the proportions of genic and intergenic nodes for the mESC networks (**Figure 3.8**). Of course, the proportion of genic nodes would be expected to be higher given the nature of a PCHi-C assay, but not drastically so given that the DNaseI capture enriches at open chromatin regions where many genes are situated.

We then charachterised both genic and intergenic nodes using four commonly used network measures in order to identify any topological differences between the two sets of nodes (**Figure 3.9**). We calculated the average degree which reflects the average number of edges, or connections, that the nodes in the network maintain. The betweenness centrality which measures the number of shortest paths that pass through a node. This measurement reflects how important a node is in connecting other nodes. The harmonic closeness centrality which is a variant of the closeness centrality for weakly connected networks. The closeness centrality is measured as the average length of the shortest path from a node of interest to all other nodes in the network. However, in a disconnected graph some of these values can be infinite. To solve this problem the harmonic centrality inverses these distances. This measurement reflects how central a node is in the network relative to all other nodes. Finally, the eigenvector centrality which measures the influence of a node within the network. This is calculated relative to the scores of other nodes in the network where a node with a high eigenvalue score connects to many other nodes with high eigenvalue scores.

For the PIC networks we observe a significant increase in the average degree and betweenness centralities of genic nodes vs intergenic nodes. We also observe a significant increase in the harmonic centrality of genic nodes from intergenic nodes in the DNaseI ChIN. However, there is no significant difference in the harmonic centrality between genic and intergenic nodes in both the monocyte (Wilcoxon rank sum test p.value = 0.09) and DNaseI ChINs (Wilcoxon rank sum test p.value = 0.42). This may be explained, at least in part, by the higher density of the DNaseI ChIN (**Table 3.5**) where an increase in edges relative to nodes would reduce the distance between any two given nodes in the network. We also observe non-significant differences in the eigenvector scores between genic and intergenic nodes for monocytes (Wilcoxon rank sum test p.value = 0.082) and neutrophils (Wilcoxon rank sum test p.value = 0.484). Since the eigenvector centrality is not optimised for weakly-connected networks this result may be explained by the DNaseI ChIN having fewer connected components and a higher percentage of nodes in the largest connected component compared to the monocyte and DNaseI ChINs. For the mESC PCHi-C and

DNaseI ChINs we observed consistent significant differences between genic and intergenic nodes for the degree, betweeness and eigenvector centrality scores where the genic nodes appear to maintain unique connectivity within the network. We were however unable to calculate the harmonic scores for the mESC DNaseI ChIN due to its size and computational limitations.

We then considered the differences of connectivity between genic and intergenic nodes when accounting for the effects introduced by baits and OE's. 3C experiments are designed in such a way that most of the interactions of baits are detected. The interactions of OE's on the other hand are only detected in the case that the chromatin fragment that they interact with is a bait. Therefore, baits are likely to be artificially more connected within the networks as a result of this bias. We therefore analysed the centrality measures (degree, betweenness, harmonic closeness and eigenvector) and compared the results between these two subgroups. The baits have a consistently higher average degree than the OE's across all five cell-types exemplifying the interaction bias towards baits in our ChIN's (**Table 3.7**). As expected, we found that almost all of the centrality scores were significantly higher for genic nodes when compared with intergenic nodes of the same type (bait and OE). However, we also observed differences between the centrality scores of baits and OE's irrespective of whether they are genic or intergenic. The harmonic closeness centrality scores are higher in OE's than baits in all five ChINs while the opposite is true for the betweenness centrality which is higher in baits compared to OE's.

As with the degree, these differences are a result of the connectivity bias that exists between baits and OE's. For example, the betweenness centrality measures the number of shortest paths that pass through a given node. I.e. how often a node acts as a bridge to connect other nodes. Where baits are the loci targeted by the 3C experiment, OE's by definition, must be connected to a bait in order to exist within the network, thus explaining the higher betweenness values of baits over OE's. The higher closeness of centralities of OE nodes means that OE nodes tend to be more central in the network. OE's are more central in the network because bait nodes that are more central in the network also have a higher degree centrality, the majority of which are OE's. For example, using the mESC PCHi-C ChIN, when ranking the bait nodes by the harmonic closeness, the top 10% of baits have an average degree of 2.09 while the bottom 10% have an average degree of 0.97. The discrepancies that exist between baits and OE's can be normalised depending on the proportion of baits and OE's that compose each ChIN and may be implemented in future work that expand upon the findings of this chapter.

Here we show that genic nodes tend to be better connected within the networks compared to intergenic nodes and we have discussed why, for some centrality scores and networks, this does not seem to be the case. These results suggest that genic nodes are more influential in the networks than intergenic nodes. It is also important to note that a proportion of intergenic nodes share the same properties as the genic nodes as measured by the centrality scores shown by a modest overlap in the distribution of each of the centrality scores (**Figure 3.10**). One reason for this is that we have not taken into account the activity of the genes and therefore inactive genes are included in the gene set. Secondly, and more relevant to the aims of this thesis is that given the importance of enhancers and the biological similarities reported between promoters and enhancers the intergenic nodes with high centrality scores may in fact be enhancers.


***Figure 3.8. The percentage of genic and intergenic nodes.*** The percentage of genic and intergenic nodes captured by the baits and the interacting other ends (OEs). Genic nodes are labelled using coordinates for the primary transcript of protein coding genes from genome build MGSCv37 (corresponding to UCSC version mm9) were downloaded from Ensembl version 75 using biomart. Excluding chromosomes X and Y.

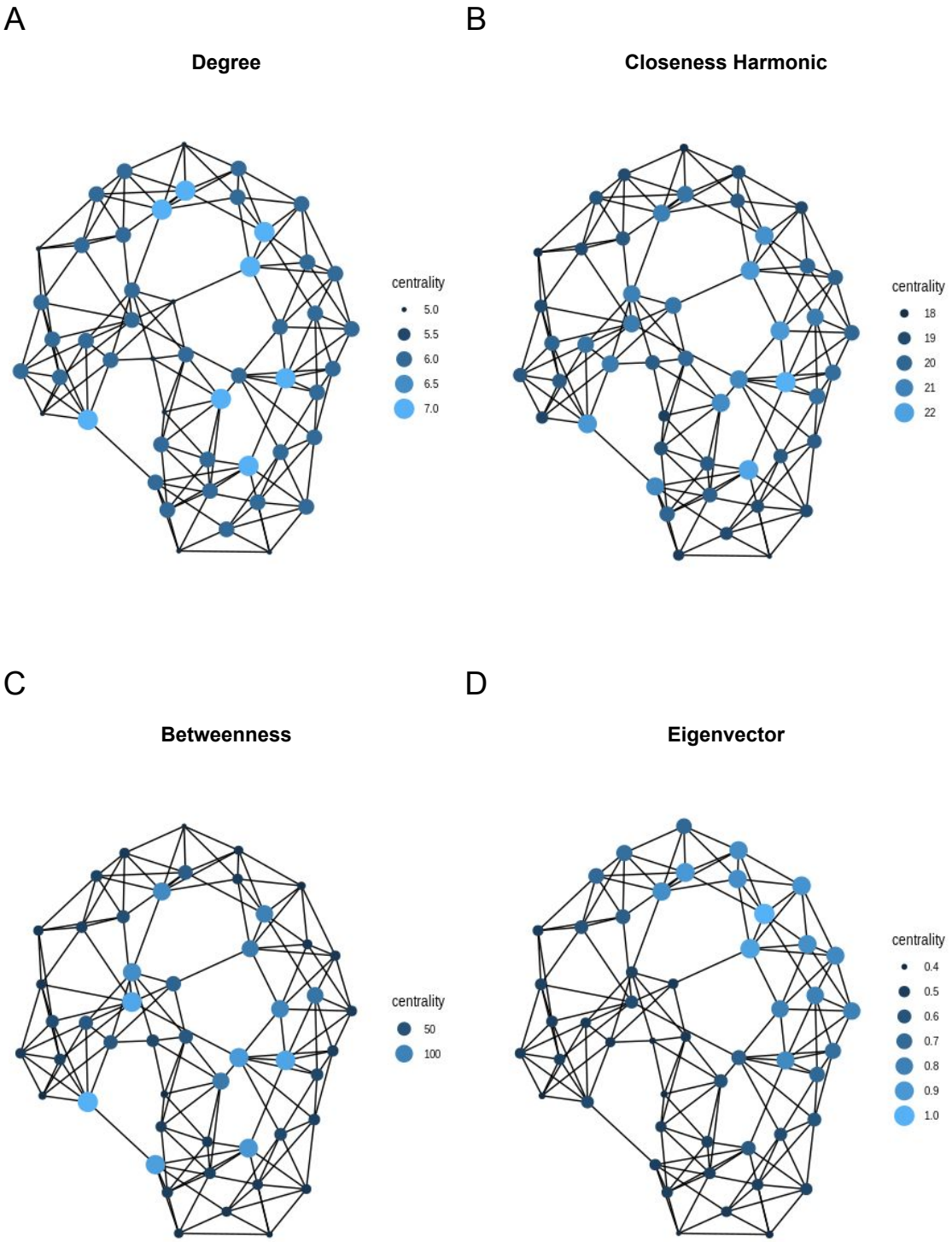



***Figure 3.9. Toy networks showing the effect of different centrality scores.*** A) Degree centrality measures the number of edges for each node. B) Closeness harmonic centrality measures how central a node is in the network relative to all other nodes. This is measured as the inverted average length of the shortest path from a node of interest to all other nodes in the network. C) The betweenness centrality that measures the number of shortest paths that pass through a node. This measurement reflects how important a node is in connecting other nodes. D) Eigenvector centrality measures the influence of a node within the network. This is calculated relative to the scores of other nodes in the network where a node with a high eigenvalue score connects to many other nodes with high eigenvalue scores.

| Genic (Intergenic) | Degree | Harmonic | Betweenness | Eigenvector |
|---|---|---|---|---|
| Monocytes (PCHi-C) | 2.98 (2.20) | **217 (215)** | 12,738 (3,487) | **0.000550 (0.000487)** |
| Baits | 14.47 (11.60) | 199 (186) | 55,170 (30,150) | **0.002380 (0.000789)** |
| OE | 1.84 (1.77) | 221 (215) | 2,379 (2,452) | 0.000105 (0.000475) |
| Neutrophils (PCHi-C) | 2.94 (2.21) | 195 (179) | 6,718 (1,932) | **0.000612 (0.000440)** |
| Baits | 13.53 (10.62) | 176 (136) | 27,151 (18,140) | **0.002603 (0.001016)** |
| OE | 1.87 (1.80) | 200 (181) | 1,285 (1,181) | **0.000082 (0.000413)** |
| CD4+ T-Cells (PCHi-C) | 3.71 (2.53) | **487 (487)** | 30,047 (5,524) | 0.000426 (0.000401) |
| Baits | 18.07 (13.29) | 441 (431) | 143,973 (49,767) | **0.002017 (0.000460)** |
| OE | 2.09 (1.96) | 497 (489) | 5,293 (3,945) | 0.000080 (0.000399) |
| mESC (DNaseI) | 13.93 (10.67) | NA | 825,498 (323,468) | 0.000555 (0.000107) |
| Baits | 22.28 (15.70) | NA | 1,779,626 (847,202) | 0.001004 (0.000181) |
| OE | 2.47 (2.20) | NA | 39,468 (67,310) | 0.000190 (0.000071) |
| mESC (PCHi-C) | 3.07 (1.31) | 913 (796) | 337,033 (54,791) | 0.000686 (0.000380) |
| Baits | 5.11 (3.25) | 877 (421) | 643,357 (548,604) | **0.001096 (0.000035)** |
| OE | 1.43 (1.29) | (941) (798) | 87,557 (51,722) | 0.000352 (0.000382) |

***Table 3.7. Summary table of mean centrality scores between genic and non-genic nodes.*** Harmonic scores for the mESC DNaseI network are not given due to its size and computational limitations

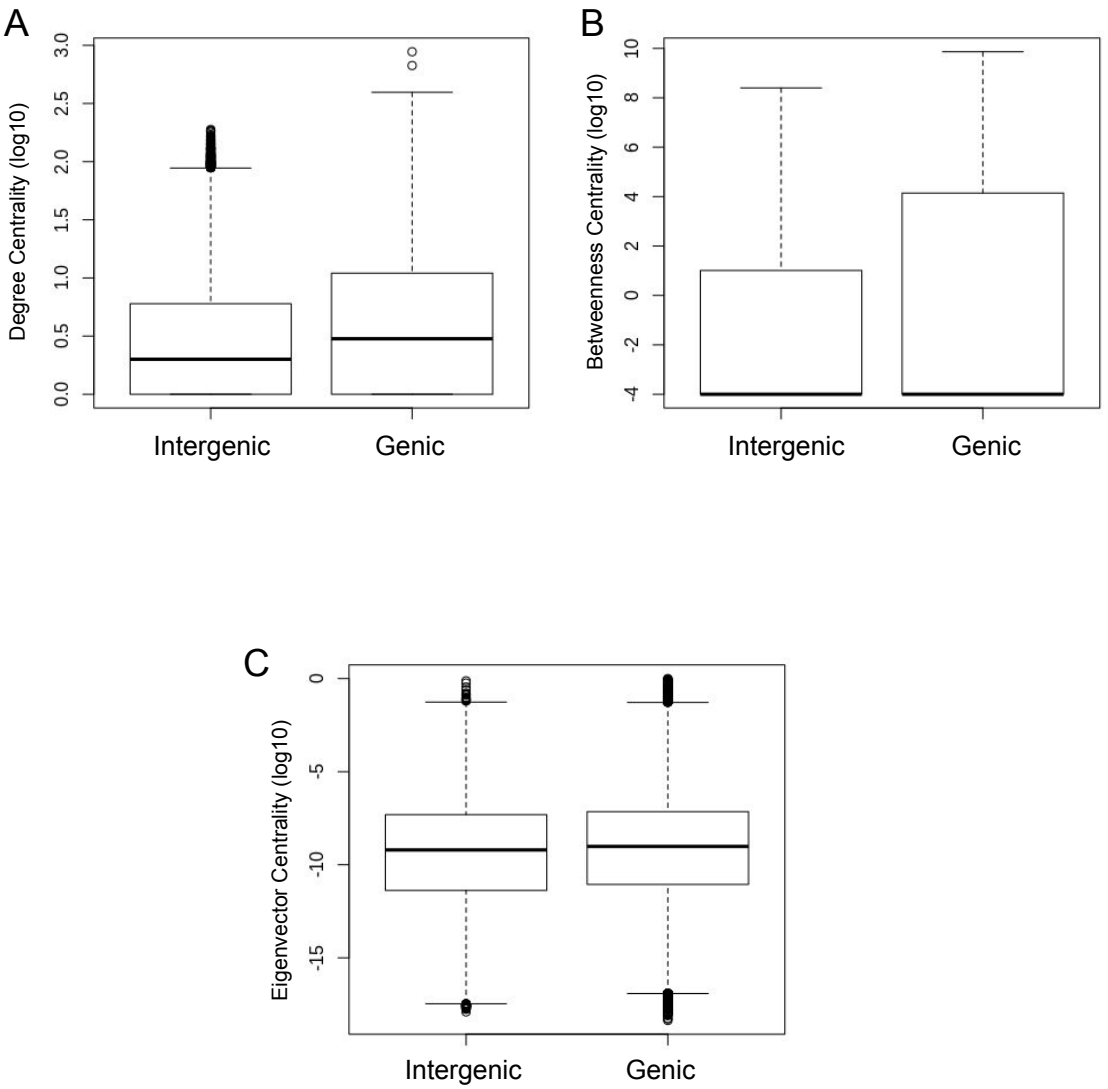



***Figure 3.10. Box-plot of centrality scores for genic and intergenic nodes in the mESC DNaseI network***

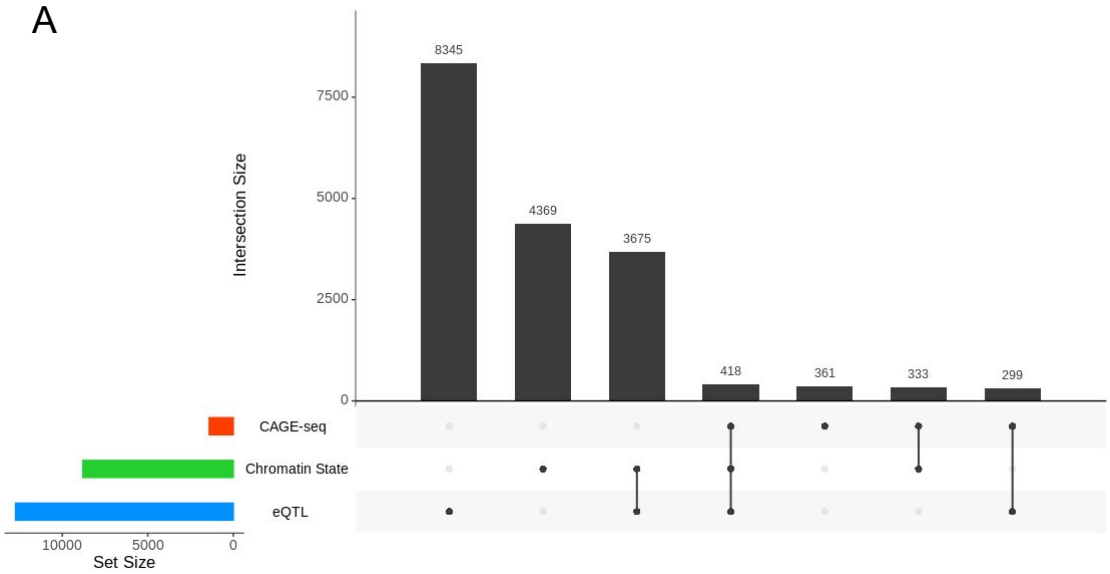
## 3.4.6 There is a low agreement of enhancer associated features in nodes

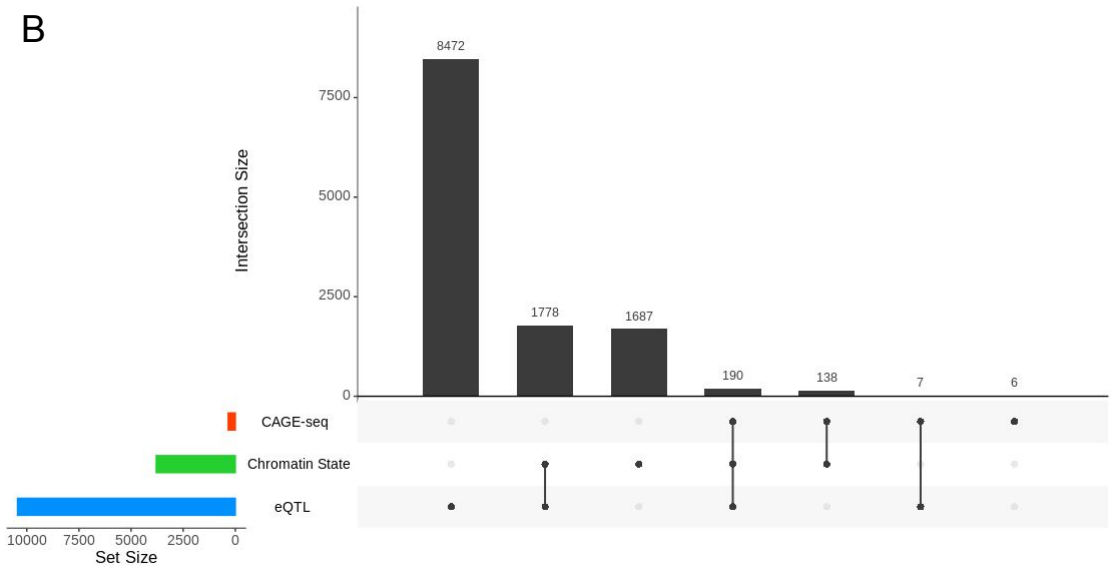
To test the hypothesis that integenic nodes with similar topological characteristics to genes represent enhancers we first identified which nodes contained enhancers. Chromatin states were initially used in **chapter 2** to annotate the nodes as intergenic enhancer nodes or intergenic non-enhancer nodes. While histone marks are a common feature used in identifying enhancers, there are many enhancers that are either not defined by these histone marks or are missed due to technical inefficiencies. To improve the labelling of enhancer nodes we expanded the annotation of the nodes to include six more enhancer associated features for the mESC networks, P300, CAGE-seq, Starr-seq and and three RNA polymerase II (RNAP II) complexes, RNAP s2p, RNAP s5p and RNAP s7p (**see chapter 1.3**). We also included CAGE-seq and matched cell-type specific eQTL data for the three immune cells. These enhancer associated features are not completely reliable, nor do they universally represent all enhancers. These features may also erroneously identify enhancers. Were these features completely accurate and representative of all enhancers the set of nodes defined as enhancers by chromatin states would be identical to the set of enhancer nodes defined by CAGE-seq, and so on. When we look at the overlap of enhancer associated features (defined by their presence in the same node) we can see that this is not the case (**Figure 3.11**).

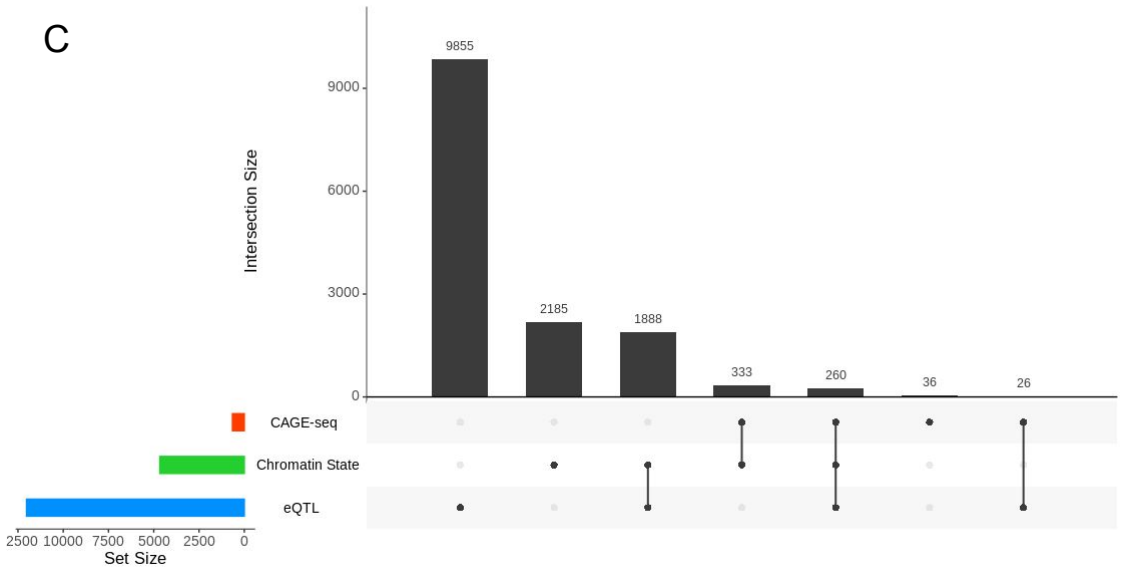
Between the three different primary immune cell networks the set of nodes that are defined by all three annotations account for 3.9%, 1.6% and 1.8% for monocytes, neutrophils and T-cells. This set represents the highest confidence of enhancer activity. The largest set of annotated intergenic nodes are defined solely by eQTLs with 46% in the DNaseI ChIN, 73% in the DNaseI ChIN and 68% in the DNaseI ChIN (**Figures 3.11a, 3.11b and 3.11c**). The eQTL's represent loci that are transcriptionally sensitive to SNP's and can include enhancers. For each of the cell-types 34%, 19% and 18% of eQTLs overlap in various combinations with the CAGE-seq and chromatin state annotations indicating likely enhancer activity. The remaining eQTLs identify additional loci that may contain enhancers that are not annotated by either CAGE-seq or chromatin state data. Of course, eQTLs also represent many other loci that can subtly affect gene expression. For the purpose of these analyses eQTL's were included as they modulate gene expression, if not directly.

In the mESC networks we find that the total number of RNAPs2p annotations is higher than RNAPs5 in the DNaseI ChIN (**Figure 3.11a**) whereas the frequency of RNAPs5p is higher than RNAPs2p in the DNaseI ChIN (**Figure 3.11b**). This is to be expected, where s5p is typically localised to promoters. The largest group in both networks are defined by the three variants of RNA Pol II. The next largest group is defined by a single annotation, RNAPs2p which occurs at sites of transcription, elongation or termination. As with the PIC networks we generally see the highest percentage of overlap between the features that are more abundant. The smallest group of enhancer nodes are defined by CAGE-seq in the primary immune cells while CAGE-seq along with P300 and Starr-seq account for the smallest groups in the mESCs. This is due to CAGE-seq and Starr-seq being much lower throughput and coverage than chromatin states and eQTLs while P300 ChIP-seq peaks are much less abundant. This highlights an important tradeoff in using different types of features to identify enhancers. CAGE-seq, Starr-seq and P300 are often deemed to be more reliable indicators

of enhancer activity; this means that they typically result in much fewer false positives when identifying enhancers. However, these features also miss many possible enhancers either because of the lower throughput and sparser genome coverage when compared with the eQTLs and chromatin states or, as mentioned previously, these features are not universal among enhancers. In using multiple features a much broader set of enhancers can be established. A more nuanced approach using different combinations of features can then be used to validate predictions (**See Chapter 5**).


A
Intersection Size
8345
4369
3675
418
361
333
299
7500
5000
2500
0
CAGE-seq
Chromatin State
eQTL
10000
5000
0
Set Size


B
Intersection Size
8472
1778
1687
190
138
7
6
7500
5000
2500
0
CAGE-seq
Chromatin State
eQTL
10000
7500
5000
2500
0
Set Size


C
Intersection Size
9855
2185
1888
333
260
36
26
9000
6000
3000
0
CAGE-seq
Chromatin State
eQTL
2500
10000
7500
5000
2500
0
Set Size

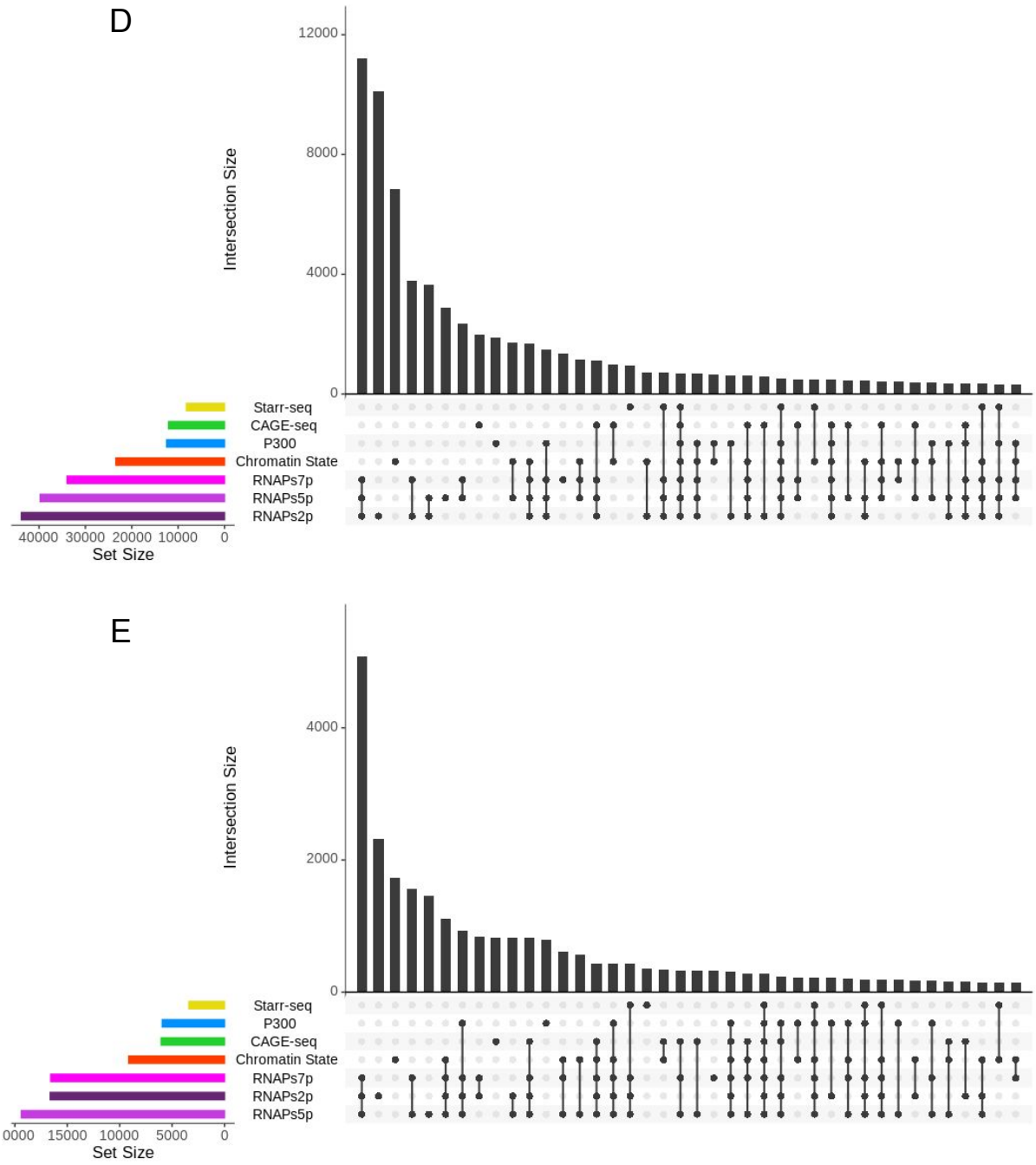


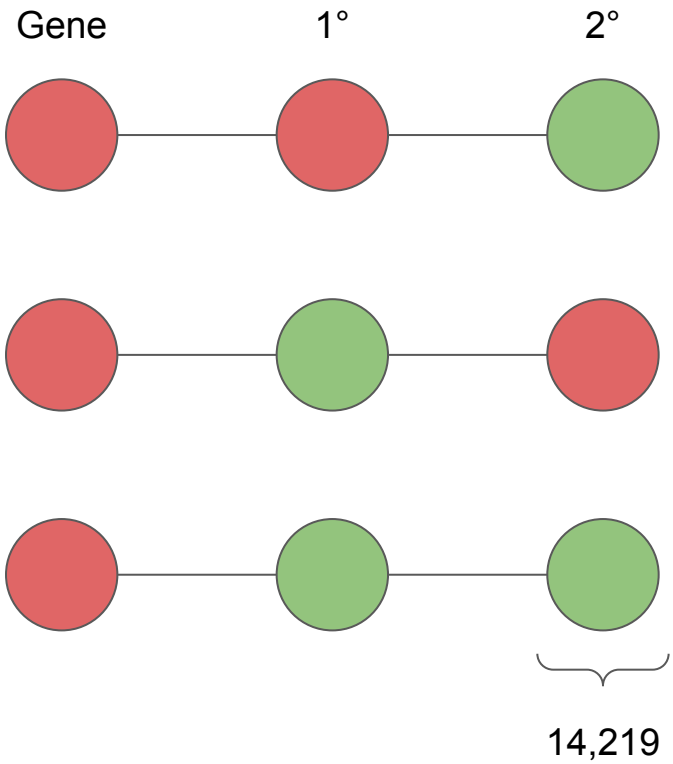

***Figure 3.11. UpsetR plot showing the overlap between enhancer features*** A) Monocytes, B) Neutrophils, C) CD4+ T-Cells, D) mESC DNaseI and E) mESC PCHi-C

## 3.4.7 The DNaseI capture Hi-C ChIN contains more regulatory nodes than the Promoter capture Hi-C ChINs

We next used the enhancer associated features to identify enhancer nodes and understand how the number of enhancers captured varies between cell-types and more interestingly between capture types. Based on the presence of at least one enhancer annotation, in the primary immune cell networks we identified 17,983 (54%), 12,280 (46%) and 14,586 (42%) intergenic enhancer nodes for monocytes, neutrophils and CD4+ T-cells, respectively (**Figure 3.12**). The differences observed in the proportions of intergenic nodes that are labeled as enhancers shows that the monocyte genome contains more enhancers than neutrophils and CD4+ T-cells. We observe that the DNaseI ChIN contains more genic nodes than the DNaseI ChIN that could explain the higher number of enhancer nodes. This is consistent with some previous findings (Rico et al. 2017). However, the observed number of genic nodes in the CD4+ T-cell is higher than those in the DNaseI ChIN. Although we do not know what proportion of these genes are active.

For the mESC DNaseI ChIN, 29,375 (53%) of intergenic nodes contain at least one enhancer annotation; while for the mESC DNaseI ChIN this is 6,204 (43%). The mESC DNaseI ChIN not only captures more enhancers overall but captures more as a proportion of the total. The higher number of enhancers identified within the mESC DNaseI ChIN compared to the mESC DNaseI ChIN is due to more enhancer regions being captured as shown by 34% of mESC DNaseI baits annotated with as enhancers compared to <0.0001% for the mESC PCHi-C baits (**Figure 3.12**). This is likely due to the mESC DNaseI ChIN covering a larger proportion of the total genome (**See 3.4.1**). It also suggests that many regulatory nodes are not directly connected to genic nodes. Indeed, we found that in the mESC DNaseI ChIN 14,219 (48%) of intergenic enhancer nodes were indirectly connected to a genic node by a degree of 2 or more (**Figure 3.13a**). These nodes interact in the second degree or more with promoters via another non-promoter node as shown by node A (**Figure 3.13b**). These nodes are only present in the DNaseI capture ChIN as all baits in the PCHi-C are promoters and therefore all intergenic nodes must be connected to a genic node. This can occur biologically as a result of transient interactions with promoters (Benabdallah et al. 2019) or a complete lack of interaction such as with enhancer chains ((Benabdallah et al. 2019; W. Song, Sharan, and Ovcharenko 2019). These results show that the mESC DNaseI capture ChIN is able to capture more enhancer nodes as well as a more diverse set of enhancers that are otherwise lost in PCHi-C ChINs.


***Figure 3.12. The percentage of intergenic enhancer and intergenic non-enhancer nodes.*** The percentage of intergenic enhancer and intergenic non-enhancer nodes captured by the baits and the interacting other ends (OEs). Enhancer nodes are defined by the presence of at least annotation defined by enhancer chromatin states, P300, CAGE-seq, Starr-seq and and three RNA polymerase II (RNAP II) complexes, RNAP s2p, RNAP s5p and RNAP s7p for the mESC ChINs. For the primary immune cells, enhancer nodes are defined by the presence of at least one annotation of enhancer chromatin states, CAGE-seq or eQTLs.



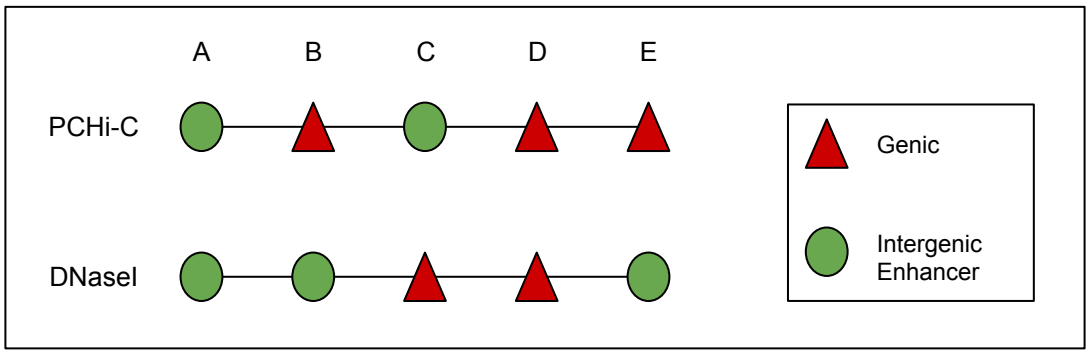




***Figure 3.13. Indirect enhancer-promoter pairs.*** A) Schematic showing the number of intergenic enhancer nodes that are not directly interacting with a genic node. For the mESC DNaseI ChIN, 14,219 intergenic enhancer nodes do not interact directly with genic nodes. B) Indirect interactions between genic and intergenic nodes can only exist in the DNaseI ChIN as the PCHi-C ChIN only identifies promoter-OE interactions by design.

## 3.4.8 Enhancer nodes maintain unique topological characteristics from non-enhancer intergenic nodes

To determine if intergenic enhancer nodes maintain distinct topological characteristics to non-enhancer intergenic nodes we calculated the degree, harmonic closeness (for the DNaseI ChINs), betweenness, and eigenvector centralities as done previously with the genic nodes (**Table 3.8**), compared to non-enhancer intergenic nodes. Enhancer nodes were defined by the presence of at least one annotation.

We observe an increase in the centrality scores in enhancer nodes compared to non-enhancer nodes. For the primary immune cells a significant increase in the centrality measures of the degree (Wilcoxon rank sum test for monocytes, neutrophils and T-cells = p-value = 2.2e-16), harmonic closeness (Wilcoxon rank sum test for monocytes and neutrophils p-value = 2.2e-16, T-cells p-value = 3.1e-14), betweenness (Wilcoxon rank sum test for monocytes and neutrophils p-value = 2.2e-16, T-cells = p-value = 2.2e-16) and eigenvector (Wilcoxon rank sum test for monocytes and neutrophils p-value = 2.2e-16, T-cells p-value = 0.00012). For the mESC networks the same significant differences were observed between enhancer and non-enhancer nodes for the centralities of degree (Wilcoxon rank sum test for PCHi-C and DNaseI p-value = 2.2e-16), betweenness (Wilcoxon rank sum test for PCHi-C and DNaseI p-value = 2.2e-16) and eigenvalue (Wilcoxon rank sum test for PCHi-C and DNaseI p-value = 2.2e-16). As with the genic nodes, we also tested the differences between the centrality scores of both the baits and the OEs. Results were consistent with the genic nodes where the harmonic centrality was found to be higher in enhancer OEs than enhancer baits. Likewise, the degree, betweeness and eigenvector scores mirrored the same pattern as the genic nodes where scores for the enhancer baits were higher when compared to enhancer OEs.  These results suggest that the topological characteristics measured by these centrality scores are more unique genes and enhancers and that intergenic enhancer nodes and gene nodes tend to share many topological characteristics within the networks.

| Enhancer (Non-Enhancer) | Degree | Harmonic Closeness | Betweenness | Eigenvector |
|---|---|---|---|---|
| Monocytes (PCHi-C) | 2.36 (1.86) | 222 (207) | 3526 (3443) | 0.000757 (0.000175) |
| Baits | 13.61 (8.90) | 195 (174) | 43,421 (12,380) | 0.001378 (0.000000) |
| OE | 1.89 (1.61) | 223 (207) | 1,864 (3,123) | 0.000731 (0.000182) |
| Neutrophils (PCHi-C) | 2.48 (1.95) | 191 (169) | 2662 (1318) | 0.000801 (0.000138) |
| Baits | 12.99 (8.45) | 166 (108) | 22,997 (13,675) | 0.001591 (0.000487) |
| OE | 1.97 (1.66) | 193 (172) | 1,672 (771) | 0.000762 (0.000122) |
| CD4+ T-Cells (PCHi-C) | 2.71 (2.08) | 505 (474) | 7547 (4038) | 0.000768 (0.000131) |
| Baits | 14.76 (11.66) | 470 (389) | 70,396 (26,903) | 0.000404 (0.000523) |
| OE | 2.18 (1.80) | 506 (477) | 4,738 (3,372) | 0.000785 (0.000119) |
| mESC (DNaseI) | 10.62 (2.63) | NA | 571,553 (74,722) | 0.000161 (0.000053) |
| Baits | 17.48 (7.43) | NA | 983,475 (212,061) | 0.000203 (0.000080) |
| OE | 2.57 (1.99) | NA | 87,620 (56,720) | 0.000113 (0.000049) |
| mESC (PCHi-C) | 1.44 (1.20) | 913 (707) | 83,562 (33,001) | 0.000638 (0.000185) |
| Baits | 4.58 (2.53) | 619 (316) | 1,028,513 (292,101) | 0.000091 (0.000005) |
| OE | 1.43 (1.19) | 914 (709) | 78,829 (31,154) | 0.000641 (0.000186) |

***Table 3.8. Summary table of centrality scores between enhancer and non-enhancer nodes***

## 3.4.9 Some centrality scores can be used to classify enhancer nodes

Given the unique topological features of enhancer nodes we then attempted to predict intergenic nodes as enhancers using the degree, betweenness and eigenvector centrality scores in both the PCHi-C and DNaseI mESC derived networks. Nodes were labelled as enhancer and non-enhancers based on the presence of a minimum of one enhancer associated feature of any type. The performance of each centrality score in correctly classifying the nodes as enhancers was measured by the precision and recall at incremental thresholds from the maximum to the minimum centrality scores. For the mESC DNaseI ChIN the results show that both the degree and betweenness centrality scores show modest performance in predicting intergenic enhancer nodes where the degree centrality AUPRC = 0.55 and betweenness centrality AUPRC = 0.50 compared to a baseline figure of 0.43 (**Figure 3.14**). The eigenvector centrality performs poorly with an AUPRC = 0.42, below the baseline. This would suggest that the eigenvector centrality is not well suited for identifying enhancer nodes despite the mean score being higher in enhancer nodes vs non-enhancer nodes. However, the eigenvector centrality score should not be discarded as this is only a simple, one dimensional analysis of the feature. Indeed, some features can only be described in a multidimensional space and as such we are now investigating the classification value of all of these features (**see Chapter 6.1**).

The results for the mESC DNaseI derived network were more consistent with the AUPRC of all three centrality scores above the baseline (degree AUPRC = 0.75, betweenness AUPRC = 0.71, eigenvector AUPRC = 0.54, baseline = 0.5). The increase in the AUPRC for each of the centrality scores of the mESC DNaseI ChIN is likely a combination of factors including the higher coverage, superior connectivity and closer representation to the true mESC ChIN. AUC is often used in the literature to describe the predictive performance of features (Thibodeau et al. 2017). However, AUC neglects the class imbalance that often occurs in these datasets. That is to say the number of positive vs negative labels are not 1:1. This can artificially inflate the area under the curve and leads to overstating the classification performance of these features or models. It has been proposed previously that precision-recall curves are a much better indicator of classification performance for binary classification (Takaya Saito 2015). For a single feature an AUPRC of 0.75 and 0.71 could be considered very good, particularly when SVM and CNN's when trained on multiple features achieve a AUPRC between 0.78 and 0.9 (Min et al. 2017).

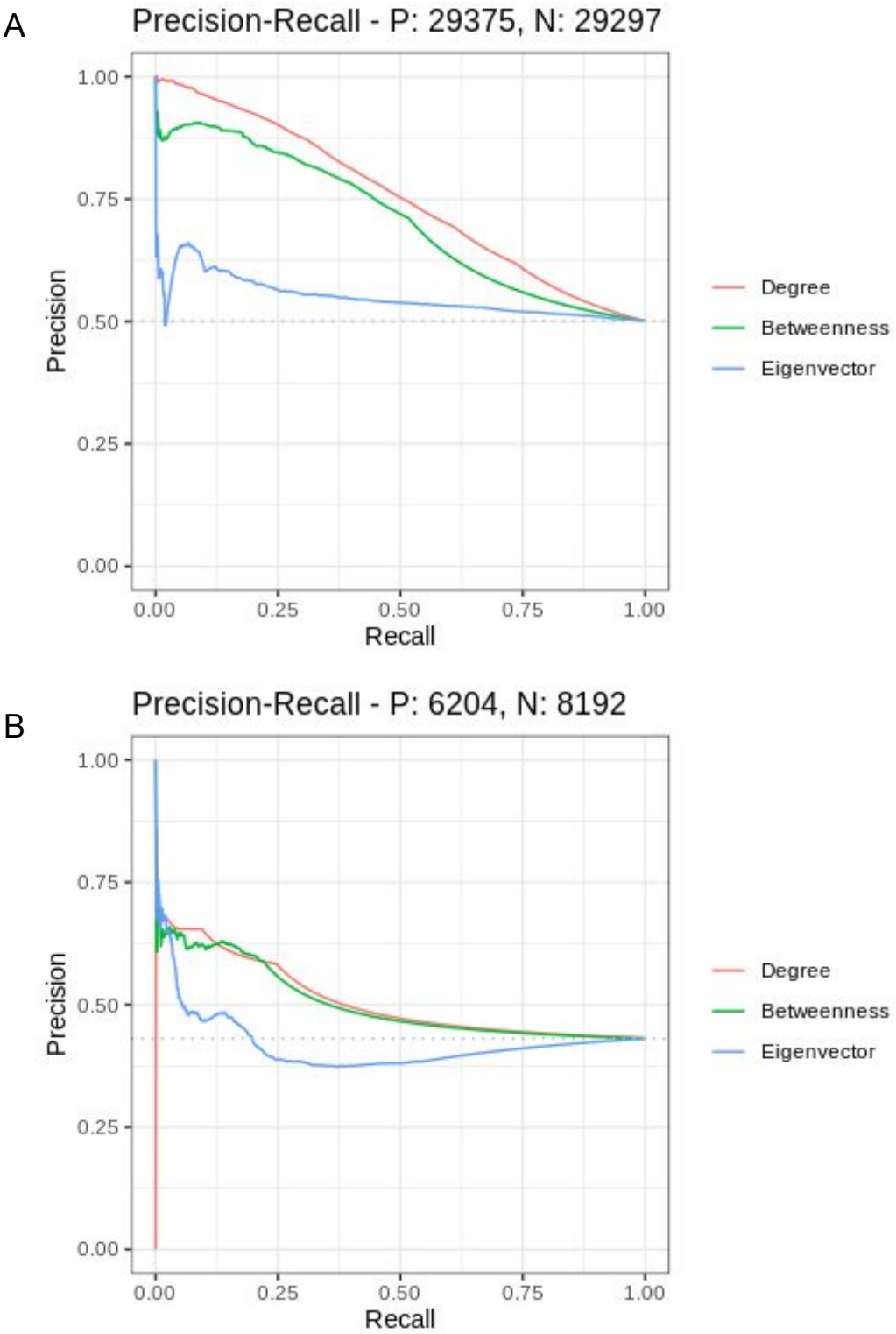



***Figure 3.14. Precision-recall curves for the classification of enhancer nodes using centrality measures***. A) Precision-recall curves for the degree (red), betweenness (green) and eigenvector (blue) centralities for classifying enhancer nodes defined by at least one annotation tested on the mESC DNaseI ChIN. B) mESC PCHi-C ChIN.

# 3.5 Discussion

The primary findings of this chapter are that (i) enhancers and genes within our networks tend to contain distinct topological features when compared to non-enhancer and intergenic nodes and that (ii) these features can be used to predict intergenic enhancer nodes. This suggests that enhancers and genes are more connected within the 3D chromatin structure than other loci. This commensurate with the looping mechanisms that localise enhancers and gene promoters resulting in an increased contact frequency (Schoenfelder and Fraser 2019). Furthermore, it is widely accepted that chromatin structure plays a role in gene regulation. However, to what extent is debatable. Some studies have shown that the disruption of TADs does not lead to wide scale changes to gene expression (Ghavi-Helm et al. 2019). Although this may suggest that mechanisms other than the chromatin topology may regulate gene expression, other explanations such as redundancy in the networks are equally plausible (Ghavi-Helm 2020). In fact, these mechanisms are shown in the scale-freeness and small worldness of the ChINs that are characteristically resilient to perturbation.

Surprisingly, there is currently no previous literature that addresses the topological features of genes and enhancers in 3C derived ChINs, even though the analysis of network characteristics of genomic features is not new (**see Introduction**). By representing 3C data as ChINs the topological properties can be reconciled with the biological function. We show that enhancer nodes can be predicted using centrality measures. The local measurement of the degree centrality and the global measurement of the betweenness, harmonic closeness and eigenvector centralities provide a tangible link between local and global chromatin organisation and gene regulation. The centrality measures used provide a useful feature that can be leveraged in future to improve the prediction of putative enhancer sequences. In order to do so, it would be beneficial to assess other network properties in addition to the centrality measures outlined in this chapter, to further characterise enhancers. These may include clustering approaches to identify communities of nodes for example as we know that biochemically similar loci tend to form localised structures via homotypic attraction (Robson, Ringel, and Mundlos 2019) and that genes found within the same connected component tend to have similar regulation patterns (Viksna et al. 2019). These additional features can then be combined with the centrality scores to train a machine learning model (Lim et al. 2018). Network analysis of enhancers such as this will benefit greatly from future improvements in 3C technologies as the resolution, sensitivity and coverage increase. More faithful network models of chromatin conformation can be created that will reduce the bias observed in current iterations. This does come with the caveat of an increased requirement of computational resources, which as we saw with the mESC mcHi-C dataset can be a limitation. Alternatively, or better still in addition to this, error estimations of the centrality scores can also be estimated using experimental replicates or between individual cells in the case of scHi-C datasets. These errors could then be used to normalise the final centrality scores in order to model the noise produced by these experiments.

We utilised the 5 remaining 3C datasets to examine the differences between cell types and between different 3C capture types. Between the datasets used for analysis we observed little variability between the primary immune cell-types in the fragment sizes, which is to be expected due to all three subject to digestion with the restriction enzyme HindII to isolate the

interacting fragments of DNA. The contact distance, which is the linear distance between two interacting fragments, was also consistent between all three cell-types. The experimental and processing bias can be seen clearly when we then compare the fragment sizes and contact distances retrieved in the mESC datasets using single-cell Hi-C, DNaseI capture Hi-C and PCHi-C. Here we observed large differences between the fragment sizes and contact distance. Furthermore, in the mESC DNaseI ChIN we found there was reduced connectivity, with a substantial increase in the number of connected components compared to the mESC DNaseI ChIN. The increased number of connected components were also observed in the three PIC networks. It has been previously reported, using the PCHi-C dataset for the primary immune cells, that the occurrence of connected components can be related to biological function (Viksna et al. 2019); we show this briefly in **Figure 3.6** whereby most of the connected components were enriched for a particular chromosome. Our results suggest that the capture type used is also a significant determinant of network structure as demonstrated by the relatively small changes seen between the three PIC networks compared to the large scale changes observed between the PCHi-C, DNaseI and DNaseI ChINs in the mESCs. Overall, the type of 3C experiment carried out and the processing used on that dataset can influence both the fragment size and the enrichment of specific contact distances. Although these changes are artificial as they are introduced by experimental bias, they do not appear to significantly alter the fundamental structure of the networks as shown by the retention of scale-freeness and small worldness across all 5 ChINs. Rather, they present a different viewpoint of the chromatin architecture. There is a tradeoff between different methods that prioritise the resolution of the fragment sizes and short range interactions for wider genome coverage and longer range interactions and vice-versa. While this presents difficulty in choosing between (i) resolution to more accurately locate enhancer sequences and (ii) coverage to capture more enhancer sequences, there are potential solutions. For one, it could be beneficial in future work to combine the networks of different 3C assays carried out in the same cell-type to concurrently improve both the resolution, coverage and contact distances. This has been achieved either by combining the various 3C networks into a single consensus network (Rao et al. 2014), or alternatively, using multiplexed networks (Didier, Brun, and Baudot 2015). For the latter, distinct networks (in this case various 3C networks) are layered atop one another and layers are connected where 3C fragments overlap which should, in theory, reduce the loss of information (such as resolution) compared to simply combining the networks.

Although we observe changes in the contact distances and fragment sizes for each 3C experiment, it does not appear to affect the large-scale topology of the network. Using the mESC PCHi-C and DNaseI ChINs we were able to provide a direct comparison between two 3C methods. Across both networks we observed patterns that are consistent with scale-free and small world properties (Barabási 2009; Watts and Strogatz 1998). Both features are commonly found within complex networks, although it has been argued that current definitions of scale-freeness are too broad (Stumpf and Porter 2012). Other more rigorous definitions of scale-freeness have been presented that were not used for our analysis (Broido and Clauset 2019). **The primary objective of this work was to understand if the 3C method fundamentally disrupted the large scale organisation of the ChINs we derive from them**. The small world and scale-free properties also present an interesting theoretical framework for the organisation of chromatin and within that, the organisation of enhancers and genes. For example, the fitness model is one of the two main mechanisms proposed for the emergence of scale-free properties (Bianconi and Barabási 2001). This

describes a process in which an inherent trait of a node provides it with increased fitness over other nodes that leads to fit nodes becoming more connected than less fit nodes, such as hubs. The enhancer and genic nodes can be considered to be more 'fit' than non-enhancer and non-genic nodes as shown by their higher centrality scores. It may be therefore interesting to analyse the fitness of nodes by measuring changes in connectivity of these nodes in networks modelled from time-series 3C data which is already available (Gibcus et al. 2018). This could be interesting in the context of mitosis where the genome undergoes drastic changes in the 3D structure of chromatin (Shoaib, Nair, and Sørensen 2020). During mitotic exit the 3D structure of chromatin and regulatory circuits are re-established and the fitness of nodes may provide clues as to how this is achieved.

While we looked at potential experimental bias arising from the networks themselves our results also show that variability may also arise from the annotations used to validate our method. As is common in all enhancer prediction models, the identification of a suitable 'truth set' of validated enhancers is non-trivial. Currently gold standard approaches such as CRISPR perturbation are limited in both the scalability to assess enhancer sequences genome wide as well as the availability of this data for multiple cell-types. As there is no reliable genome-wide gold standard for enhancer identification, we identified several enhancer associated features that could be used to label the intergenic nodes as enhancers and non-enhancers. For mESC's these included histone ChIP-seq summarised in a chromatin state model, three RNAPII variants, P300 ChIPseq, Starr-seq and CAGE-seq from the FANTOM project and for the primary immune cells this included chromatin states, CAGE-seq and eQTL's (**See Methods: Data**). The reason why multiple enhancer associated features are needed is due to the heterogeneity of enhancers and this can be observed by the lack of full agreement in the overlaps between each of the features (**Figure 3.11**). This could be further improved by labelling the nodes using additional annotations. For example, CTCF plays a role in establishing a range of chromatin structures including the establishment of chromatin loops (Sjoerd Johannes Bastiaan Holwerda 2013). By using multiple features we are able to capture a wide range of distinct enhancer classes. However, a caveat of using such broad definitions is that it can result in many false positives. For this reason it is important to acknowledge that they are putative enhancers as there is no direct evidence that they are functional, this is commonly known as validation creep (Halfon 2019). The low agreement between features highlights the need to uncover new features of enhancers, something that we demonstrate could be achieved using centrality measures and indeed the propagation method outlined in **chapters 4 and 5**.

# Chapter 4: Integrating gene expression with chromatin interaction networks: The development of 3D-SearchE

---

## 4.1 Introduction

The increased availability of whole genome sequencing (WGS) and functional annotations means that identifying functional non-coding variants is becoming more feasible. Often, understanding whether non-coding variants are functional relies on determining whether they occur within regulatory elements such as enhancers. However, current definitions of enhancers are broad and non-specific while their mode of action is poorly understood. Additionally, non-coding variants can occur at loci that induce structural changes in the chromatin organisation. This can affect the localisation of enhancers to gene promoters. The ability to identify these regions and define whether a variant is functional or not relies solely on the annotations available for the cell-type in question. We have shown in **Chapter 3** that network theory approaches are a potentially useful tool in further annotating the genome by deciphering the complex relationships between the chromatin architecture genes and enhancers. However, these approaches neglect a key bit of information, the expression of the genes. One area of particular interest in structural genome biology is how gene expression and chromatin conformation are linked (Dekker and Misteli 2015). Naturally, this relationship includes gene regulatory elements such as enhancers. As explained in previous chapters the conformation of chromatin mediates the flow of information between enhancers and genes by localising the two in 3D space or insulating them from one another depending on the specific expression patterns required between cell-types and during the life cycle of those cells.

Network analysis becomes a particularly powerful approach when the nodes and edges of a network are labeled with *a priori* information that describe the biological properties of the fragments and interactions they represent. In the previous chapter we show that by labelling the nodes with enhancer data can facilitate a better understanding of how enhancer and genic nodes maintain distinct connectivity patterns within the ChINs. This analysis links two of the main components that control gene regulation; enhancers and chromatin architecture. Importantly though, it does not account for the expression of genes. We hypothesised that the integration of RNA-seq data would improve the characterisation and identification of enhancer nodes. To link the properties of the chromatin architecture with enhancers required measuring centrality scores for each enhancer node and measuring the differences observed against non-enhancer nodes (**see Chapter 3)**. In order to link gene expression with enhancer nodes and the chromatin architecture the process is more complicated. Here, the gene expression in all of the genic nodes must be reconciled with all of the enhancer nodes while taking into account the chromatin architecture that links them all together. To achieve this, we adopted a network propagation approach.

Network propagation can be thought of as the transfer of heat from one radiator to many others via copper pipe. The heat transferred to a group of 2 radiators that are directly connected to the heat source will be higher than the heat transferred to a group of 5 radiators. Similarly, a radiator connected to the heat source via another radiator will receive less heat than a radiator connected directly to the heat source. The conductivity of the copper pipes can also be determined by a value called alpha to control how far the heat is able to travel from the source. For networks and propagation the same logic is used to determine the topological properties of nodes where heat can be any value, the radiators can be any entity and represented by nodes while alpha is often user defined based on contextual considerations.

In recent years network propagation has emerged as the state-of-the-art in many fields and particularly in the investigation of PPI networks to identify associations between genes and novel drivers of disease, (Peña-Castillo et al. 2008; Navlakha and Kingsford 2010). For example, it has shown particular promise through the propagation of mutational frequencies across PPI networks in order to identify previously unreported cancer drivers (Hofree et al. 2013). Fundamentally, network propagation aims to describe the associations and relationships between the nodes in the network. This type of analysis is not limited to the study of biological systems either. The most famous is the PageRank algorithm which was first employed by Google (Page et al. 1999). PageRank works by returning websites (represented as nodes) in order of their importance. Imagine, for website A its importance is a function of the number of websites linking to it, in this example two with B and C, and the importance of B and C determined by the number and importance of the websites linking to them. Using such a process enables all of the websites to be ranked relative to one another. Network propagation has the distinct benefit of incorporating *a priori* information about the nodes; meaning that some nodes are more important than others and as such connections to these nodes are considered more favourable by the propagation algorithm (Cowen et al. 2017). We proposed that a network propagation algorithm was therefore ideal to rank the intergenic nodes according to their connectivity to genic nodes (with the assumption that ChINs are gene centric). In this case, the RNA-seq values would act as the *a priori* information. Here we propose a methodology to apply network propagation across ChINs to rank intergenic nodes and hypothesise that higher ranking intergenic nodes represent loci harbouring enhancers. In this chapter we outline the development of a method to integrate gene expression data with ChINs by propagating RNA-seq values in order to rank intergenic nodes.

# 4.2 Development of 3D-SearchE

## 4.2.1 Using Network Propagation to Integrate Gene Expression and 3C Data

Our network-based propagation algorithm was provided by the lab of Professor Rosalba Giugno (University of Verona) and was originally designed for smoothing mutational frequencies across PPI networks. These algorithms exist in various guises, most notably as

Google PageRank as mentioned previously. Here, we outline the adaptation of this algorithm to integrate and propagate RNA-seq values across ChINs to generate an **imputed activity score** (IAS) at intergenic nodes. Gene expression data for the primary immune cells were downloaded from (L. Chen et al. 2016) and the mean expression across 200 individuals were used. Gene expression data for the mESC's were downloaded from (Kolodziejczyk et al. 2015) using the serum population and the mean expression per gene. Both RNA-seq datasets were integrated into the ChINs and analysed in the same way (see **chapter 5** for the results from mESCs).

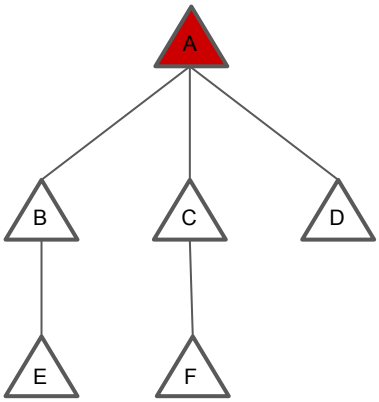
The propagation process is defined by several components and parameters. Firstly, the interactions between all the nodes in the network (**Figure 4.1a**) are defined by the adjacency matrix $A$ (**Figure 4.1b**). In this case the $A$ is symmetric as the network is unweighted; all of the edges are bidirectional. $A$ is then normalised using a symmetric normalisation method defined in by the equation $A = D - 0.5 * A * D - 0.5$ (**Figure 4.1e**) where D is the diagonalised degree sum of $A$ (**Figure 4.1c**). The result is a degree normalised transition matrix $W$' that defines the probabilities of transitioning from a one node to another (**Figure 4.1d**).

Each node also contains an *a priori* value, in our case these are set by the expression values of each gene as defined by an normalised RNA-seq dataset. The *a priori* values are then mapped to the appropriate nodes defined by $Y: V -> [0, MAX]$ where $0$to $MAX$ is given by the expression of each gene (**Figure 4.1f**). The amount of gene expression that can be propagated from node A to the rest of the network is defined by two hyperparameters. The first is the alpha value, $\alpha$. It is able to restrict the propagation of expression values anywhere between 0 to 100% and is defined by a unit interval [0,1] (**Figure 4.1g**). The second parameter is defined as the number of times the algorithm is iterated and can also be set to 1 to limit the amount of propagation or more to increase the amount of propagation (**Figure 4.1g**).

All of the inputs are defined with $\alpha$, $W$' and $Y$which can be used to determine the function F given by $Ft := \alpha W'Ft - 1 + (1 - \alpha)Y$ where F(v) defines an imputed expression value at any given intergenic node, taking into account both the global topology of the network and the relative expression levels of all the genes in the network. This algorithm is best described as simulating a process of propagating information across the network, starting from node A and where node A in the network contains an *a priori* value (**Figure 4.1h**). The propagation algorithm then propagates this value from A to the rest of the network (**Figure 4.1h**). This process results in all intergenic nodes that are connected to a genic node receiving an imputed activity score (IAS) that is determined by both the activity of the connected genes and the topology of the ChIN. A visual example of IAS enrichment in the largest connected component of the monocyte ChIN can be seen in **Figure 4.2**. The IAS can then be used to rank intergenic nodes in order to predict which nodes contain enhancers.

A)

**Network Propagation**




B)

**Adjacency Matrix**

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| 0 | 1 | 1 | 1 | 0 | 0 |
| 1 | 0 | 0 | 0 | 1 | 0 |
| 1 | 0 | 0 | 0 | 0 | 1 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |

C)

**Diagonal Matrix**

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| 3 | 0 | 0 | 0 | 0 | 0 |
| 0 | 2 | 0 | 0 | 0 | 0 |
| 0 | 0 | 2 | 0 | 0 | 0 |
| 0 | 0 | 0 | 1 | 0 | 0 |
| 0 | 0 | 0 | 0 | 1 | 0 |
| 0 | 0 | 0 | 0 | 0 | 1 |

D)

**Transition Matrix**

| A | B | C | D | E | F |
|---|---|---|---|---|---|
| 0 | 0.3 | 0.3 | 0.3 | 0 | 0 |
| 0.5 | 0 | 0 | 0 | 0.5 | 0 |
| 0.5 | 0 | 0 | 0 | 0 | 0.5 |
| 1 | 0 | 0 | 0 | 0 | 0 |
| 0 | 1 | 0 | 0 | 0 | 0 |
| 0 | 0 | 1 | 0 | 0 | 0 |

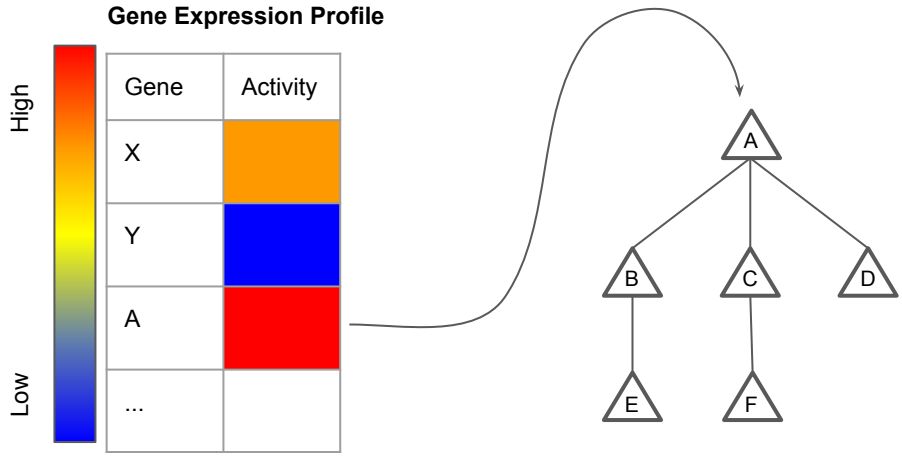
E)

**Equations**

$$A = D - 0.5 * A * D - 0.5$$

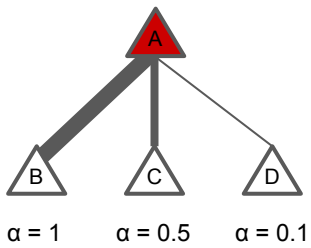
$$Ft := \alpha W' Ft - 1 + (1 - \alpha) Y$$

F)




G)

**Alpha Value**




**Iteration Value**

| | A | B | C | D | E | F | |
|---|---|---|---|---|---|---|---|
| | 100 | 0 | 0 | 0 | 0 | 0 | $I_0$ |
| | 82.21 | 5.59 | 5.59 | 5.48 | 0.56 | 0.56 | $I_1$ |
| | 69.45 | 9.38 | 9.38 | 9.01 | 1.39 | 1.39 | $I_2$ |

H)

**Network Propagation**

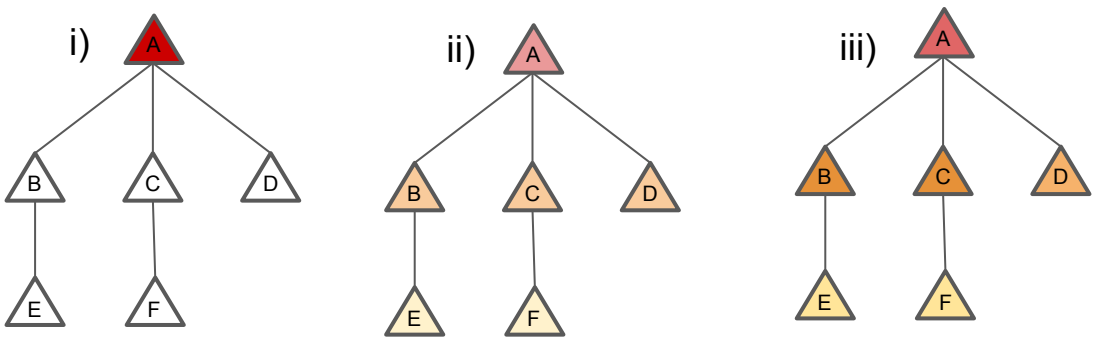



***Figure 4.1. Network propagation schematic*** **A) i)** A toy model of network propagation run with two iterations with an insulating score of 0.2. A single gene node A contains an *a priori* expression value. aii) In the first iteration, the value from A is smoothed to its. aiii) In the second iteration the values are further smoothed and important nodes such as B and C are ranked higher. **B) i)**This process is achieved by representing the interactions between nodes of the network in an adjacency matrix. **ii)** The degree of each node is represented in the diagonal matrix. **iii)** The probabilities of transitioning from one node to another is calculated from the diagonal matrix and the adjacency matrix using equation 1. These results are then represented in the degree-normalised transition matrix. **iv)** The expression value for each node at the start and after each of the two iteration values after propagation using equation 2.

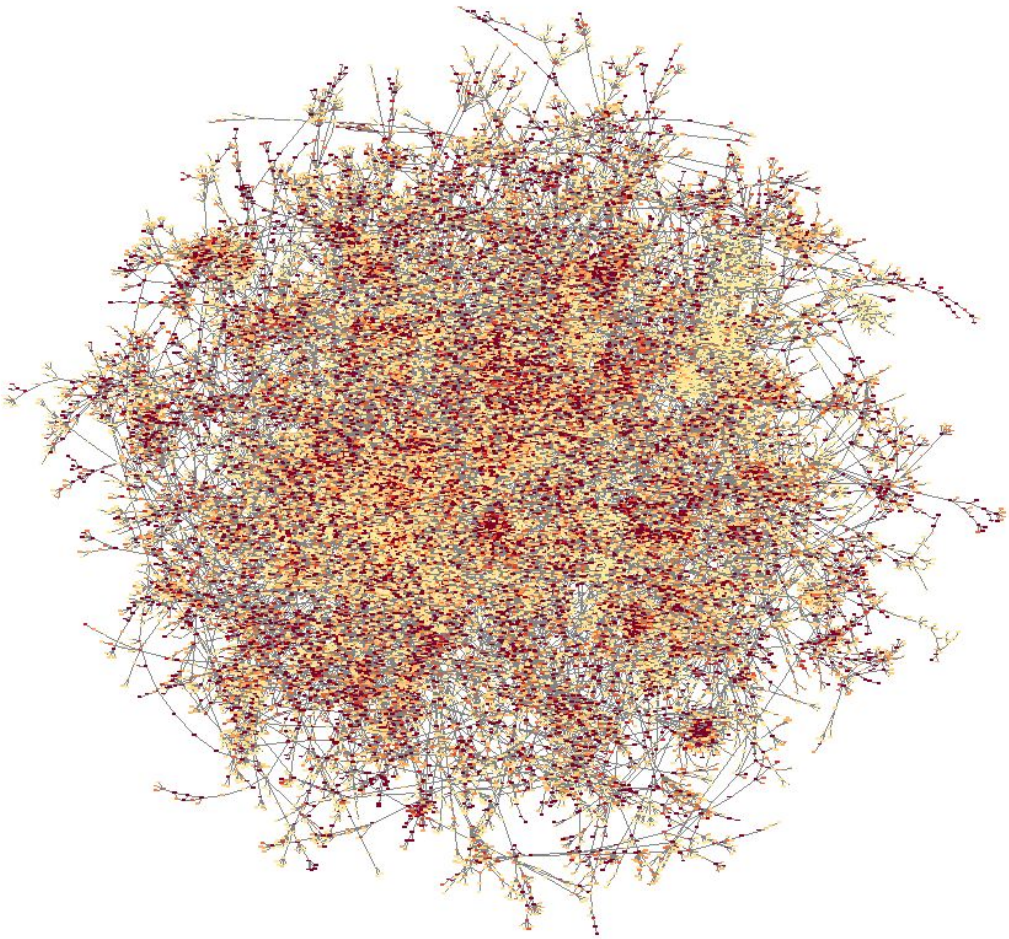


***Figure 4.2. IAS in the largest connected component of the monocyte PCHi-C network***. Nodes are organised based on an Edge-weighted Spring-Embedded Layout where the frequency of interaction denotes the edge weight. Nodes are then coloured from yellow to red based on the imputed activity score calculated by 3D-SearchE.

## 4.2.4 Primary immune cell chromatin interaction networks

We initially used the promoter capture Hi-C (PCHi-C) dataset containing 17 primary immune cell-types with matched gene expression data to prototype our method. Each of the 17 cell types originates from a common multipotential hematopoietic stem cell (HSCs) and therefore each network will share common interactions. These cells are then differentiated into several lineages to form a selection of specialised blood cells. The 3D-folding of chromatin and cell-type specific chromatin contacts play a major role in this process as they coordinate the localisation of enhancers to genes resulting in cell-type specific gene expression. We first investigated how the proportion of shared and cell-type specific interactions vary between the ChINs of three cell types. We generated our initial results using the networks of monocytes, neutrophils and CD4+ T-cells as these contained matched gene expression data. Both monocytes and neutrophils originate from the common myeloid progenitor cell. While T-cells originate from the common lymphoid progenitor cell. We then calculated the percentage of shared chromatin interactions between these three cell-types (**Table 4.1**). Typically, cells with a common lineage tend to share a higher proportion of their interaction. For example, Naive_CD4+ T-Cells share just over a third (36.1%) of their interactions with naive B cells. This is to be expected as they both originate from the common lymphoid progenitor (**Figure 4.3**). This is comparably higher than the percentage of interactions Naive CD4+ T-Cells share with cells originating from the common myeloid progenitor. However, monocytes share 41.6% of their interactions with neutrophils while a higher percentage of 51.8% with T-cells. Neutrophils share 48.5% of their interactions with monocytes and 42.9% with T-cells. While T-cells share 27.0% with monocytes and 19.2% with neutrophils. These data show that all three cell-types share many chromatin interactions with each other; monocytes share a higher proportion of their interactions with T-cells than neutrophils, despite monocytes and neutrophils being more closely related. Interestingly, the chromatin of neutrophils is involved in a peculiar physical function known as neutrophil extranuclear trap (NET) that can trap, neutralise and kill pathogenic material (Papayannopoulos 2018). This may explain, in part, why neutrophils appear as distinct from monocytes as they are from T-cells despite the differences in the lineage.

As well as testing the performance of the method in the individual cell-type networks, we also tested the ability of our model to identify cell-type specific enhancer nodes from two consensus networks (**Figure 4.4**). The use of consensus networks were included to understand how the predictive performance of propagation is affected in a ChIN consisting of monocyte, neutrophil and T-cell interactions. Secondly, to understand if cell-type specific RNA-seq could be propagated on a ChIN composed of the interactions of the other two cell-types. This would be particularly useful in cases where 3C data is unavailable for particular cell-types.

The first consensus network consisted of the nodes and edges from the monocyte, neutrophil and CD4+ DNaseI ChINs. This network contained a total of 361,221 unique edges and 158,538 unique nodes. The second was composed of 17 different cell-types resulting in a total number of 728,835 interactions composed by a total of 253,146 unique nodes, of which 20,817 contained promoters**.**

| | Monocytes | Neutrophils | Naive_CD4 + T-Cells |
|---|---|---|---|
| Total Number of Interactions | 171,430 | 147,195 | 329,439 |
| **Monocytes** | **1.000** | **0.485** | **0.270** |
| **Neutrophils** | **0.416** | **1.000** | **0.192** |
| **Naive_CD4+ T-Cells** | **0.518** | **0.429** | **1.000** |
| Macrophages_M0 | 0.528 | 0.360 | 0.256 |
| Macrophages_M1 | 0.492 | 0.334 | 0.243 |
| Macrophages_M2 | 0.514 | 0.357 | 0.258 |
| Megakaryocytes | 0.424 | 0.312 | 0.253 |
| Endothelial_precursors | 0.334 | 0.241 | 0.204 |
| Erythroblasts | 0.435 | 0.341 | 0.259 |
| Foetal_thymus | 0.450 | 0.365 | 0.338 |
| Total_CD4_MF | 0.499 | 0.417 | 0.473 |
| Total_CD4_Activated | 0.517 | 0.437 | 0.467 |
| Total_CD4_NonActivated | 0.502 | 0.424 | 0.475 |
| Naive_CD8 | 0.527 | 0.446 | 0.498 |
| Total_CD8 | 0.518 | 0.428 | 0.478 |
| Naive_B | 0.521 | 0.427 | 0.361 |
| Total_B | 0.548 | 0.448 | 0.381 |

***Table 4.1. The percentage of shared interactions for monocytes, neutrophils and CD4+ T-cells***

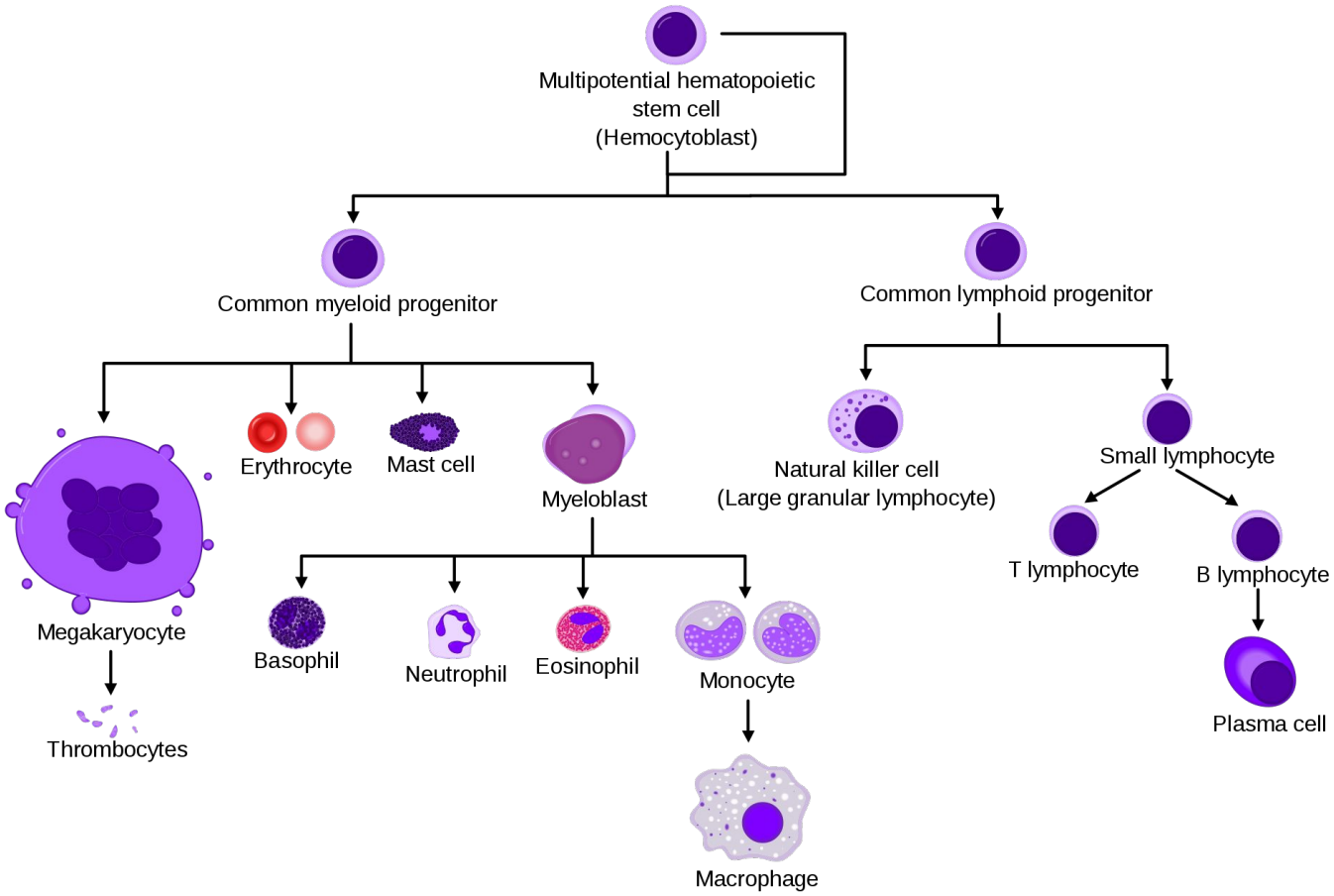



***Figure 4.3. Differentiation lineage of multipotential hematopoietic stem cells.*** Adapted from https://en.wikipedia.org/wiki/Haematopoiesis#/media/File:Hematopoiesis_simple.svg

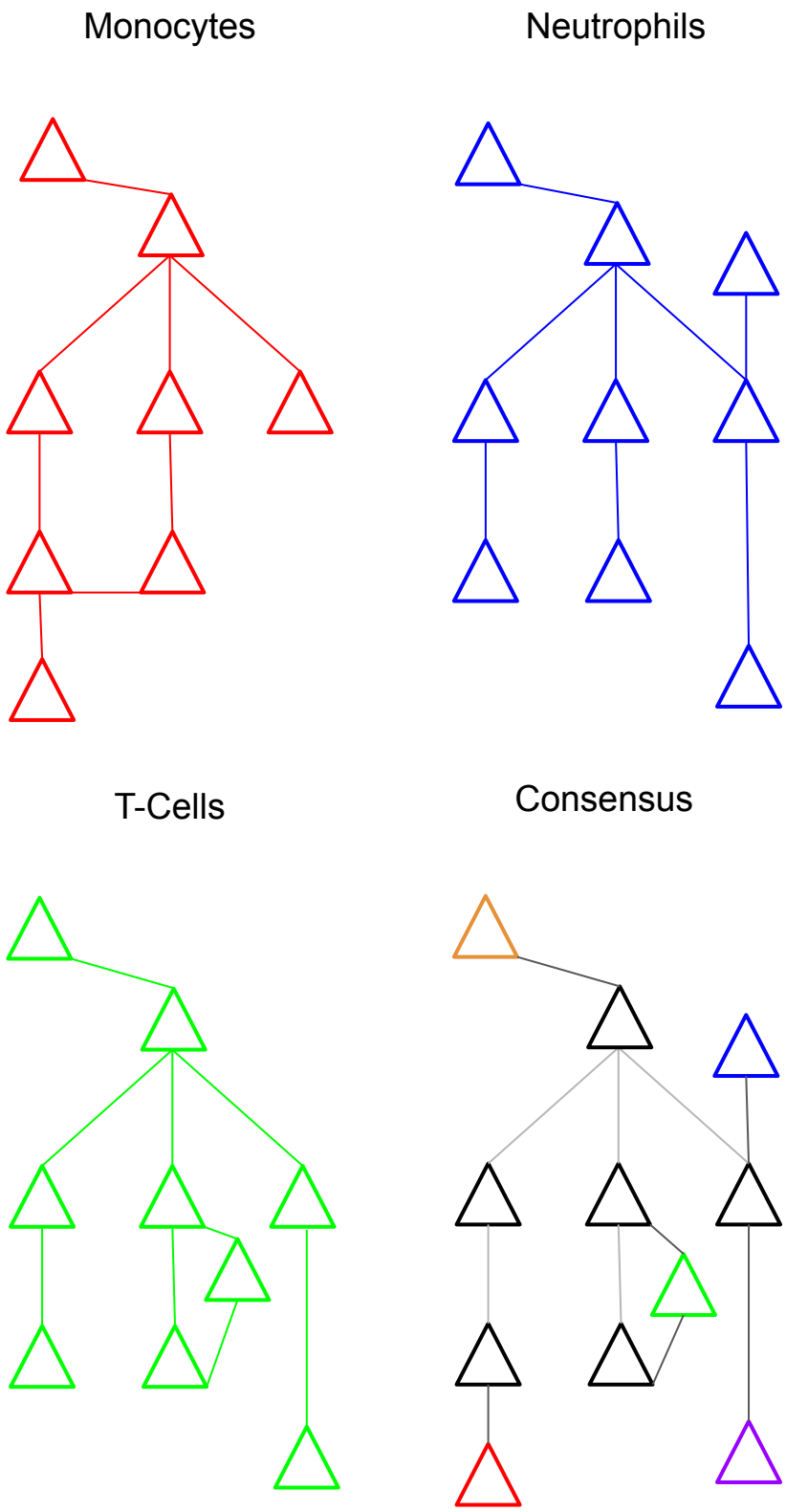



***Figure 4.4. A consensus network of monocytes, neutrophils and CD4+ T-cells***. The networks of monocytes, neutrophils and CD4+ T-cells are collapsed into a single network that contains nodes unique to each cell-type (red, blue and green), nodes unique to two cell-types (purple) and nodes present in all three cell-types (orange).

## 4.2.3 Mapping gene expression data to the network

We implemented our propagation method using the matched RNA-seq gene expression data and ChINs for each of the three cell-types; monocytes, neutrophils and T-cells. Assigning the gene expression to the ChINs is not trivial as genes can map to multiple 3C fragments (**Figure 4.5a**). This can lead to some genes being annotated to multiple nodes leading to the gene expression values being multiplied by the number of nodes they map to. There are two ways in which the RNA-seq data can be assigned to the genic nodes. One is to use the TSS coordinates meaning the gene will only ever map to a single node. The other is to use the whole gene body (WG) coordinates which for large genes can result in mapping to multiple nodes. The TSS mapping was used initially to simplify the model, however the method was designed to incorporate whole gene mapping for future testing.

In the case of WG mapping the problem of gene expression values being multiplied by the number of nodes they map to persists. There are two solutions to this problem. One is to simply divide the gene expression value by the number of nodes the corresponding gene maps to. The other is to introduce the genes as nodes within the network and propagate their values into the ChIN. Here we opted for the latter and treated the gene as individual nodes within the ChINs (**Figure 4.5b**). Each gene node has an associated RNA-seq score that reflects the activity of the gene. This serves as the *a priori* information that is to be propagated from gene nodes (circular nodes that represent the genes) to genic nodes (square nodes that represent the 3C fragments that genes map to) and then across the network to intergenic nodes (triangle nodes) (**Figure 4.5b**). This approach was taken as it could be incorporated into the existing framework more easily. In this chapter we explore the results for TSS mapping. Here, the gene expression values could be assigned directly as the TSS will only map to a single node. However, we retained the same protocol for both the TSS and WG mapping procedures for direct comparison. The results for WG mapping can be seen in **chapter 5**.

The network propagation algorithm uses an undirected network as input. As such, the flow of information, in this case RNA-seq values, are propagated in all directions. Because of this, the genes nodes can receive a proportion of the propagated scores following propagation. As the gene nodes are not part of the true topology of the network and genes are already represented in the network by the genic nodes we developed a two-step propagation strategy. In simple terms, we generated two networks: one containing the genes connected to the genic nodes (circle and square nodes) and the second the chromatin interaction network (square and triangle nodes) (**Figure 4.6a**). In the first step, the RNA-seq scores were propagated to the genic nodes; thus normalising by degree. In the second step, the activity scores in genic nodes appropriated from the first step were propagated across the ChIN (**Figure 4.6b**). The details of how this is achieved using the propagation algorithm are outlined as follows:

We define a set of associations $GF$ representing genes and the chromatin fragments they map to and a set $FF$ of interacting chromatin fragments as defined by the 3C data. A matrix $M$of the cell-type specific RNAseq counts defines the expression profile of each gene $(p1)$ where $MAX$ is the highest count value in $M$. And two networks $G1 = (V1, E1, w1)$ and $G2 = (V2, E2, w2)$ where V1 is the set of unique names in GF, E1 is the set of unique associations in GF, w1 is a weight function denoting the reliability of each interaction and a

fragment-fragment interaction network G2=(V2,E2,w2) where V2 is the set of unique names in FF, E2 is the set of unique associations in FF and w2 is a weight function denoting the reliability of each interaction.

The aim is to impute an activity score (IAS) for $Vfi$ based on gene expression$M$and the topology of the chromatin interaction network $G2$. However, as a single gene can map to multiple chromatin fragments, and many genes can map to a single fragment we employ a two step propagation strategy . Here we propagate the values of $M$from $Vg$to $Vfg$across $G1$ in step one and from $Vfg$to $Vfi$ across $G2$ in step two. Here, $Vfg$acts as as the intermediary between $G1$and $G2$where $V1 \cap V2 = \{Vfg : Vfg \in V1 \; and \; Vfg \in V2\}$.

Following propagation the intergenic nodes can be compared relative to one another by their imputed activity scores (**IAS**) reflecting their probability of being an enhancer node (**Figure 4.6c**).

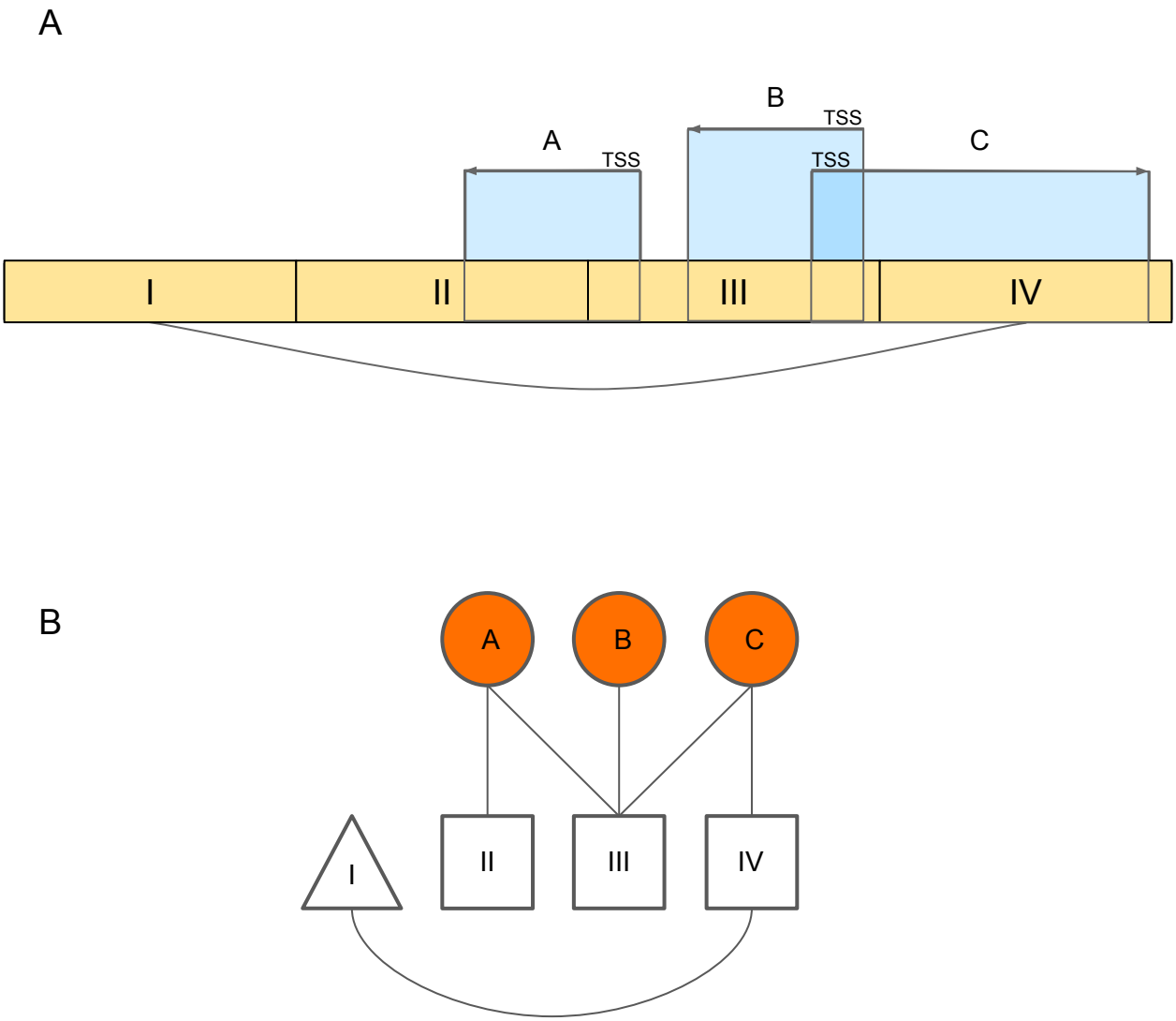



***Figure 4.5. Mapping of genes to 3C fragments.*** A) Genes can be mapped to a single 3C fragment using the TSS or by the whole gene body which results in some genes mapping to multiple 3C fragments. B) Genes are represented as circles, the 3C fragments they map to as squares while intergenic 3C fragments are represented as triangles.

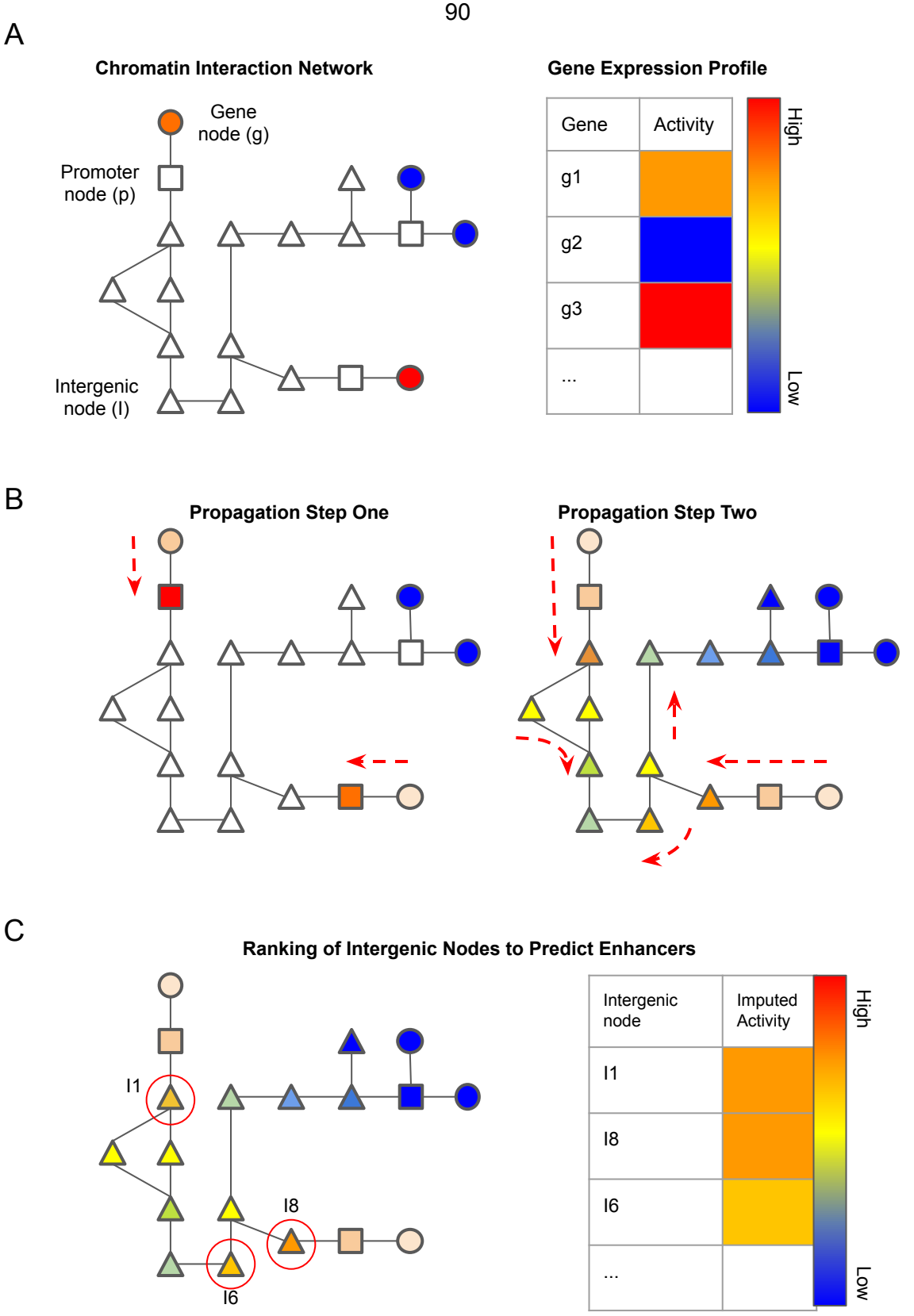
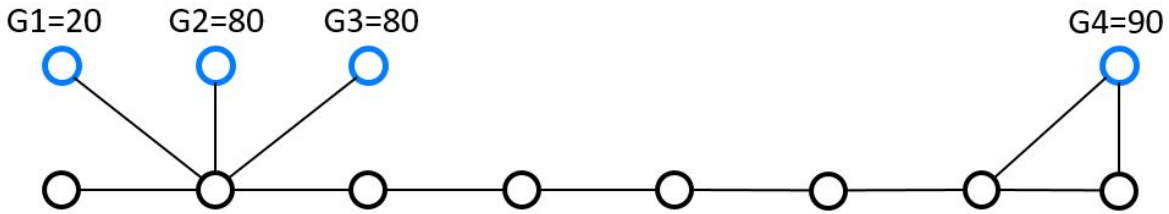



**Figure 4.6.** ***Schematic diagram of the network propagation used to impute activity values at intergenic nodes.*** **A)** Genes are mapped to nodes representing genic chromatin fragments. Each gene has an associated gene activity value determined by RNA-seq data. **B)** Gene activity is propagated from gene nodes to genic chromatin nodes in propagation step one. Activity scores are then imputed in intergenic chromatin nodes by propagating the scores from genic chromatin nodes. **C)** Ranking of non-genic nodes by the imputed activity score to identify high confidence enhancer nodes.

## 4.2.4 Parameter Optimisation to Reflect Biologically relevant Enhancer Distances

There are two hyper-parameters that can influence the propagation of gene expression across the network; alpha and the number of iterations. The alpha score is often called the insulating parameter, and restricts the amount of propagation across the network. The number of iterations determines how many times the algorithm is executed to propagate gene expression across the network. To be able to apply our model to various 3C networks and cell-types we initially optimised the parameters to reflect the general biological properties of enhancer promoter interactions rather than optimising the parameters for each individual ChIN. Here we limited the propagation of gene expression to nodes at 2 and 3 degrees away the most and maximised the difference in IAS between the nodes (**Figure 4.7**). This would ensure that nodes that are proximal to genic nodes were prioritised ahead of nodes located at larger differences that are less likely to impose a regulatory effect on a gene.

In the example network given there are 3 genic nodes (F2, F7 and F8) and 5 intergenic nodes (F1, F3, F4, F5 and F6). Each gene connected to a genic node is given an arbitrary value between 20 and 90 and these are then propagated across the network with various settings. With very strict settings such as with 2 iterations and an alpha of 0.1 the majority of the expression is retained in the intergenic nodes that are direct neighbours of the genic nodes, F2, F7 and F8, with very little spreading to the intergenic nodes that are connected indirectly. With more lenient settings such as with 10 or 100 iterations and an alpha of 0.5 we can see that while the indirect intergenic nodes receive a higher value, nodes such as F5 and F6 receive the same score. This is despite each of these nodes being located at different distances to the genic nodes. The ‘goldilocks’ parameters were found to be 2 iterations with an alpha value of 0.2. This alpha value retained high IAS in nodes near genic nodes. This is useful considering enhancers will tend to be proximal to genes in 3D-space in order to activate gene expression. It also maximised the differences in the scores between each of the nodes to better resolve intergenic nodes of different topology and proximity to genic nodes. These parameters were again shown to be optimal when testing the performance in predicting enhancer nodes on the real ChINs (see 4.4.9).




| | F1 | F2 | F3 | F4 | F5 | F6 | F7 | F8 |
|---|---|---|---|---|---|---|---|---|
| Niter=2 α=0.1 | 0.88 | 29.6 | 0.89 | 0.06 | 0.03 | 0.39 | 7.93 | 7.74 |
| **Niter=2 α=0.2 | 2.93 | 49.5 | 2.99 | 0.42 | 0.23 | 1.40 | 14.1 | 13.3 |
| Niter=10 α=0.5 | 14.2 | MAX=69.8 | 23.5 | 19.8 | 17.7 | 17.4 | 27.3 | 18.6 |
| Niter=100 α=0.5 | 11.2 | 56.4 | 22.5 | 22.5 | 22.4 | 22.4 | 33.5 | 22.3 |

***Figure 4.7. Propagation with four different parameters across a toy network.*** RNA-seq from G1, G2, G3 and G4 are propagated using four different parameters to model the effects of low to high propagation. (Data and figure produced by Luca Guidice)

## 4.2.5 Annotation of the nodes using chromatin states

In order to identify whether IAS is able to identify enhancers we annotated the networks using a chromatin state model from the publication of Carrillo-de-Santa-Pau et al, (Carrillo-de-Santa-Pau et al. 2017). Using a multivariate hidden Markov model, ChromHMM (Ernst and Kellis 2017) provides an 11 state classification; transcription (2 states), heterochromatin (4 states), repressed promoter (1 state), enhancer (2 states) and active promoter (2 states) (**Figure 4.8a**). The genome is segmented into these states by learning the various occupancy combinations of 6 histone marks, H3K4me1, H3K27ac, H3K4me3, H3K27me3, H3K9me3 and H3K36me3. Enhancers, for example, are characterised by H3K4me1 such as in state 8 (**Figure 4.8a**) and sometimes in combination with H3K27ac as in state 9 (**Figure 4.8a)**. As discussed previously, regions with a combination of H3k4me1, H3K27ac and H3K4me3 can represent enhancer regions (**See Chapter 1: Histone modifications**) however this state was not included as an enhancer in the initial analysis.

The chromatin state profile of the genomes of each of the cell-types show that heterochromatin is by far the most abundant chromatin state, covering up to 85% of the entire genome (**Figure 4.9**). Interestingly, the chromatin profile of monocytes (**Figure 4.9a**) contains ~5% less heterochromatin than neutrophils (**Figure 4.9b**) and T-cells (**Figure 4.9c**). Instead, the monocytes contain a higher percentage of both enhancer and transcription chromatin states suggesting that monocytes contain more active chromatin. They do not, however, contain a higher percentage of active promoter states, nor a lower percentage of repressed promoter states meaning the increase in active chromatin is not necessarily a consequence of a higher number of active genes.

The PCHi-C dataset enriches specifically for promoters and the chromatin they interact with. To understand how the chromatin profile of this dataset diverges from the genome wide chromatin profile we then annotated the nodes with the chromatin states and calculated the percentage composition of chromatin states for each node. In each cell-type we observe a decrease in the percentage of heterochromatin states, most notably in monocytes where we observe a percentage point drop of 11.44 (**Figure 4.10a**); twice that of neutrophils with 5.42 (**Figure 4.10b**) and 26 times larger than T-cells with a 0.44 decrease (**Figure 4.10c**). This reduction is largely shifted towards a slightly higher enrichment of chromatin with transcription states and a much larger enrichment of enhancer states. Results show that the percentage of the total genome labelled as enhancers stands at 3.75% for monocytes, 1.31% for neutrophils and 2% for CD4+ T-cells. This percentage increases in the chromatin regions captured by PCHi-C with 12.23% for monocytes, 4.18% for neutrophils and 4.8% for CD4+ T-cells. These results confirm that the ChINs contain enhancers that can now be used to validate the predictions of our model.

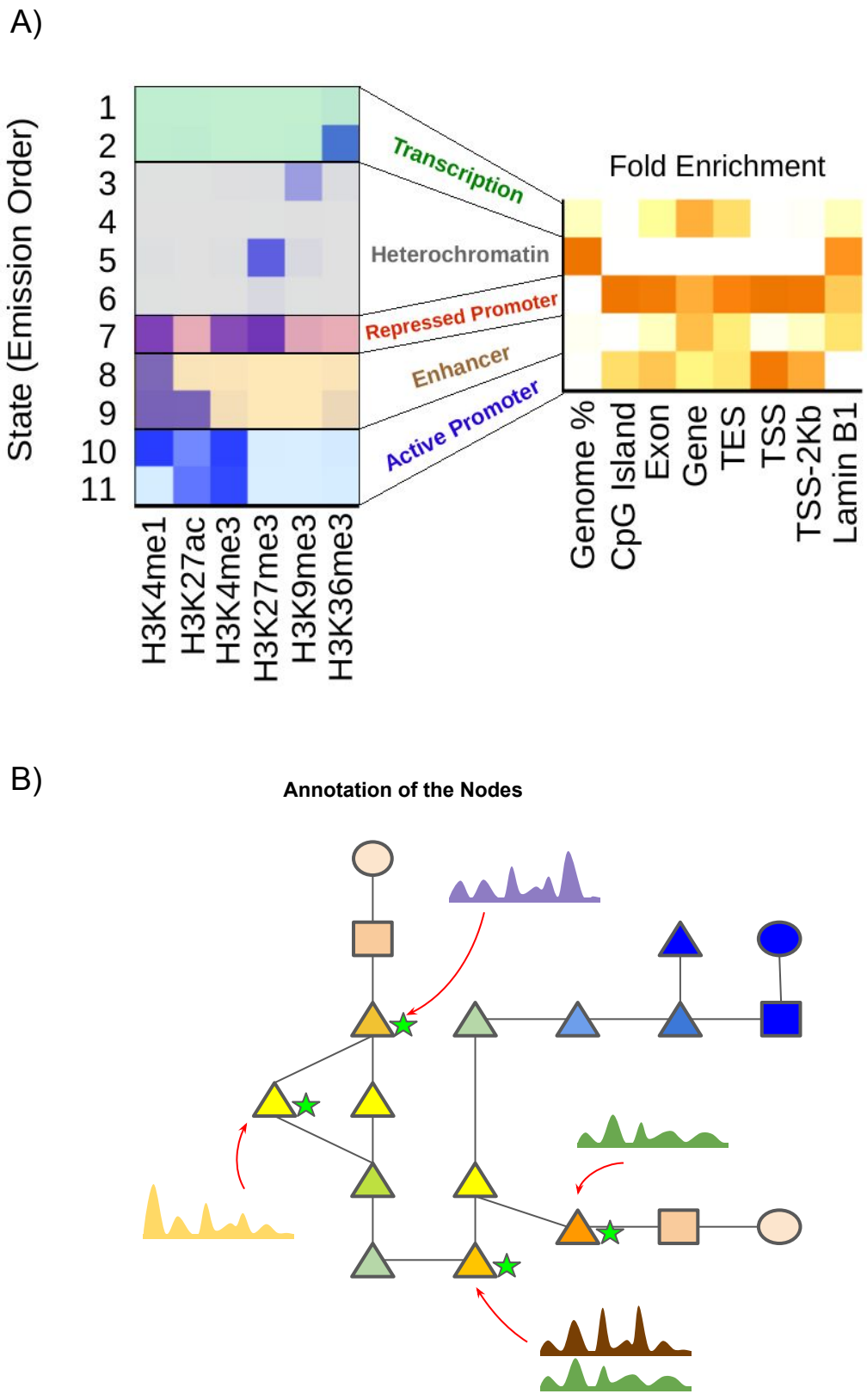



***Figure 4.8. Annotation of nodes using chromatin states defined by combinations of histone modifications.*** **A)** Publicly available 11 chromatin state model based on six histone modifications. (Figure from Carrillo-de-Santa-Pau et al. 2017) **B)** Chromatin states were annotated to the nodes based on the genomic coordinates. Nodes labelled with enhancer chromatin states are denoted with a green star.

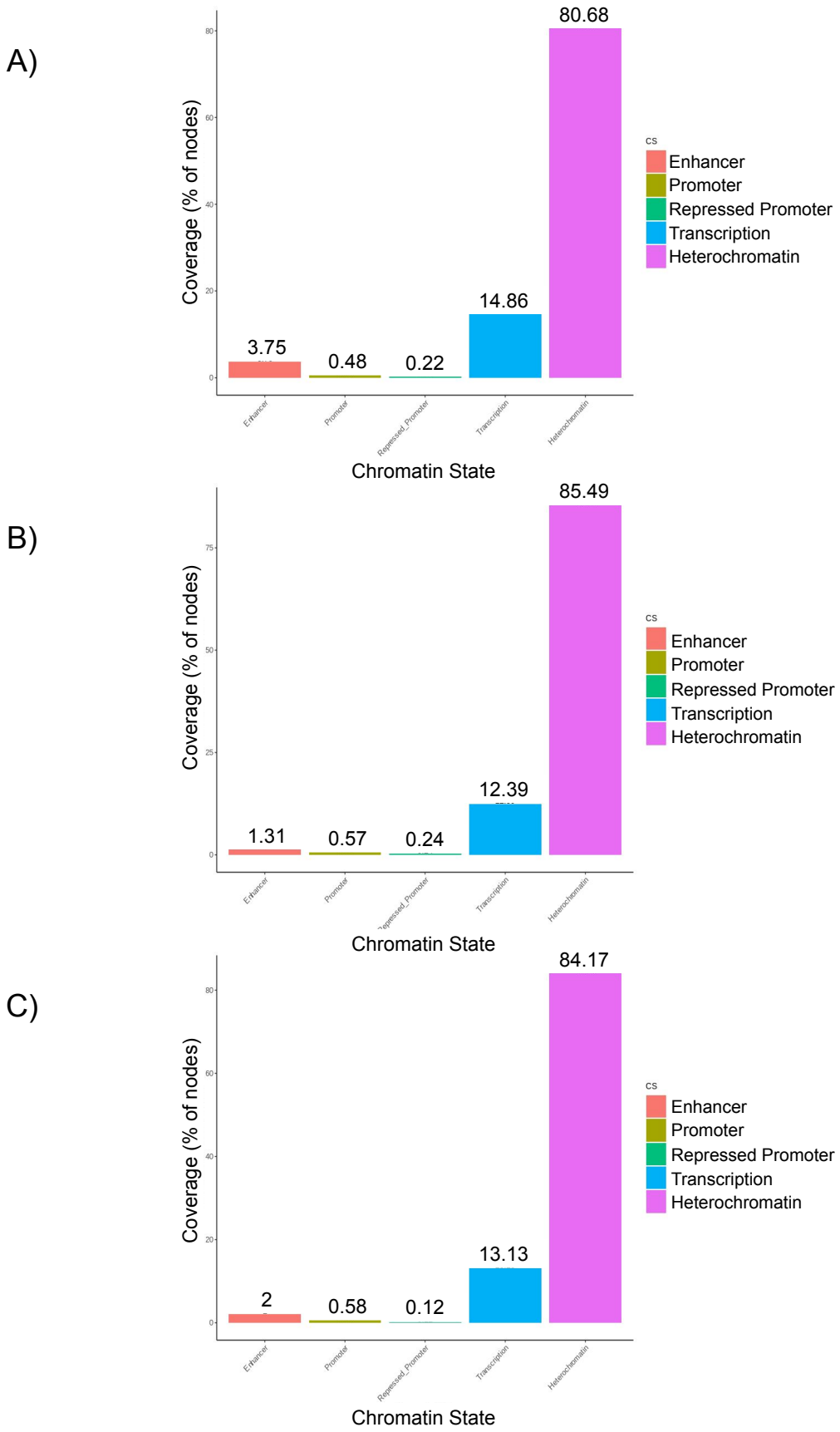



***Figure 4.9. Genome wide chromatin state composition for monocytes, neutrophils and CD4+ T-Cells.*** The percentage of each chromatin state genome wide for monocytes (A), neutrophils (B) and T-Cells ©.

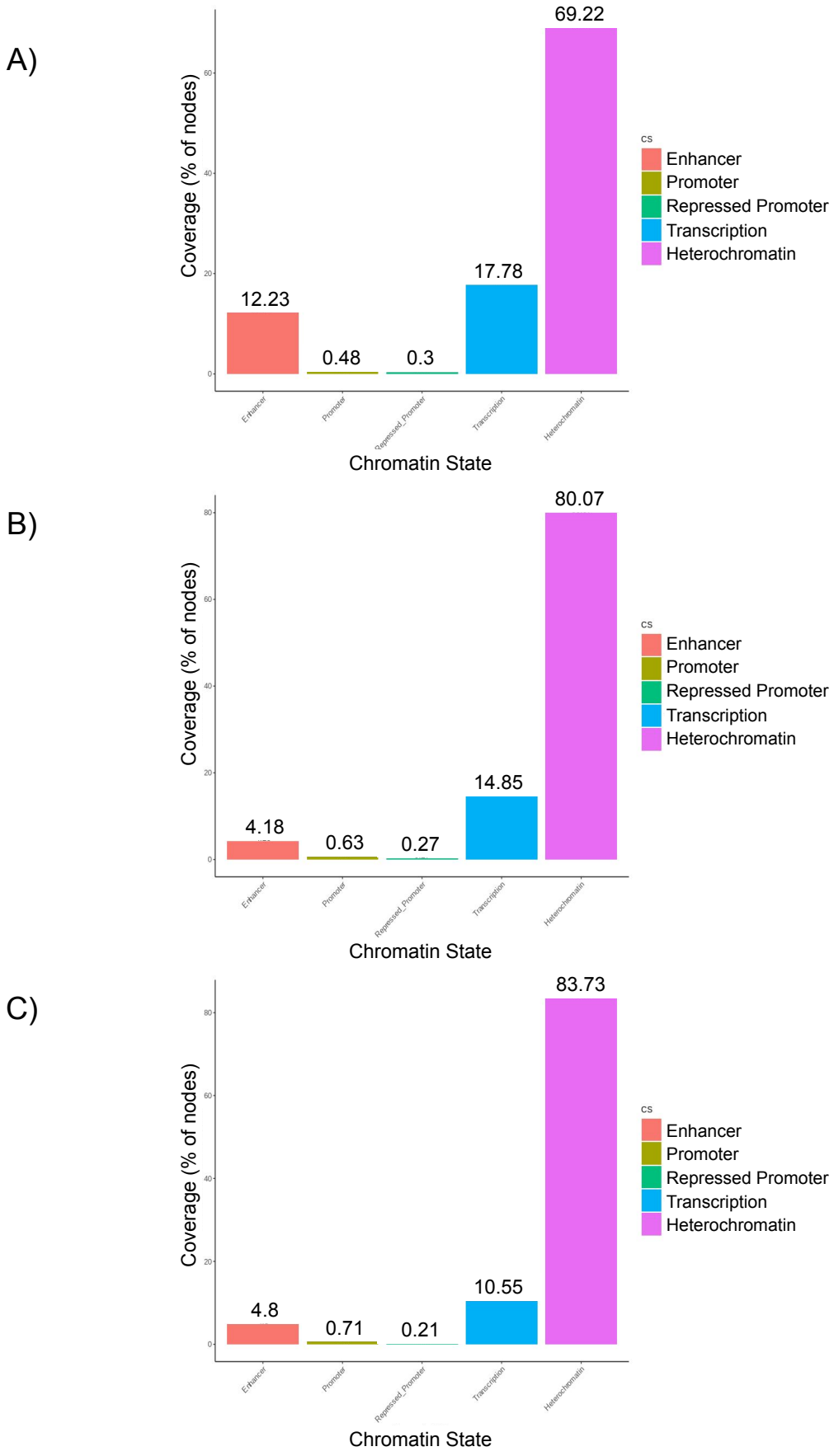



***Figure 4.10. PCHi-C chromatin state composition for monocytes, neutrophils and CD4+ T-Cells.*** The percentage of each chromatin state in loci captured in the PCHi-C data set for monocytes (A), neutrophils (B) and T-Cells (C).

### 4.2.6 Initial results

We generated an initial set of results using the monocyte derived network due to the high enrichment of enhancers as defined by the chromatin states. Genes were mapped to the network as described previously using the TSS coordinates and propagation of the gene expression was carried out using two iterations and an alpha of 0.2 on the monocyte specific network. We first tested how well the IAS metric correlated with each of the chromatin states: Do nodes with a higher percentage of enhancer states receive a higher IAS from the propagation? Intergenic nodes were selected and the IAS and chromatin state coverage were plotted for each state (**Figure 4.11**). The correlation between IAS and each chromatin state was then calculated where 1 indicates a perfect correlation and -1 a perfect anti-correlation. For the two enhancer chromatin states E8 and E9 we observed a correlation of 0.024 and 0.015, respectively. While for the repressed heterochromatin state E5 we observe an anti-correlation of -0.009. While a pattern emerges with a correlation between IAS and enhancer chromatin states and an anti-correlation with heterochromatin states the evidence is weak.

We again tested our model using the same parameters and monocyte gene expression, this time using the consensus network of monocytes, neutrophils and T-cells (**Figure 4.12**). In this experiment the correlation between IAS and the two enhancers states E8 and E9 increased by 0.038 and 0.028 to 0.062 and 0.043, respectively. While the anti-correlation between IAS and E5 increased by -0.001 to -0.01. Again, a pattern of correlation with the enhancer states is seen, accompanied by an anti-correlation with the repressed heterochromatin state E5. The use of a consensus network appears to marginally improve the correlation between IAS and enhancer chromatin states. We therefore opted to use the consensus three cell-type network for the next set of analyses.

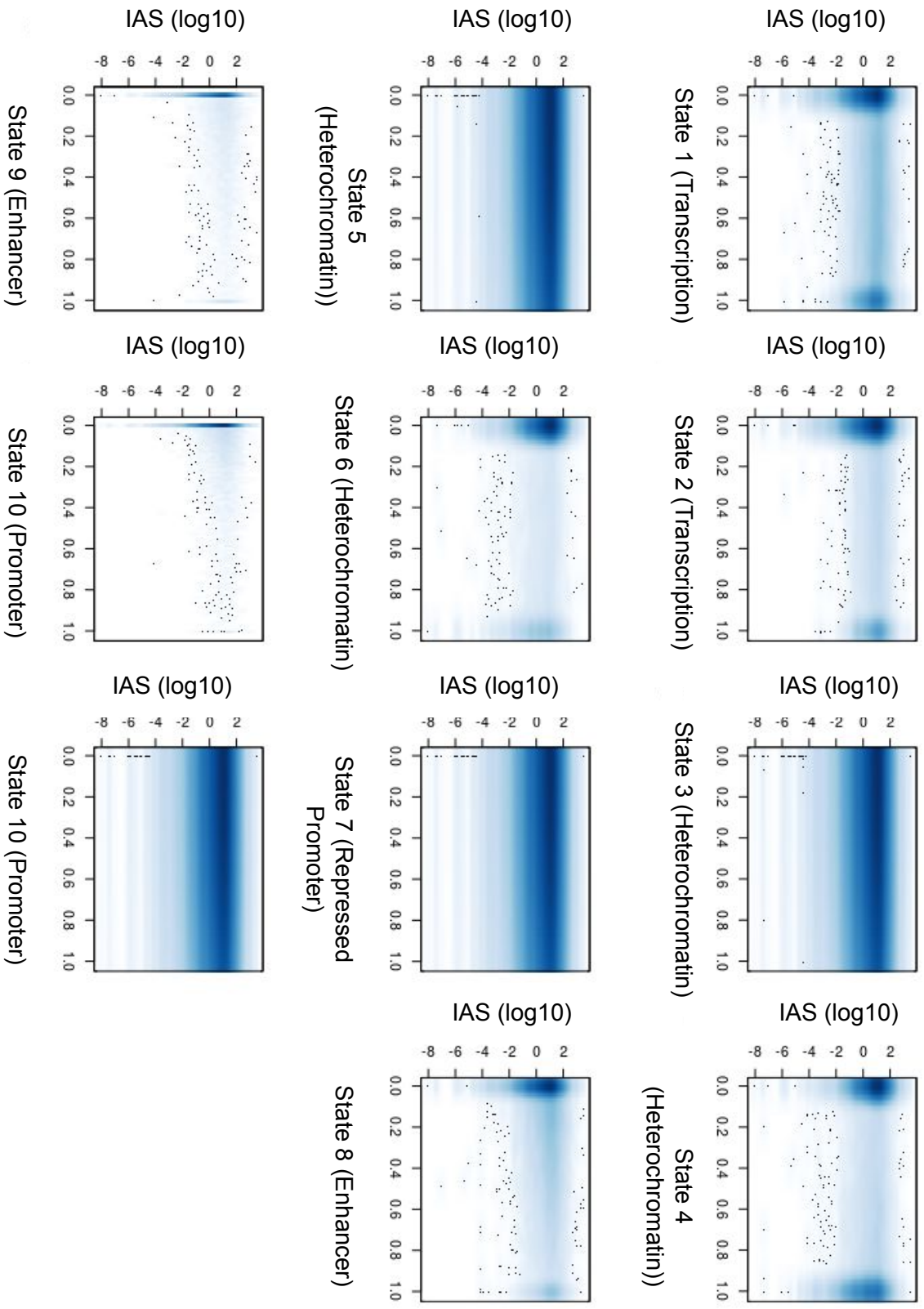



***Figure 4.11. Scatter plots of IAS vs the percentage of each chromatin state across the nodes of the cell-type specific networks.***

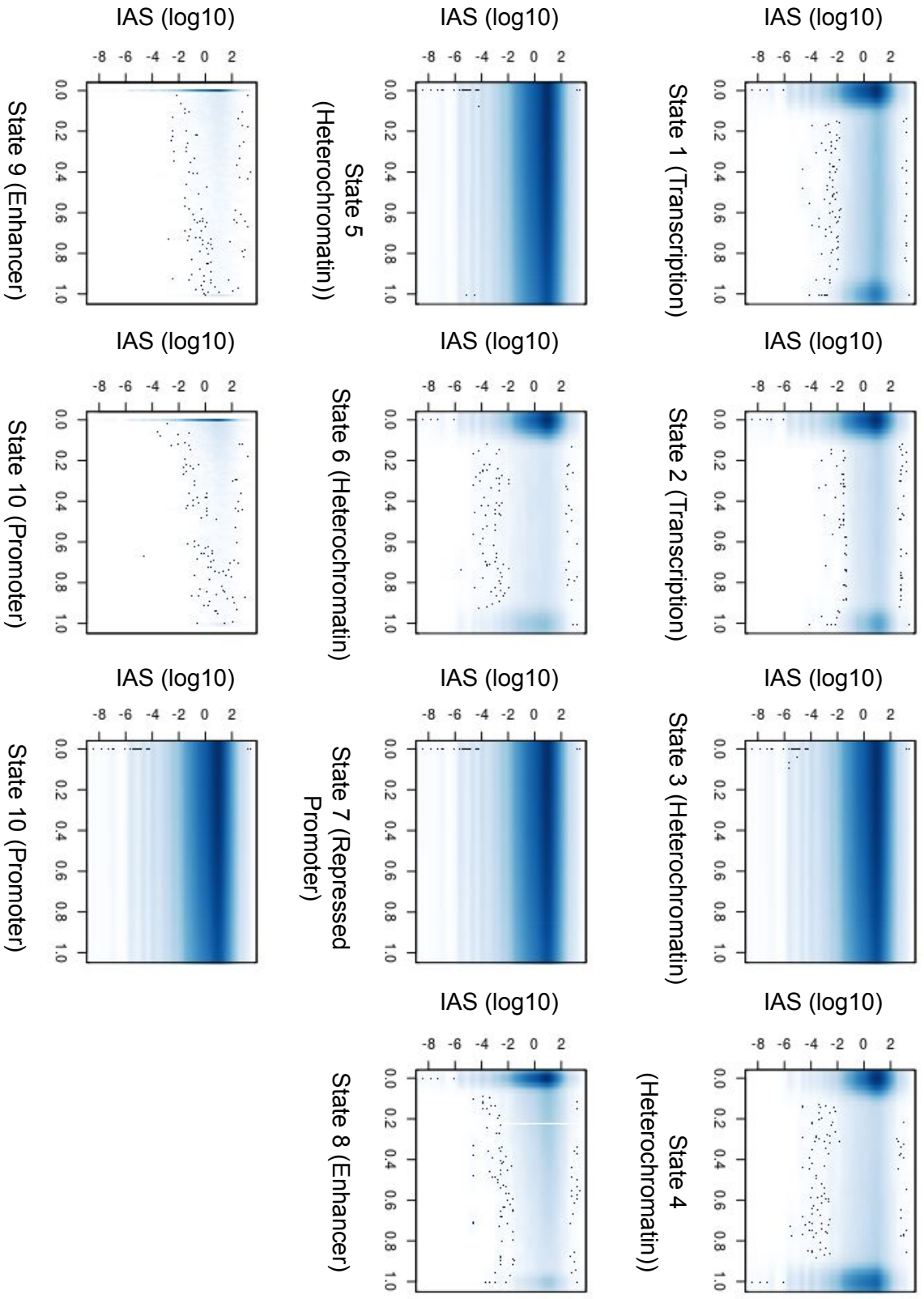



***Figure 4.12. Scatter plots of IAS vs the percentage of each chromatin state across the nodes of the three cell-type consensus networks.***

## 4.2.7 Labelling of nodes into active and inactive

Despite showing a promising difference between the enhancer states and heterochromatin states the actual correlation between each of the states and IAS is relatively low. While the low correlation can be due to inefficiencies in the model, both the labelling of the nodes and the testing methods themselves may also be less than optimal. We noticed in the previous analysis that the distribution of the coverage of chromatin states for the fragments appeared to be bimodal.

Upon further investigation, for some chromatin states including the enhancer states, the average distribution of the states across the nodes are bimodal (**Figure 4.13**): the majority of nodes with these states cover either less than 10% or more than 90% of the chromatin. In this case, Spearman correlation would be a more appropriate measure of correlation due to the non-normal distribution. The next problem arises from the non-uniform chromatin fragment sizes from the PCHi-C dataset. Because we are measuring the correlation between the percentage coverage of each chromatin state and IAS nodes representing, for example, a large 10kb fragment with 10% coverage of E8 contains a larger enhancer region than a small 1kb fragment with 10% coverage of E8. This is further compounded by the fact that nodes representing larger fragments tend to have a higher degree, or number of connections (**See chapter 3.**). This results in this group of nodes generally receiving a higher IAS value than smaller fragments (**See chapter 5**). In addition to this, we also observe a slight correlation between IAS and other states, most notably E10 which can be defined as a hybrid E-promoter or less active enhancers (**See chapter 1: Enhancers**). This highlights both the ambiguity in defining enhancers based on classical histone modifications and the potential to mislabel intergenic nodes as non-enhancers when using a single method for validation.

For these reasons we collapsed the promoter, enhancer and transcription states into a single ‘active’ state and the heterochromatin states and repressed promoter states into an ‘inactive’ state. Active and Inactive nodes were determined by a simple majority, i.e. if more than 50% of the node was active it was labelled as active. Using this method of labelling we were able to minimise the effects of variable enhancer distribution and fragment sizes, while reducing the number of false negatives. Improvement in the labelling of nodes is investigated and discussed in further detail in **chapter 5**.

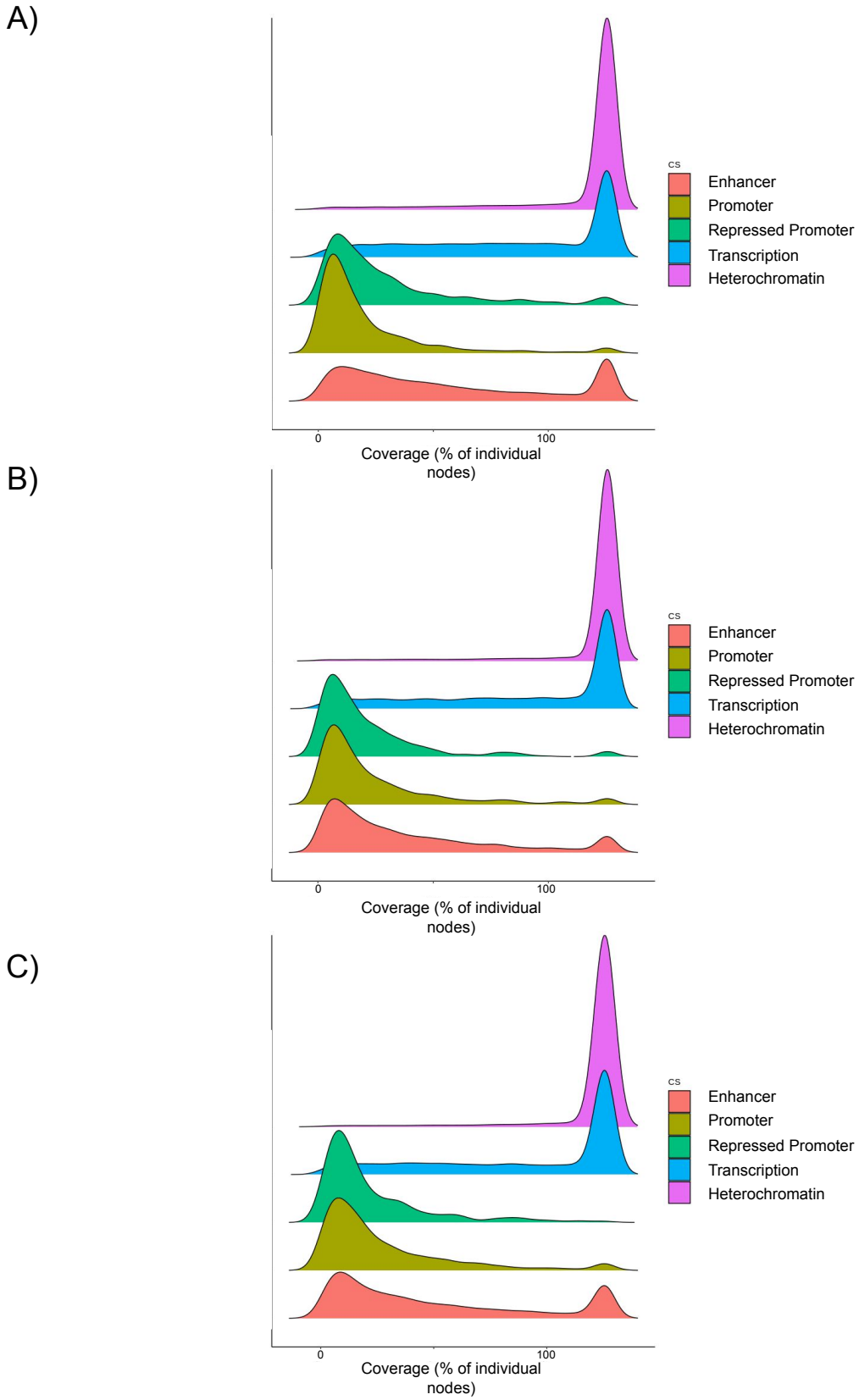



***Figure 4.13. Distribution of chromatin state annotation across the nodes of the monocyte, neutrophil and T-cell networks*.** The distribution of each chromatin state for every node generated from the fragments of the PCHi-C dataset for monocytes (A), neutrophils (B) and T-Cells (C).

## 4.2.8 IAS is enriched in nodes labelled as active

Using the active and inactive labels we then compared the relative levels of IAS in each of the two groups and extended the analysis to include the cell-type specific networks of neutrophils and T-cells. Here we observe that IAS in nodes annotated as being active are significantly higher (Kolmogorov-Smirnov test Monocytes p.value = <2.2e-16, Neutrophils p.value = <2.2e-16 and T-Cells p.value = <2.2e-16). The IAS is 2.89, 4.34 and 4.04 fold higher than the inactive nodes for monocytes, neutrophils and t-cells, respectively. Higher levels of IAS in active intergenic nodes show that our model is able to discriminate between active and inactive nodes (**Figure 4.14**).

These results confirm that our model is, to some degree, able to identify active nodes as defined by chromatin states with gene expression and 3C data with no other *a priori* knowledge or input. This is achieved by leveraging the underlying relationship between gene expression and the complex interactions of chromatin. One of the relationships that our model is able to take advantage of is the tendency of chromatin to interact with regions of a similar composition. When testing the correlation between genic nodes, which are inherently active, and both active and inactive intergenic nodes we see this effect. Across all three cell-types a positive correlation is seen in the interactions between genic nodes and active intergenic nodes and a negative correlation between genic nodes and inactive intergenic nodes (**Table 4.2**). This means that the propagated gene expression data, by virtue of the network architecture, is more likely to reach an active intergenic node. The underlying relationships that our model is able to leverage are explored and discussed in further detail in **chapter 5**.

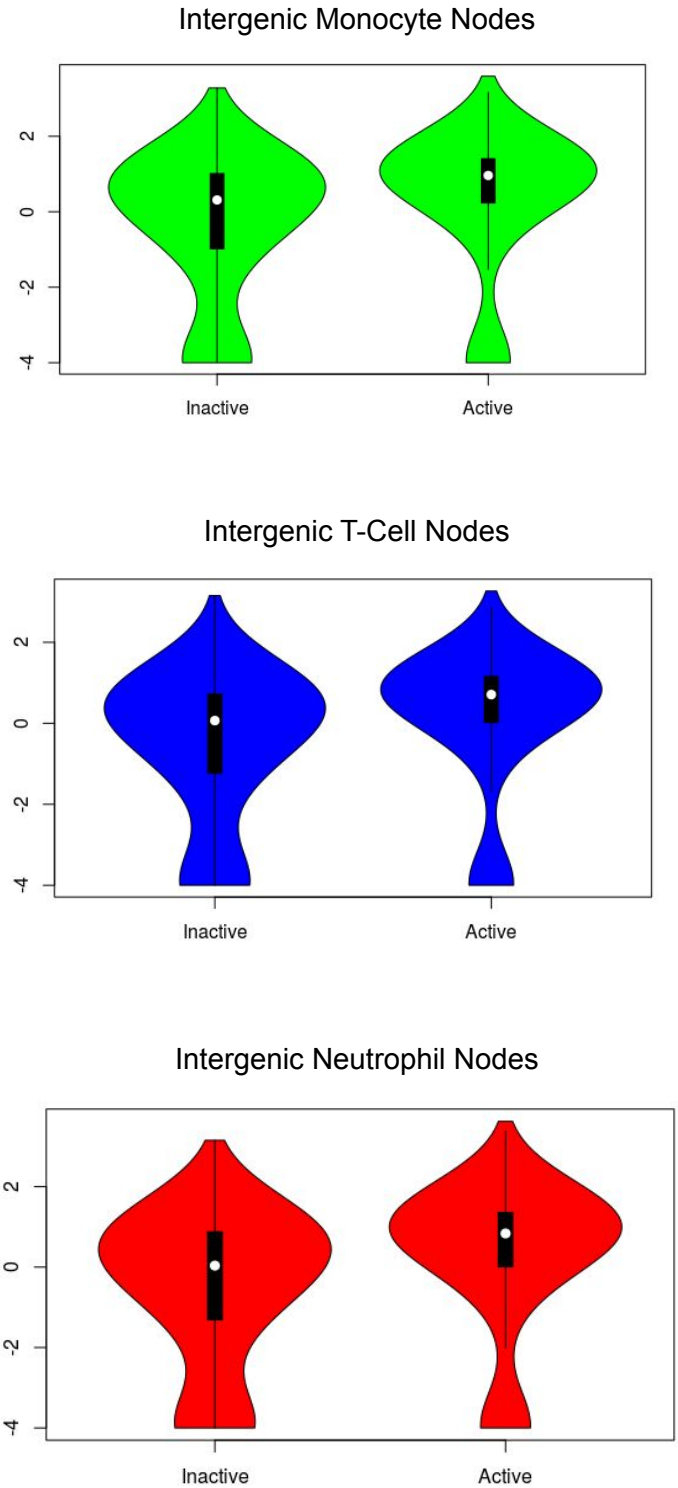



**Figure 4.14. IAS in nodes annotated as active and inactive.** Active and inactive intergenic nodes are defined by the percentage of chromatin states for transcription, enhancers and promoters covering >=50% of the chromatin represented by the node. Inactive nodes are defined by the percentage of of chromatin states for heterochromatin and repressed promoter covering <50% of the chromatin represented by the node. A significant difference is observed between each of the groups (Kolmogorov-Smirnov test Monocytes p.value = <2.2e-16, Neutrophils p.value = <2.2e-16 and T-Cells p.value = <2.2e-16).

| Cor.test on fragment chromatin state interactions | Intergenic Active | Intergenic Inactive |
|---|---|---|
| Monocytes | 0.239 | -0.239 |
| Neutrophils | 0.252 | -0.252 |
| T-Cells | 0.251 | -0.251 |

***Table 4.2. The correlation between active genic nodes and intergenic active and intergenic repressive nodes in the three cell-type networks for monocytes, neutrophils and T-cells.*** The Pearson correlation between active and inactive nodes Active and inactive nodes are defined by the percentage of chromatin states for transcription, enhancers and promoters covering >=50% of the chromatin represented by the node. Inactive nodes are defined by the percentage of of chromatin states for heterochromatin and repressed promoter covering <50% of the chromatin represented by the node.

## 4.2.9 IAS is predictive of active nodes

Following on from the previous results we carried out a more comprehensive analysis, including the cell-type specific and 17 cell-type consensus networks along with two additional parameters for propagation. RNA-seq values mapped to nodes containing the TSS of the corresponding gene and for each cell-type, network and parameter combination IAS was calculated. Initially this method of mapping RNA-seq values to nodes was chosen to avoid the multiple mapping of RNA-seq values of genes that span more than one node in the consensus networks. We used the active and inactive labels as described previously to validate our results.

Following propagation ROC curves were produced and the area under the curve (AUC) calculated to determine the ability of the model to identify enhancer nodes across the cell-type specific networks, the three-cell consensus network and the 17 cell-type consensus network **(Table 4.3)**. Results show that in all three types of networks and in all cell-types the propagation set with an alpha value of two and an iteration value of two returned better performance. Interestingly, the consensus network made up of the interactions of the monocytes, neutrophils and CD4+ T-Cells yielded a better AUC score than the individual networks.

Although this is a commonly used metric, an often overlooked caveat is that when the labelling set is imbalanced, the results of ROC curves can be biased. In this case the percentage of nodes labelled as active is 48%, 37% and 40% for monocytes, neutrophils and T-cells. Because of this, the predictive performance is likely underestimated for neutrophils and to a lesser extent T-cells, using the AUROC metric. A more comprehensive analysis of IAS in predicting enhancer nodes is provided in **chapter 5**.

| | Cell-specific | Consensus (3 cell-types) | Consensus (17 cell-types) |
|---|---|---|---|
| Mono Niter2a1 | 0.604 | 0.648 | 0.645 |
| Mono Niter2a2 | 0.605 | 0.650 | 0.648 |
| Mono Niter10a5 | 0.603 | 0.640 | 0.628 |
| Neut Niter2a1 | 0.617 | 0.658 | 0.654 |
| Neut Niter2a2 | 0.617 | 0.660 | 0.656 |
| Neut Niter10a5 | 0.617 | 0.656 | 0.644 |
| T-cell Niter2a1 | 0.649 | 0.654 | 0.655 |
| T-cell Niter2a2 | 0.650 | 0.656 | 0.656 |
| T-cell Niter10a5 | 0.646 | 0.649 | 0.640 |

***Table 4.3. Predictive performance measured by the area under the ROC curve.*** The area under the ROC curve was calculated to measure the predictive performance of our model across 3 cell-type specific networks for monocytes, neutrophils and T-cells as well as, 3 cell-type and 17 cell-type consensus networks. The expression data of monocytes, neutrophils and T-cells were propagated across their respective cell-type specific networks as well as both consensus networks. For each propagation we also tested three parameters (first column).

## 4.2.10 Mapping RNA-seq to nodes by the coordinates of the TSS or whole gene body differentially affects the imputed score in intergenic nodes

Gene enhancers typically target the promoter by the transcriptional start site. For this reason we initially utilised this mapping procedure as outlined in **chapter 3** to assign RNA-seq values to the nodes. However, we can also map RNA-seq values to nodes using the coordinates of the whole gene. Here we revisit the mapping procedure and investigate the effects of using each type of mapping on a toy network.

The mapping function $Y: V \rightarrow [0, MAX]$ defines the mapping of gene expression from an RNA-seq data set to genic nodes in the propagation algorithm. This requires the annotation of the genic nodes using gene data from Ensembl (**See Chapter 3:Methods**). One notable consideration for this is the gene size in relation to the size of the 3C fragments. For the DNaseI capture experiment the median fragment size is 2.8kb. Compare that to the median gene size in mice of 5.2kb we start to see a potential problem in the assignment of genes to genic nodes. For example, take gene X with an associated RNA-seq value of 100 where the TSS maps to a single node, node A (**Figure 4.15a**). In this scenario genic node A would start with a value of 100 if the values were assigned directly based on the presence of that gene. This offers the most simplistic method of assigning RNA-seq values to genic nodes. However, when using the whole gene body, genes can map to multiple nodes. Take gene Y , with an associated RNA-seq value of 100 that maps to two gene nodes, nodes A and D (**Figure 4.15b**). In this scenario nodes A and D would each start with a value of 100; because gene Y mapps to two nodes it's associated RNA-seq value has been doubled. This poses the potential problem of artificially inflating the RNA-seq values for genes that map to multiple nodes. In the mESC DNaseI ChIN for example, there are 23,506 genes that map to 117,546 nodes meaning that each gene maps to 5 nodes on average.

In order to test the effects of TSS and WG mapping we propagated the expression of genes X and Y across the toy network with one iteration and an alpha of 0.2 using both mapping types (**Figure 4.15c**). The imputed activity scores were then used to rank the intergenic nodes for each mapping procedure . As expected the WG mapping resulted in slightly higher final imputed activity scores for the intergenic nodes. However, the final ranking of the intergenic nodes were identical. These results show that the mapping procedure, whether TSS or WG, does not have an effect on the ranking of intergenic nodes when the gene body node D is only connected to the TSS node A. This is because all of the intergenic nodes are still in topologically identical positions in the network in relation to the source of where the values are being propagated from making the duplication of RNA-seq score arbitrary. To understand this concept we can calculate the distance of each intergenic node to the gene node using the shortest path measure. For the intergenic nodes B and C the shortest path to the genic nodes A and D are 1 and 2 respectively, while for E and F it is 2 and 4. This is reflected in the final rankings of these nodes whereby B and C are ranked joint 1st, with E and F ranked joint second.

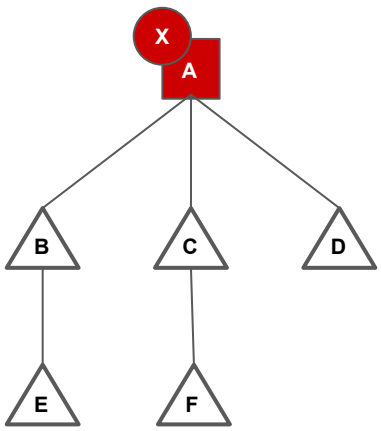
We then considered a scenario whereby gene Z maps to nodes A and E (**Figure 4.15d**) . In this case, the shortest path to the gene nodes would change depending on the use of TSS

or WG assignment (**Figure 4.15d**). Nodes B, C and D all have a shortest path of 1 to node A (TSS) while node F has a shortest path of 2. While to node E (Gene body) the distance is 1, 3, 3, and 4 for B, C, D and E, respectively. Here we can see that using either TSS or WG mapping results in different rankings of nodes B and C. For TSS mapping C is ranked joint 1st with B whereas for WG mapping node B is ranked ahead of node C.

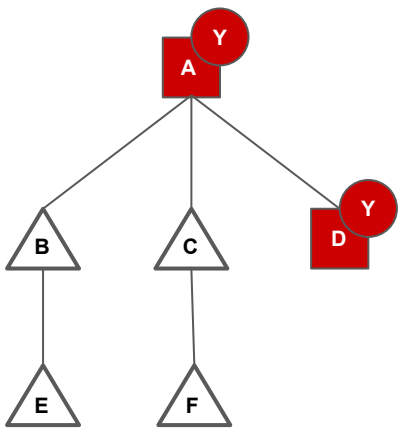
We have shown how using WG mapping can separate nodes more clearly than TSS mapping based on their connectivity with genic nodes. Here we discuss the framework developed to assign the RNA-seq scores to each genic node using the WG mapping process. Using WG mapping can result in inflated values for genes that span more than one node compared to smaller genes and therefore creating a bias with respect to large genes following propagation. There are technical solutions to solve this problem, such as dividing the RNA-score associated with each gene by the number of nodes it maps to. Instead of adding a new normalisation step into our workflow we consolidated the normalisation process into our existing framework. The normalisation process can also be imagined as normalising the RNA-seq values by the number of connections it has to nodes. As such we included genes as nodes in the network where their edges connected them to the nodes they mapped to. We then used the propagation algorithm to propagate the RNA-seq value for each gene from the gene node to the genic node(s) that they map to (**Figure 4.16**).

Finally, we applied these principles to the mESC DNaseI ChIN. The results confirm that WG mapping (AUPRC = 0.650) results in an improvement in the classification of enhancer nodes over the use of TSS mapping (AUPRC = 0.629) (**Figure 4.17**). We therefore opted to use the WG mapping as the default mapping procedure for 3D-SearchE.

A)



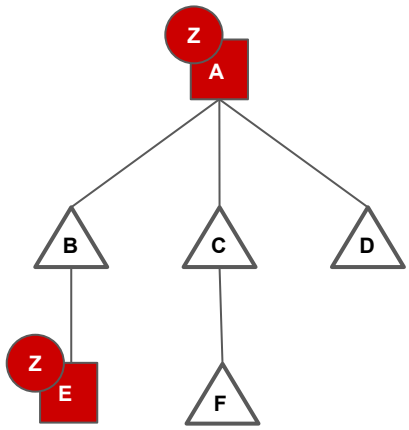

B)




C)

| Expression | | | | | | |
|---|---|---|---|---|---|---|
| **A** | **B** | **C** | **D** | **E** | **F** | |
| 82.21 | 5.59 | 5.59 | 5.48 | 0.56 | 0.56 | $X_{TSS}$ |
| 98.66 | 6.71 | 6.71 | 86.58 | 0.67 | 0.67 | $Y_{WG}$ |
| N/A | 1st | 1st | N/A | 2nd | 2nd | $X_{TSS}$ |
| N/A | 1st | 1st | N/A | 2nd | 2nd | $Y_{WG}$ |

D)



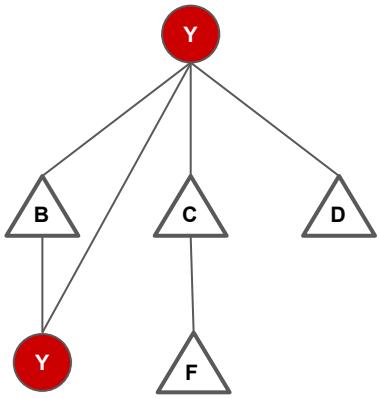

E)

**Expression**

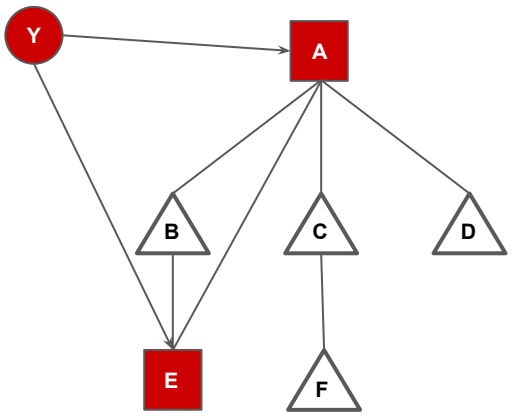
| A | B | C | D | E | F | |
|---|---|---|---|---|---|---|
| 82.21 | 5.59 | 5.59 | 5.48 | 0.56 | 0.56 | $X_{TSS}$ |
| 83.89 | 22.03 | 5.70 | 5.59 | 82.20 | 0.57 | $Z_{TSS}$ |
| N/A | 1st | 1st | 3rd | N/A | 4th | $X_{TSS}$ |
| N/A | 1st | 2nd | 3rd | N/A | 4th | $X_{TSS}$ |

***Figure 4.15. Toy networks showing the effect of whole genome vs TSS gene mapping procedures.*** A) Using TSS mapping gene X maps to node A. B) Using TSS mapping gene Y maps to nodes A and D. C) In this scenario, the resulting IAS for genes X and Y with the same starting values are ranked identically. D) Using TSS mapping gene Z mapps to nodes A and E. E) In this scenario when gene Z has the same starting value as genes X and Y the final IAS ranks node C differently.

A)




B)




C)

| | | | | | | |
|---|---|---|---|---|---|---|
| | | | Expression | | | |
| **A** | **B** | **C** | **D** | **E** | **F** | |
| 91.28 | 13.15 | 4.66 | 4.56 | 85.88 | 0.47 | Direct |
| 8.86 | 0.92 | 0.36 | 0.35 | 8.54 | 0.04 | Smoothed |
| N/A | 1st | 2nd | 3rd | N/A | 4th | Direct |
| N/A | 1st | 2nd | 3rd | N/A | 4th | Smoothed |

***Figure 4.16. Toy networks showing the effect of direct vs smoothed assignment of gene expression scores to genic nodes.*** Using smoothed gene expression prevents genes that map to multiple genes from artificially inflating their IAS values when propagating to other nodes in the network.

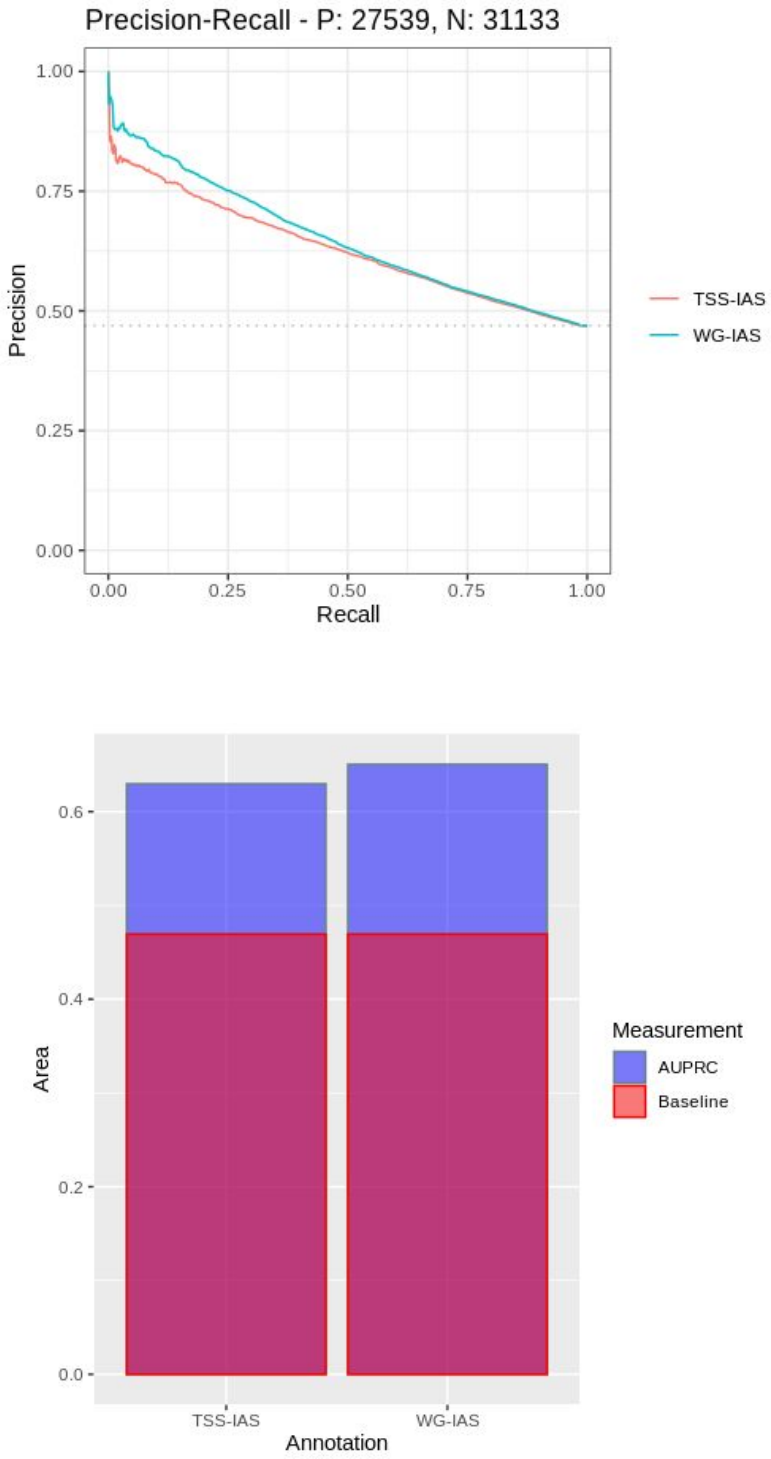



***Figure 4.17. Precision-recall curve for TSS IAS vs WG IAS.*** The area under the precision recall curve for TSS mapping is 0.629 compared to 0.650 when classifying nodes annotated with at least one enhancer feature from enhancer chromatin states, P300, CAGE-seq, Starr-seq and and three RNA polymerase II (RNAP II) complexes and RNAP s2p.

# 4.3 Discussion

A major problem in the identification of enhancers is their heterogeneity in both composition and location. In this chapter, we have outlined the development of a method that will be used to predict enhancers by leveraging the relationship between gene expression and chromatin interactions. To achieve this we first model 3C interaction data as networks. We then utilise a network propagation algorithm based on random walks to both integrate gene expression data within the network. In doing so, we are then able to impute an activity score (IAS) in intergenic nodes by propagating the gene expression values from genic nodes across the network. The IAS is a metric that summarises the importance of an intergenic node in the context of its proximity to all of the genic nodes within the network, the relative expression of those genes as well as the topology of the entire network.

The use of network propagation has been used previously in other fields. In cancer research for example, network propagation has been used to identify previously unknown proteins involved in the progression of cancers by propagating mutational frequencies across a protein-protein interaction network (Hofree et al. 2013). This method relies on the assumption that the interactions of the biological components have a functional effect. We utilise this same concept using chromatin-chromatin interaction networks and propagating gene expression values instead of mutational frequencies. Here, the underlying assumption is that the activity, defined by the gene expression, of one chromatin fragment is influenced by, or can influence the activity of all other chromatin fragments. This assumption is merited by what we know about chromatin organisation and gene regulation. For example, chromatin organisation is known to show preference to chromatin of similar composition; distinct A/B compartments contain majority euchromatin and heterochromatin (S. Wang et al. 2016). At the more local level, transcriptional factories have been hypothesised to localise and influence the transcriptional activity of co-expressed genes (Sutherland and Bickmore 2009). The hierarchical organisation of chromatin suggests that there exists a global influence on gene expression and network propagation provides a useful tool to model this.

Our results show that by computing IAS using the topology of the chromatin interaction networks (ChINs) and global gene expression we can independently identify nodes with active chromatin states. There are several limitations in both the model and the testing methodology that can be improved. The 3C data used to create the ChINs contain various biases. These include the resolution of the chromatin fragments and the enrichment for specific regions of chromatin. The resolution of the fragments can vary and can be determined by the capture type, for example the type of restriction enzymes used and the accessibility of said restriction sites. This can lead to the undersampling of interactions in more connected regions that are captured as larger fragments. This information is partly captured by the interaction frequency and in theory could be used to weight the edges of the network prior to propagation, although this was not considered at the time.

The flavour of 3C can also greatly affect the enrichment of specific regions of chromatin. In our networks we use a PCHi-C, or promoter capture Hi-C, dataset to create our networks. This has the advantage of capturing enhancers, due to their likely proximity to genes. However, by enriching for chromatin in the proximity of promoters only a selective sample of global chromatin topology is captured. Interestingly, this is possibly highlighted by the

marginal increase in performance of the model across the three cell-type consensus network. By including the nodes and edges of closely related cell-types we are also ‘filling in the gaps’ of some chromatin regions missing in the dataset either through the enrichment of specific chromatin regions or inefficiencies in the experimental protocol and/or the *in silico* post-processing of the data. In theory, using a more expansive 3C capture such as Hi-C (all against all) should yield better predictions. The analysis of networks derived from different 3C experiments and their advantages and limitations are provided in **chapter 3**.

Our network propagation algorithm also uses RNA-seq data as an input. In order to identify enhancer nodes within the network we used the expression levels of genes as our *a priori* information to then impute the activity scores at intergenic nodes. Because we use gene expression data the starting values at genic nodes are not uniform. Therefore, intergenic nodes proximal to genic nodes of high and low expression could receive a different IAS. However, provided that gene expression is controlled, at least partially, by the organisation, folding and looping of chromatin to localise enhancers to promoters, the relative differences in gene expression levels should reflect the relative differences in the network topology (Panigrahi and O’Malley 2021). However, this still poses a problem when assessing the performance of the model at a global level. The effects of differentially expressed regions of the network are examined in further detail in **chapter 5**.

The inputs used can include bias and noise into the model. However, the methods used to test the model at this stage were not ideal and also skewed the results. Correlation of IAS with the chromatin state coverage was not a good measure of performance and ROC curves are undermined by the imbalance in the positive and negative labels. Related to this problem is the process of reliably identifying enhancer nodes. Chromatin states remain one of the most popular methods for identifying enhancers due to their ability to profile chromatin with genome wide coverage and their relative availability for many cell-types (Ernst and Kellis 2017). As highlighted in the introduction and an important motivation for this thesis, chromatin states and many other enhancer associated features do not capture the full repertoire of enhancers for any genome. Without a reliable ground truth the true false negative rate of our model cannot be truly measured. To alleviate some uncertainty we can incorporate a more comprehensive and diverse set of enhancer associated features. In the following chapter we include CAGE-seq, eQTL, Starr-seq, RNA Pol II Chip-seq and P300 Chip-seq data data sets.

The results of this chapter provide a proof of concept and a first step in the prediction of enhancers by first identifying active regions of chromatin. In the following chapter we further investigate the use of network propagation as a tool to identify enhancers while also evaluating other network theory approaches.

# Chapter 5: 3D-SearchE: A method to leverage gene expression and chromatin architecture to classify intergenic enhancers

---

## 5.1 Introduction

Currently, enhancer associated features such as histone modifications, coactivators and transcription are used in lieu of any definitive and universal enhancer feature. It is therefore paramount to identify new enhancer features that can be used to further characterise them. In **chapter 3** we have demonstrated the potential of using network theory to identify the topological properties of intergenic enhancers. However, there is currently a dearth of computational tools designed to leverage biological data using network based approaches.

Here we present 3D-SearchE - a network analysis tool that will be made available to the scientific community to propagate gene expression data across 3C networks in order to identify enhancer regions of interest. 3D-SearchE can be found at https://github.com/manind95/3D-SearchE/tree/main/propagation_v13 and will be made publicly available in the near future. In this chapter we present the work that builds on the initial proof-of-concept presented in **chapter 4.** In **chapter 4** we outlined a method to integrate gene expression data with chromosome conformation data in order to classify enhancers. We presented an initial proof of concept that demonstrates the feasibility of using network propagation as a tool to classify nodes as active enhancer loci as defined by chromatin states.

## 5.2 Aims

In this chapter we expand on the work of **chapter 4** by measuring the performance of 3D-SearchE in classifying enhancer nodes using the PIC and mESC networks using the enhancer annotations from **chapter 3**. The aims of this chapter were to A) Further validate the use of network propagation as a tool to classify enhancer nodes using additional enhancer features. B) Understand the mechanisms that underpin the performance of the classifications and where classification performance may be lost. C) Develop a method to link putative enhancers with their cognate genes. D) Improve the classification performance of 3D-SearchE.

# 5.3 Results

## 5.3.1 Imputed activity scores generated by 3D-SearchE are significantly higher in enhancer nodes with multiple annotations

In **chapter 3** we showed that enhancers were uniquely connected within the ChINs. In **chapter 4** we propagated gene expression across the three immune cell ChINs and demonstrated that this can also be used to identify enhancer nodes. Using this propagation approach, in this chapter we outline the development of 3D-SearchE. 3D-SearchE works by calculating an **imputed activity score (IAS)** at intergenic nodes based on the transcriptional activity of gene nodes and the topology of the network. It then identifies likely gene-enhancer pairs using a novel distance based approach to define gene regulatory neighbourhoods. Based on the results of **chapter 4** we extended our analysis to the DNaseI and promoter-capture HiC derived network from mESCs. For mESCs there is a more comprehensive and diverse set of enhancer annotations available that include the histone modifications and FANTOM5 CAGE-seq data available for the immune cells in addition to P300 ChIP-seq, Starr-seq and RNA-pol ChIP-seq. For these analyses we used only the RNA Pol II s2p variant and excluded the s5p and s7p variants. This was done as 3D-SearchE aims to identify actively transcribed intergenic regions based on the relative levels of gene transcription. The s5p and s7p were then included for comparison in **5.3.4**. For the primary immune cells we included both FANTOM5 CAGE-seq data and eQTL data.

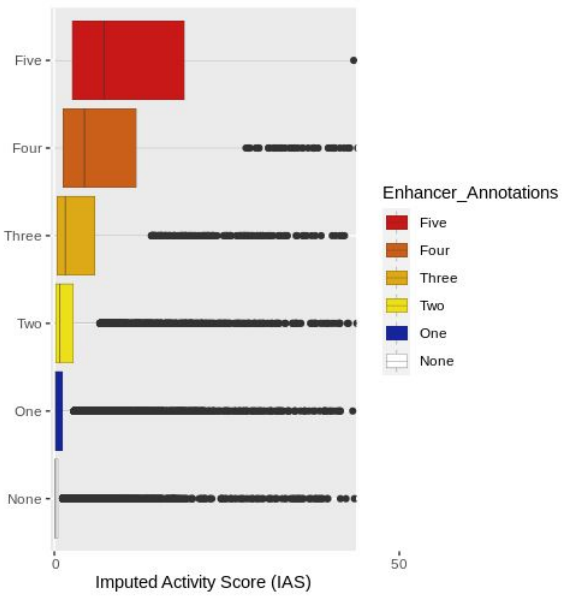
Following propagation, intergenic nodes were identified using an Ensembl gene database (**see chapter 3**). In the mESC DNaseI ChIN we found that intergenic nodes that contain at least one enhancer feature there was a significant enrichment of IAS in enhancer nodes compared to non-enhancer nodes (Wilcoxon rank sum test p.value = < 2.2e-16). We next stratified the nodes into 6 groups based on the different combinations of the 5 enhancer features. The median value of IAS in each group increased as the number of enhancer features increased. While the median score from nodes absent of any enhancer annotations to the group labelled with all five annotations increased by a factor of 46 (**Figure 5.1a**). These results show that 3D-SearchE not only enriches IAS more in enhancer nodes compared to non-enhancer nodes, but it also enriches IAS in nodes that are more likely to contain enhancers based on the presence of multiple enhancer associated features. This is despite the fact that the number of nodes decreases in each group reducing the likelihood of IAS enrichment by chance (**Figure 5.1b**).

To then understand the influence of using a different capture type and cell-types we then extended our analysis to promoter capture Hi-C (PCHi-C) ChINs. The mESC PCHiC ChIN uses the same 5 enhancer annotations used for the mESC DNaseI ChIN and can be used as a direct comparison. For the monocyte, neutrophil and T-Cell ChINs we used chromatin states, FANTOM5 CAGE-seq and eQTLs as enhancer annotations. The PCHi-C ChINs of mESCs, monocytes, neutrophils and T-cells generally result in a higher number of connected components - i.e. the network is more disconnected than the DNaseI capture (**See chapter 3**). Despite this, 3D-SearchE also performs well on these more disconnected ChINs as well as between cell-types with varying numbers and sizes of connected components. Comparing the performance of 3D-SearchE across these networks shows that despite the lower

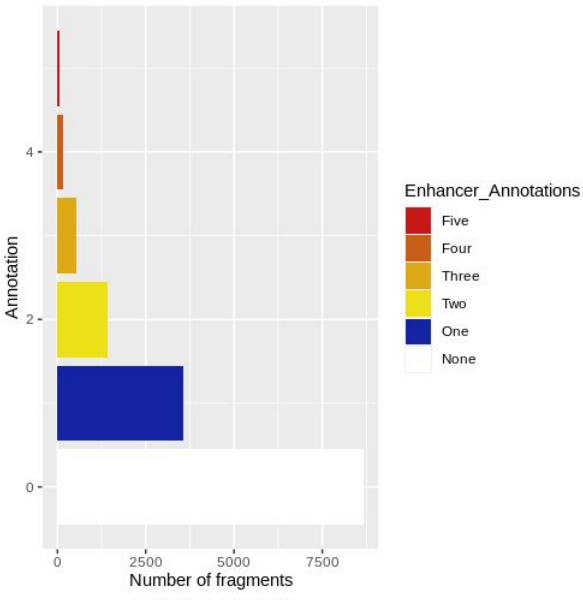
connectivity IAS is still enriched in enhancer nodes and shows a significant increase in enrichment between each group (Wilcoxon rank sum test p.value = < 2.2e-16). These results were consistent when using the PCHi-C ChINs where the IAS calculated from the mESC PCHi-C ChIN shows a 14-fold enrichment between enhancer nodes with 5 annotations compared to the non-enhancer nodes. For monocytes IAS showed an increased fold-change of 29X, neutrophils 17X and CD4+ T-Cells 26X between enhancer nodes with 3 enhancer annotations compared to non-enhancer annotations (**Figures 5.1c, 5.1d, 5.1e and 5.1f**).

Our results show that 3D-SearchE significantly imputes high gene expression values at enhancer nodes in DNaseI and DNaseI ChINs as well as between different immune cell types. These imputed expression values can be therefore used as a proxy of the enhancer activity at these intergenic regions that are often determined by assays such as Starr-seq (Arnold et al. 2013). This is achieved independently from any *a priori* knowledge enhancer associated features, using only gene expression and 3C contact data. These results show that 3D-SearchE can be used to leverage the intrinsic relationship between the chromatin topology and gene expression to identify which nodes harbour enhancers.

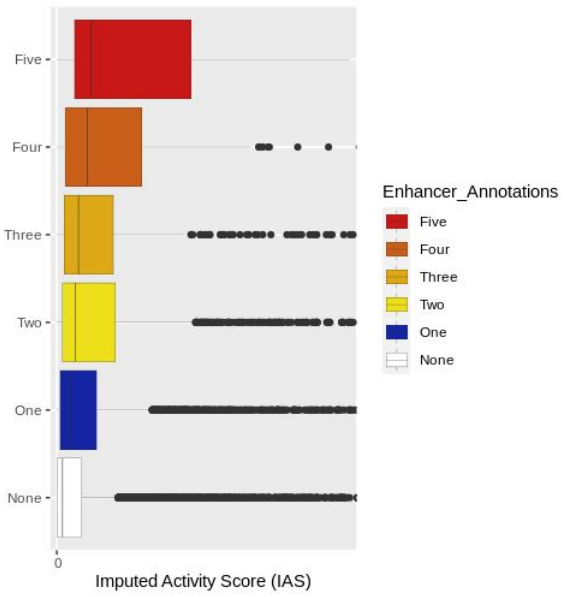
A


Five
Four
Three
Two
One
None
0
50
Imputed Activity Score (IAS)
Enhancer_Annotations
Five
Four
Three
Two
One
None

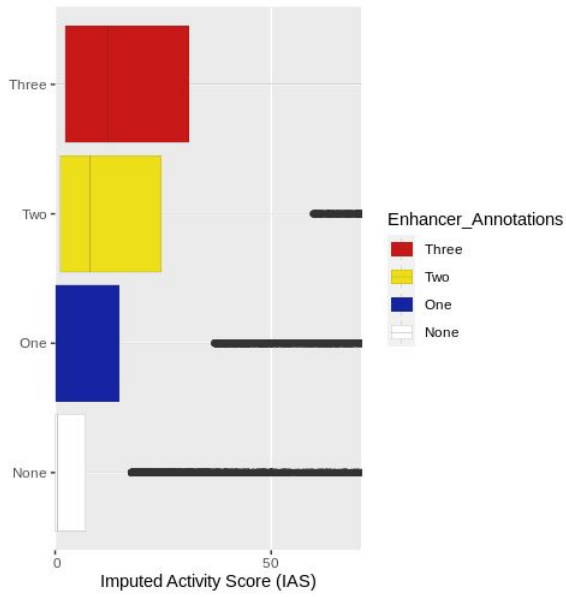
B


Annotation
4
2
0
0
2500
5000
7500
Number of fragments
Enhancer_Annotations
Five
Four
Three
Two
One
None

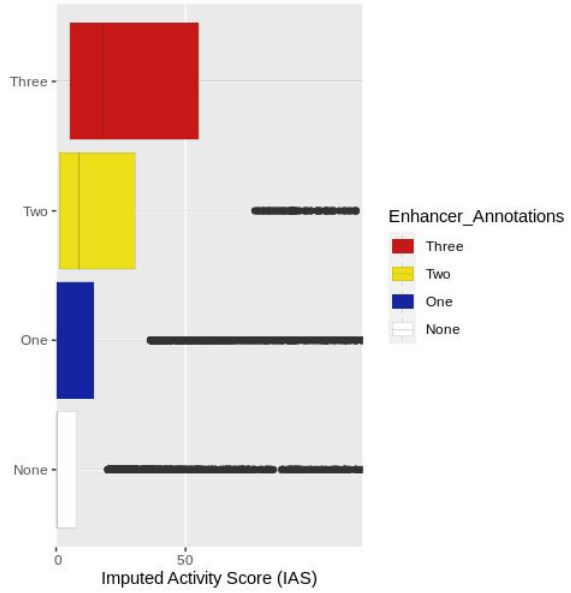
C


Five
Four
Three
Two
One
None
0
Imputed Activity Score (IAS)
Enhancer_Annotations
Five
Four
Three
Two
One
None

D


Three
Two
One
None
0
50
Imputed Activity Score (IAS)
Enhancer_Annotations
Three
Two
One
None

E




F

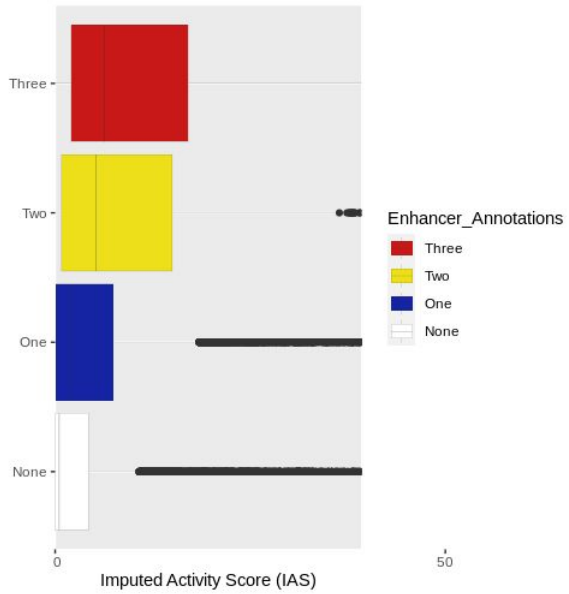



***Figure 5.1. Boxplot of IAS for groups of nodes labelled with 1 to 5 enhancer annotations.*** A) Intergenic nodes are stratified into enhancer and non-enhancer groups based on the number of enhancer associated features that were annotated previously. The imputed activity score is then plotted along the X-axis for each group of nodes. The enhancer group contain a significantly higher average imputed activity score compared to the non-enhancer group (Wilcoxon rank sum test p.value = <2.2e-16). B) The number of intergenic nodes that contain 0 to 7 enhancer annotations. C) IAS enrichment in each set for the mESC PCHI-C derived network. D) IAS enrichment in each set for the monocyte PCHI-C derived network. E) IAS enrichment in each set for the neutrophil PCHI-C derived network. F) IAS enrichment in each set for the CD4+ T-Cell PCHI-C derived network.

## 5.3.2 Precision and recall of enhancer classifications by 3D-SearchE

We have shown that IAS is significantly higher in nodes labeled with enhancer annotations. We next used these scores to plot precision-recall curves to quantify the performance of IAS in classifying nodes as enhancers and non-enhancers. This was achieved by calculating the area under the precision-recall curve (AUPRC) (**Figure 5.2**). The AUPRC quantifiably measures the ability of IAS to classify nodes as enhancers and non-enhancers by measuring the precision given by $Precision = tp / tp + fp$ and the recall given by $Recall = tp / tp + fn$ ability across incremental cut-off thresholds of IAS. At each cut-off the precision and recall is established. The precision and recall at each cutoff is then plotted that shows the trade-off between precision and recall. The area under the plotted curve can be used to summarise both, whereby a high area under the curve represents a high precision and a high recall. The use of precision-recall curves rather than a ROC curve, as used initially in **chapter 2**, was used to address the class imbalance in our labels . This is shown by the smaller proportion of enhancer nodes compared to non-enhancer nodes in the DNaseI ChINs (**see chapter 3**). In a ROC plot this ratio is assumed to be 50:50 giving a baseline of 0.5. For a precision-recall curve, this is calculated as the ratio of positive to negative labels, in this case enhancer vs non-enhancer nodes (Takaya Saito 2015).

Beginning again with the mESC DNaseI ChIN we assessed the overall performance of IAS in correctly classifying enhancer nodes. We show that IAS can be used to classify enhancer nodes containing **at least one annotation** with an AUPRC of 0.650 over a baseline of 0.469. The baseline is calculated as the ratio of enhancer nodes (n = 27,539) divided by the total number of intergenic nodes (n = 58,672). This is equal to a performance of 0.181 over baseline (**Figure 5.2a**).

For context, we then assessed the performance of each individual enhancer associated feature in classifying the other 4 features. For example, the AUPRC of P300 when classifying enhancer nodes labeled by enhancer chromatin states, Starr-seq, FANTOM5 CAGE-seq and the RNA Pol II. The AUPRC of IAS is also calculated for classifying nodes with this set of enhancer labels for direct comparison. The performance above baseline for each feature was as follows (AUPRC of IAS using the same labels in brackets): P300 = 0.11 (0.193), enhancer chromatin states = 0.101 (0.164), Starr-seq = 0.088 (0.190), CAGE-seq = 0.089 (0.187), RNAPs2p = 0.021 (0.173). In each case, IAS performs better at classifying enhancer nodes than any of the single enhancer features. This shows that both the connectivity of enhancer nodes and their proximity to genes within the network can be used to distinguish enhancer nodes from non-enhancer nodes.

We then extended the analysis to understand if our method can be used in the PCHi-C ChINs as well as between cell-types. For the mESC PCHi-C ChIN there is a 0.095 increase over the baseline (**Figure 5.2b**), a 0.086 decrease when compared to the mESC DNaseI ChIN. The differences observed between the two capture methods are likely due to the smaller representation of the mESC genome by PCHi-C compared to a DNaseI capture that covers 50% more of the total genome. This has two main effects. One is the number of potential enhancers that are captured; the proportion of intergenic nodes in the DNaseI ChIN

labelled as enhancers is 50% compared to 43% for the DNaseI ChIN. The second is that a lower coverage capture of the genome by the PCHi-C results in a less representative topology of the network. In the DNaseI ChIN we observe a lower average degree of 2.61 compared to 9.69 for the DNaseI ChIN (**Table 5.1**). We note that this does not drastically alter the density of the network which calculates the ratio between the number of actual connections vs the number of potential connections; the number of potential connections is given by the binomial coefficient of the number of nodes. This suggests that the PCHi-C isn't losing information about the topology of the network, rather it is sampling a smaller proportion of the network. We observe a similar performance for the other three PCHi-C ChINs with little variation in the performance between cell-types; monocytes 0.116, neutrophils 0.110 and CD4+ T-Cells 0.100 (**Figures 5.2c, 5.2d and 5.2e**).

As the mESC DNaseI ChIN appeared to be the best with respect to using IAS to classify enhancer nodes we then compared IAS to the, betweenness, eigenvector and degree centrality scores first investigated in **chapter 3**, now using the 5 enhancer labels to identify only active enhancer nodes. When comparing the features, IAS (AUPRC = 0.650) is comparable to the classification performance of the betweenness centrality (AUPRC = 0.676) and the degree centrality (AUPRC = 0.722) (**Figure 5.3**). This shows that IAS, the betweenness centrality and the degree centrality can be used to discriminate between enhancer and non-enhancer nodes. As discussed in **chapter 3** this cannot be used for feature selection and these features, including the eigenvector centrality which performs relatively poorly in this test (AUPRC = 0.516), may contribute differently to the performance of a machine learning classifier using all of the available features.

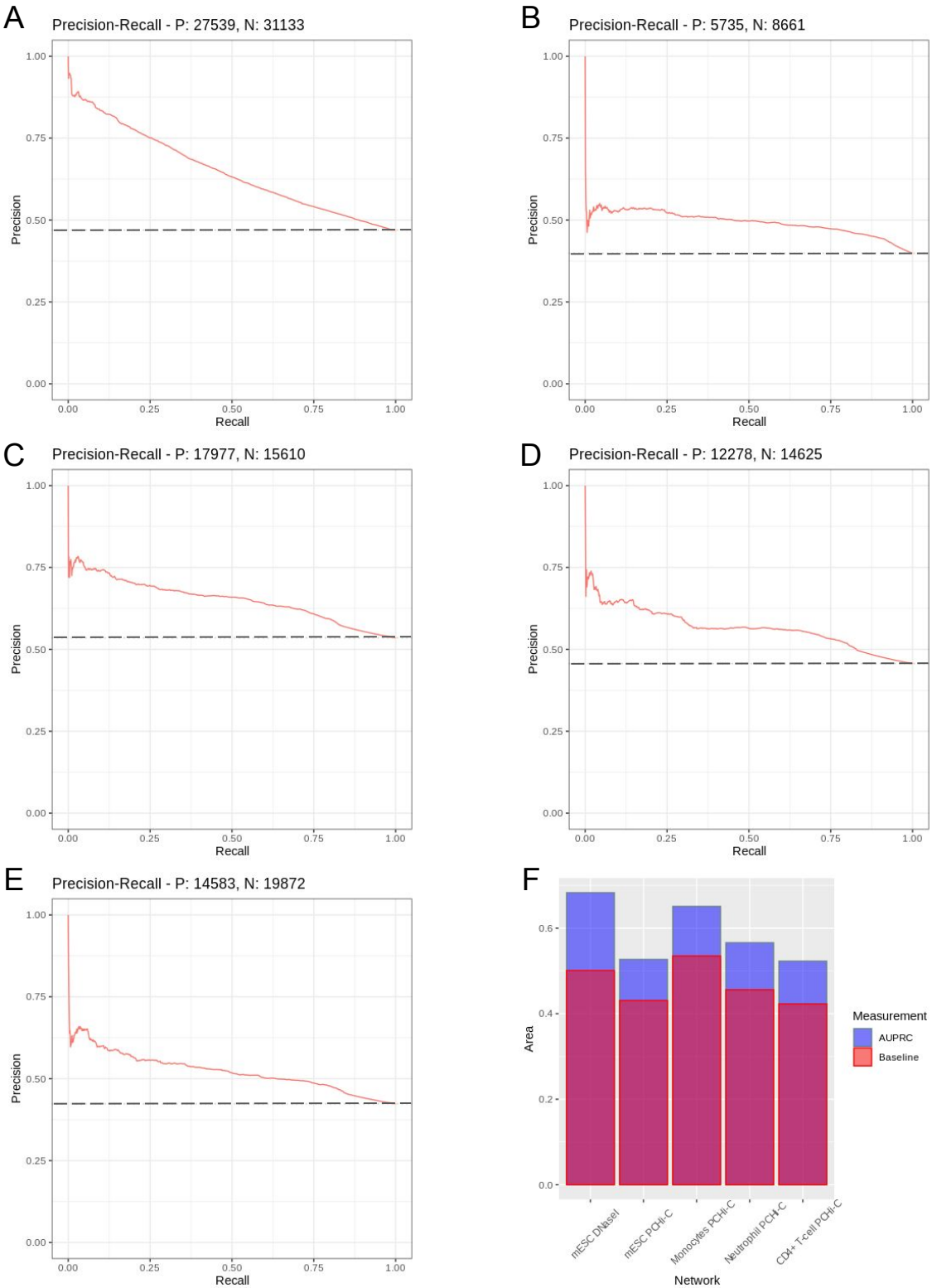



***Figure 5.2. A precision recall curve measuring the overall performance of IAS in the prediction of enhancer nodes containing a minimum of 1 enhancer feature.*** A) mESC DNaseI derived network AUPRC = 0.650 over a baseline of 0.469. B) mESC PCHi-C derived network AUPRC = 0.493 over a baseline of 0.398. C) Monocytesderived network AUPRC = 0.0.651 over a baseline of 0.535., D) Neutrophil derived network AUPRC = 0.566 over a baseline of 0.456. E) CD4+ T-Cell derived network AUPRC = 0.0.523 over a baseline of 0.423. F) An overlay barchart showing the AUPRC (blue) relative to the baseline score (red) for each of the the precision recall curves.

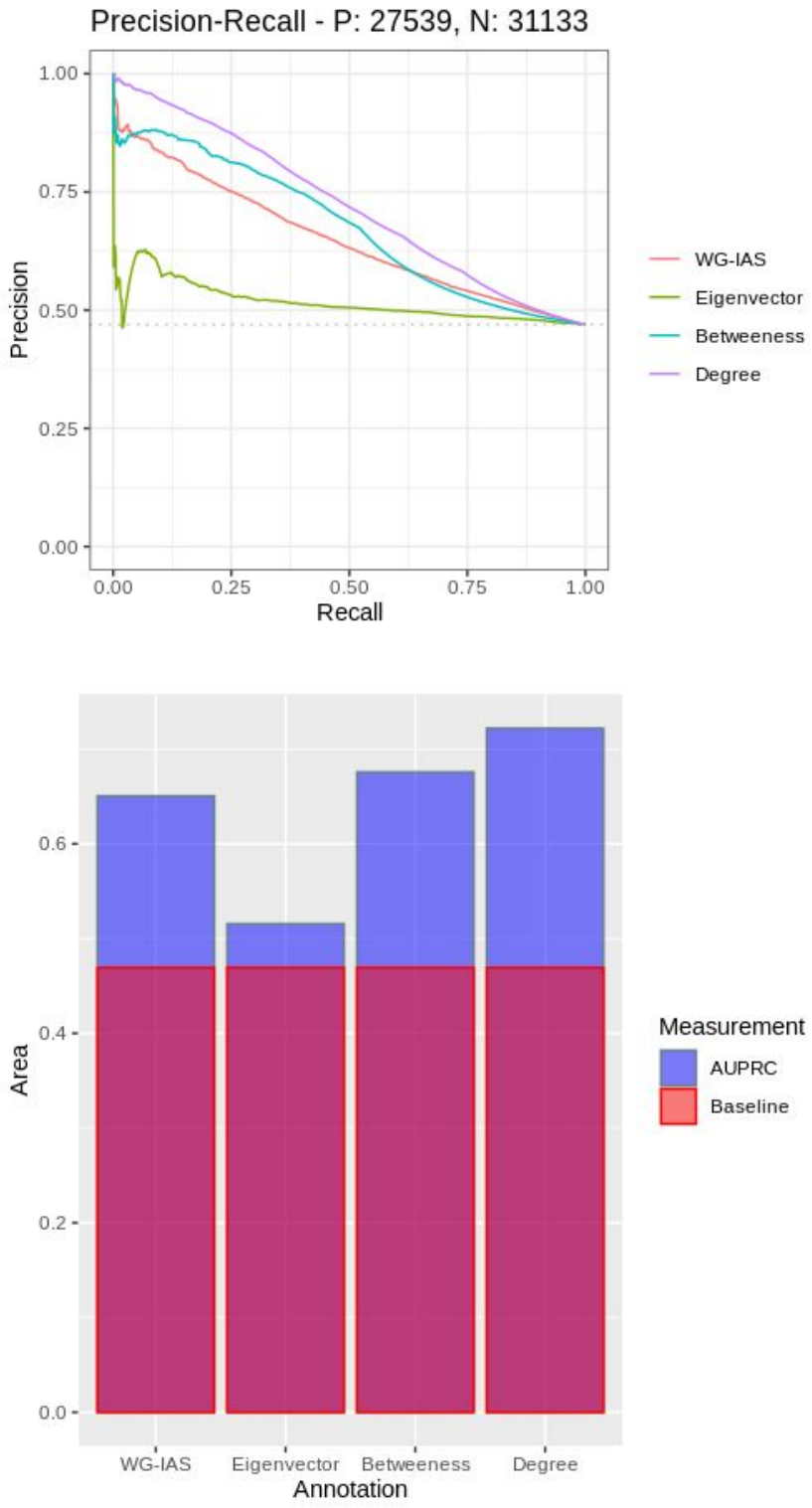



**Figure 5.3. A precision recall curve measuring the performance of IAS and three centrality scores** in the classification of enhancer nodes containing a minimum of 1 enhancer feature**.** IAS (AUPRC = 0.650), betweenness centrality (AUPRC = 0.676), degree centrality (AUPRC = 0.722), eigenvector centrality (AUPRC = 0.516).

## 5.3.3 Classification performance of IAS varies depending on the number of labels

In the previous analysis we quantified the ability of IAS to classify any node that contains at least one enhancer associated feature. For a node with multiple enhancer associated features from independent experiments that include both direct and indirect evidence for enhancers it could be argued that the probability of it containing an enhancer is higher. Since IAS is meant to reflect the probability of finding an enhancer in any given node we next looked at how IAS performs in classifying nodes with multiple enhancer associated features.

As shown in **Figure 5.1** we have already demonstrated that IAS is generally higher in nodes with an increasing number of distinct annotations. Using any combination of the five enhancer associated features in the mESC networks we again stratified the nodes into five groups based on the number of distinct enhancer associated features. We then looked at how well IAS is able to classify nodes as enhancers and non-enhancers in each of these groups by plotting the precision-recall curves.

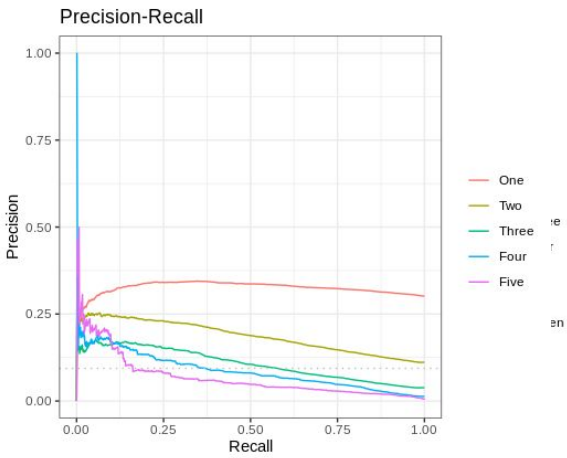
For the mESC DNaseI ChIN the model performs poorly in identifying nodes with just a single annotation with an AUPRC of 0.325 which is 0.023 above the baseline of 0.302 meaning that IAS on its own is not necessarily a good feature to be used to classify nodes with a single annotation in the mESC DNaseI ChINs (**Figure 5.4a**). However, in nodes with multiple enhancer features the model performs better with a performance over baseline of 0.075, 0.069, 0.075 and 0.062 as the number of features increases. This suggests that IAS is better at classifying nodes with multiple enhancer features than those with one and is consistent across all 5 ChINs (**Figure 5.4**).

One possible explanation for this trend is that IAS is able to identify some features better than others. We therefore looked at the composition of features that make up each group (**Figure 5.5 and 5.6**). For the mESC DNaseI ChIN, in nodes containing one feature, 46% are defined by enhancer chromatin states and the histone marks that make up each state are an indirect assessment of enhancer activity and can be found in non-enhancer regions of the genome. As more annotations are included and the AUPRC rises the percentage of enhancer chromatin states is reduced from 44% to 14.3% (**Figure 5.5**). In particular the percentage of the RNAP variants increase between the groups. Additionally, we know that RNA Pol II is able to mediate chromatin interactions and we observe the same performance patterns across the groups for the degree centrality, which is the number of interactions a node has. While the reasons are likely more complex than this, these may be contributing factors. We therefore investigated the performance of IAS in identifying each feature in further detail in **4.3.4**.

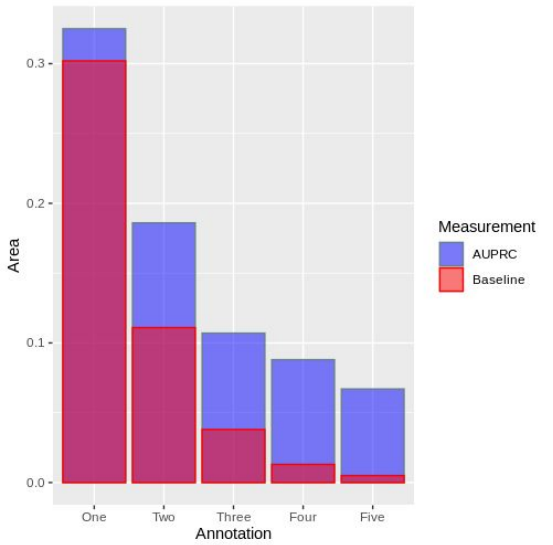
In the mESC PCHi-C we observe an even performance in classifying each of groups with the exception of those nodes that contain two enhancer associated features in which we observed a dramatic increase in performance with an AUPRC of 0.0.029 above the baseline for nodes with a single enhancer annotation; two annotations AUPRC = 0.040, three annotations AUPRC = 0.015, four annotations AUPRC = 0.011 and five annotations AUPRC = 0.006. When we extended this analysis to the primary immune cell networks we observed more even performance between each of the annotation groups suggesting that IAS is not

necessarily better at classifying between nodes containing different enhancer associated features PRC for each network even though we see a significant increase of IAS between the groups in **Figure 5.1**.

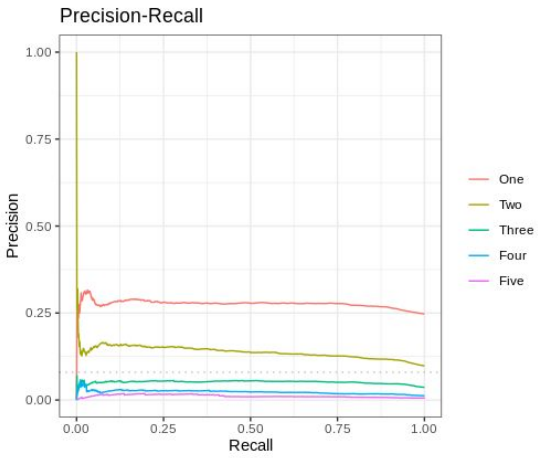
A


Precision-Recall
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
One
Two
Three
Four
Five

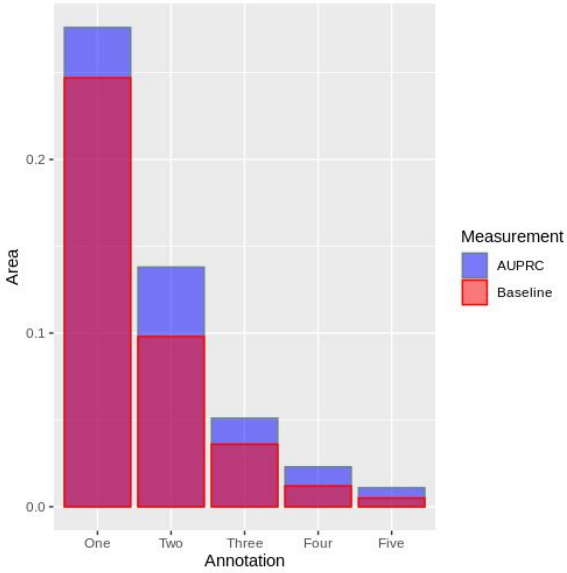
B


Area
Annotation
0.3
0.2
0.1
0.0
One
Two
Three
Four
Five
Measurement
AUPRC
Baseline

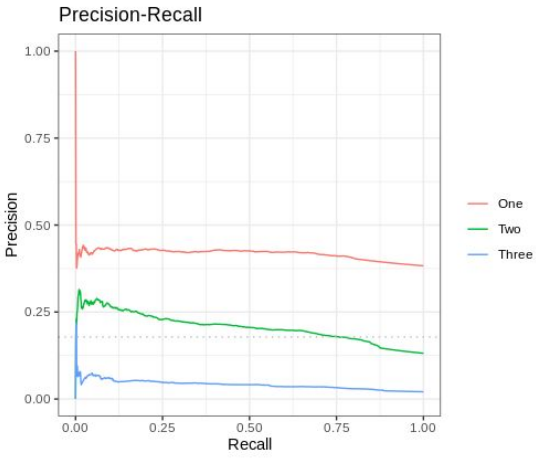
C


Precision-Recall
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
One
Two
Three
Four
Five

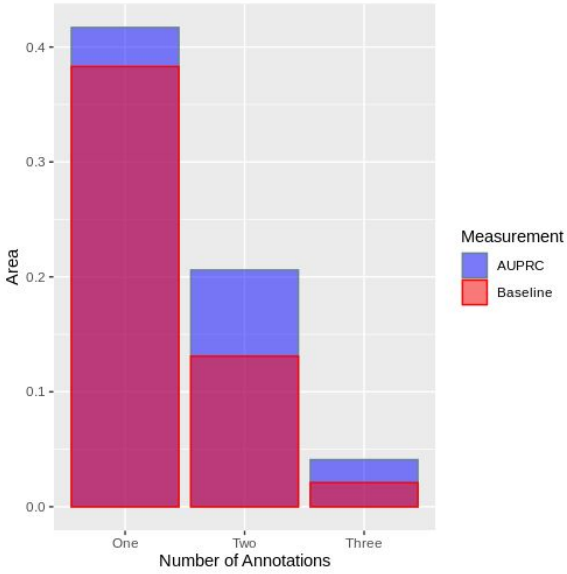
D


Area
Annotation
0.2
0.1
0.0
One
Two
Three
Four
Five
Measurement
AUPRC
Baseline

E


Precision-Recall
Precision
Recall
1.00
0.75
0.50
0.25
0.00
0.00
0.25
0.50
0.75
1.00
One
Two
Three

F


Area
Number of Annotations
0.4
0.3
0.2
0.1
0.0
One
Two
Three
Measurement
AUPRC
Baseline

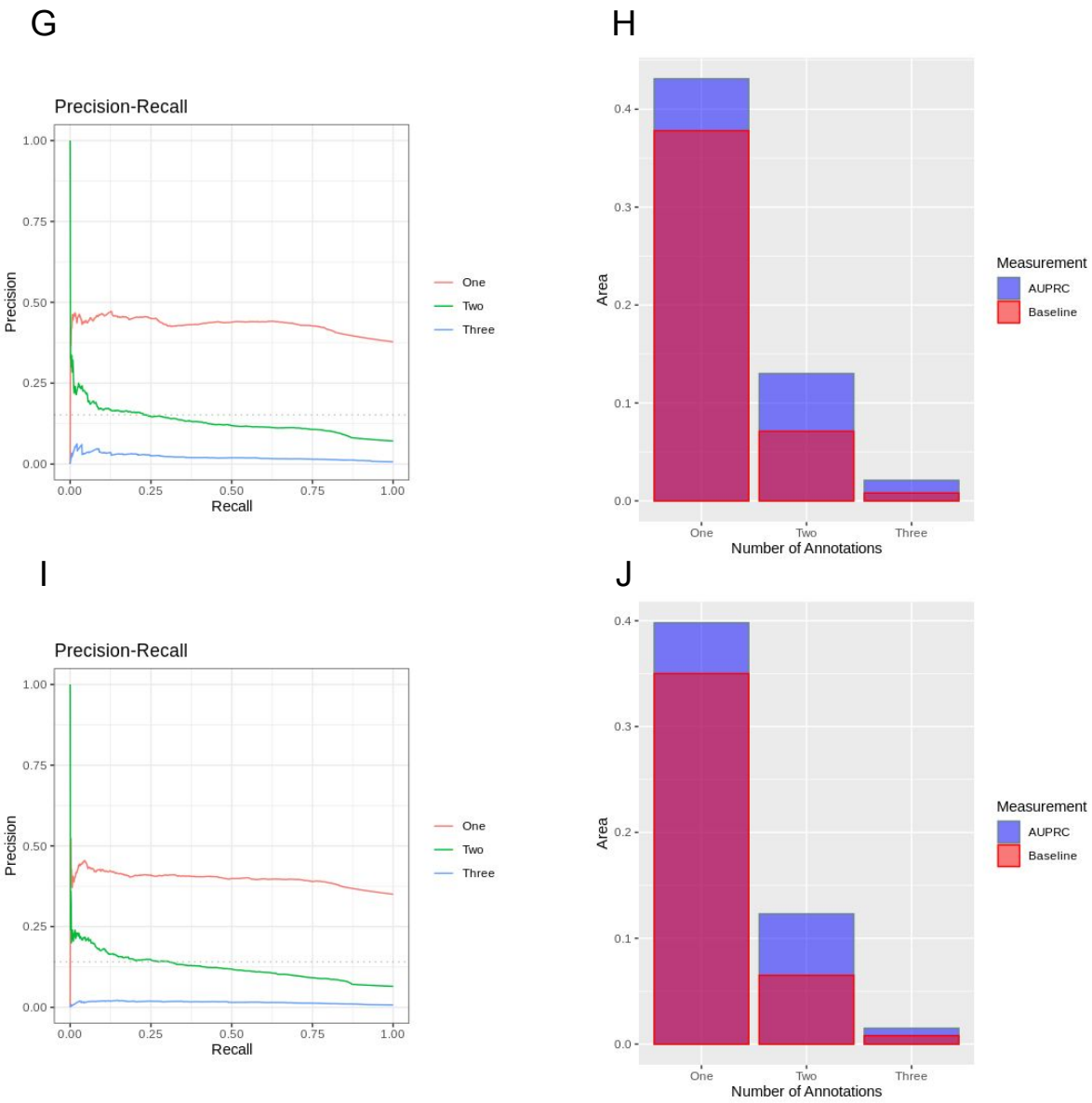



***Figure 5.4. Precision-recall for classifying combinations of multiple enhancer features***
A) A precision recall curve measuring the overall performance of IAS in the prediction of enhancer nodes containing at minimum 1 to 5 enhancer annotation for the mESC DNaseI ChIN (AUPRC over baseline = 0.023, 0.075, 0.069, 0.075 and 0.062). B) An overlay barchart showing the AUPRC (blue) relative to the baseline score (red) for the precision recall curve in part A. C and D) For the mESC PCHi-C ChIN (AUPRC over baseline = 0.0.029, 0.040, 0.015, 0.011, 0.006. E and F) For the monocyte ChIN (AUPRC over baseline = 0.034, 0.075, 0.020). G and H) For the neutrophil ChIN (AUPRC over baseline = 0.053, 0.059, 0.013). I and J) For the neutrophil ChIN (AUPRC over baseline = 0.048, 0.058, 0.007)


***Figure 5.5. Composition of annotation groups one to seven for the mESC DNaseI ChIN.*** Each pie chart represents a set of nodes with one to five enhancer annotations in the mESC DNaseI network and shows the breakdown of each feature within each group


***Figure 5.6. Composition of annotation groups one to seven for the mESC PCHi-C ChIN.*** Each pie chart represents a set of nodes with one to five enhancer annotations in the mESC promoter capture Hi-C network and shows the breakdown of each feature within each group

## 5.3.4 The classification performance of 3D-Search is superior for some enhancer annotations

The performance of IAS in classifying enhancer nodes appears to vary depending on the composition of annotations. This would suggest that IAS is better at classifying particular enhancer associated features over others. Here, we measure the AUPRC for IAS when classifying nodes containing each of the five enhancer associated features in the mESC DNaseI and DNaseI ChINs. In this analysis we included the RNA Pol II variants s5p and s7p. These two variants identify RNA Pol II at the repressed (RNAPs5p) and initiating (RNAPs7p) stages of transcription. They were included to add additional information as to whether IAS also identifies intergenic nodes that are not being actively transcribed.

3D-SearchE performs best when classifying nodes that contain the s5p modification of RNA Pol II (AUPRC over baseline = 0.226 (**Figure 5.7**)); this is closely followed by the s7p modification (AUPRC over baseline = 0.193). Both the s5p and s7p modifications are known to be enriched at the start of genes where transcription is initiated. This corroborates findings that s5p and s7p tend to have a high betweenness, whereas s2p nodes tend to be more peripheral (Pancaldi et al. 2016). Therefore, these nodes will tend to be better connected within the networks and closer to active genes resulting in higher IAS scores relative to other nodes. This idea is partly corroborated by the lower performance of the s2p variant of RNA Pol II which is enriched as the polymerase molecule transits across the gene body (AUPRC over baseline = 0.098). 3D-SearchE also performed well in classifying nodes with enhancer chromatin states (AUPRC over baseline = 0.190) and nodes with CAGE-seq annotations (AUPRC over baseline = 0.115). While the performance was lower when classifying nodes with P300 (AUPRC over baseline = 0.095), and Starr-seq (AUPRC over baseline = 0.089) annotations.

The performance of 3D-SearchE in identifying each of the features using the PCHi-C ChIN were comparatively worse (**Figure 5.8**). The best AUPRC was achieved when classifying the enhancer chromatin states (AUPRC over baseline = 0.080). The classification of the s5p (AUPRC over baseline = 0.071) and s2p (AUPRC over baseline = 0.047) RNA Pol II modifications were also lower. Much of the loss in performance can be related to the connectivity differences between the two networks as shown in **chapter 3**. The expression spreads more evenly across the DNaseI ChIN than the DNaseI ChIN owing to the sparser connectivity observed in the DNaseI ChIN compared to the DNaseI ChIN. With more nodes sharing similar IAS values it becomes more difficult to discriminate between nodes and thus, reducing the performance of the model.

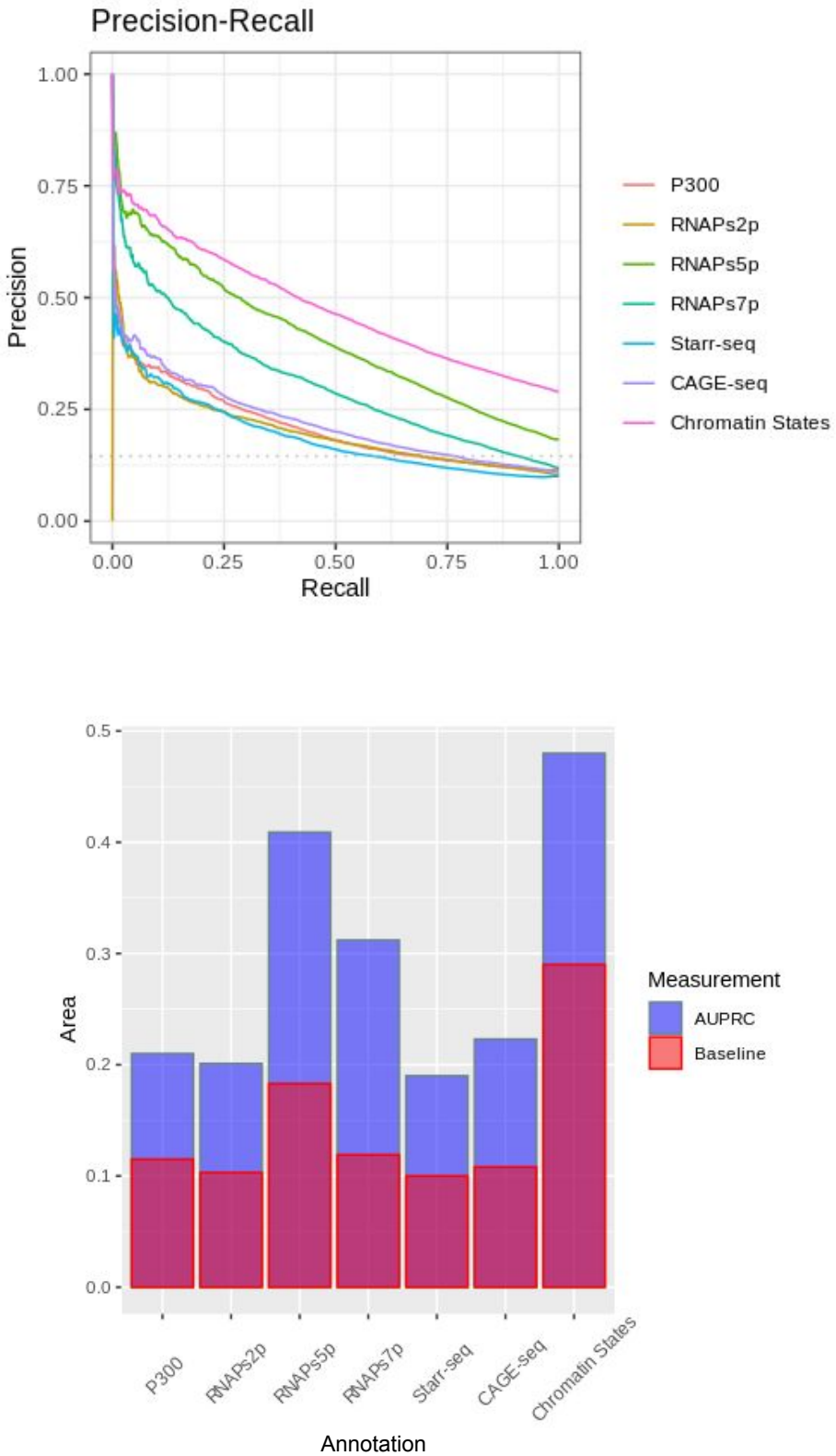



***Figure 5.7. Precision recall curve for IAS classifying each individual annotation in the mESC DNaseI ChIN.*** The imputed activity scores produced by 3D-SearchE were benchmarked using a precision-recall curve in their ability to correctly classify nodes containing each of the seven enhancer annotations in the mESC DNaseI ChIN. P300 (AUPRC over baseline = 0.095), RNAPs2p (AUPRC over baseline = 0.098), RNAPs5p (AUPRC over baseline = 0.226), RNAPs7p (AUPRC over baseline = 0.193), Starr-seq (AUPRC over baseline = 0.089), CAGE-seq (AUPRC over baseline = 0.115), Chromatin states (AUPRC over baseline = 0.190).

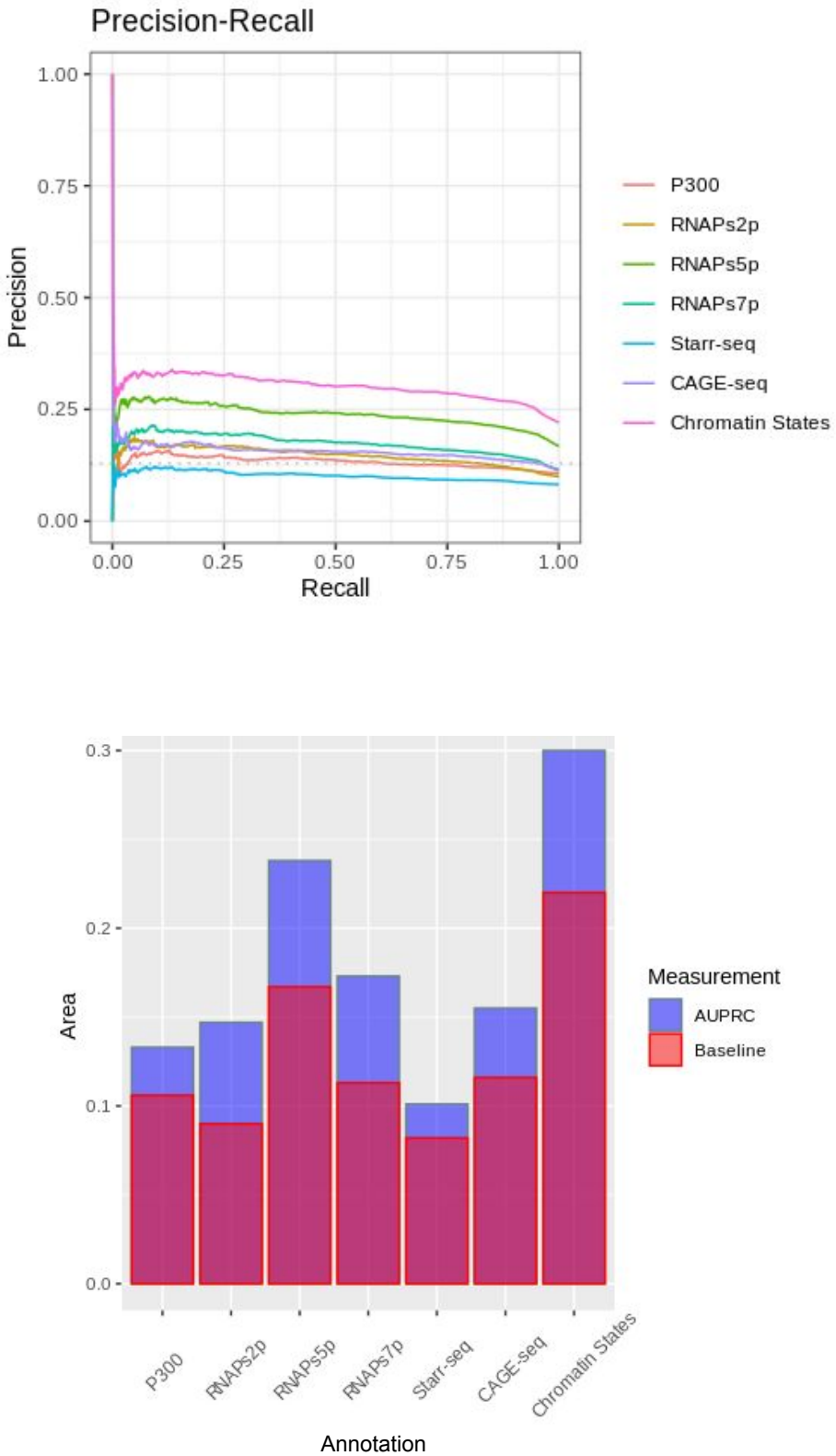



***Figure 5.8. Precision recall curve for IAS classifying each individual annotation in the mESC PCHi-C ChIN.*** The imputed activity scores produced by 3D-SearchE were benchmarked using a precision-recall curve in their ability to correctly classify nodes containing each of the seven enhancer annotations in the mESC PCHi-C ChIN. P300 (AUPRC over baseline = 0.095), RNAPs2p (AUPRC over baseline = 0.047), RNAPs5p (AUPRC over baseline = 0.071), RNAPs7p (AUPRC over baseline = 0.064), Starr-seq (AUPRC over baseline = 0.016), CAGE-seq (AUPRC over baseline = 0.042), Chromatin states (AUPRC over baseline = 0.080).

### 5.3.5 Chromatin topology and gene expression are crucial features to classify enhancers

3D-SearchE is unique to other topological measures in that it takes into account the proximity of intergenic nodes with all other genic nodes in the network. It is able to successfully classify enhancer nodes because enhancer nodes both maintain distinct connectivity patterns with genic nodes and are typically more proximal to genic nodes when compared to non-enhancer nodes. Indeed, the distinct connectivity patterns of enhancer nodes are demonstrated in **chapter 3**. For example, in the mESC DNaseI ChIN, gene nodes maintained 11.42 connections on average, while intergenic nodes maintained an average of 6.63. Of these intergenic nodes, enhancer nodes maintained 13.54 connections compared to 2.98 of non-enhancer nodes. However, the specificity of the connections in the network is essential to transmit information from enhancers to promoters which is accounted for by the propagation of gene expression. To demonstrate that this is an important feature in identifying enhancers we implemented a degree preserving randomised rewiring algorithm to shuffle the edges in the network while maintaining the degree of each node(**Figure 5.9a**) . Results showed that following the randomisation the ability of the model to classify enhancer nodes was lost with an AUPRC = 0.501 over a baseline of 0.501 (**Figure 5.9b**).

Our model also leverages the global expression of genes to classify enhancers. The classification of enhancers across the network is driven by simultaneously propagating the expression profiles of all the genes that map to the network. To understand the effects of multi-gene propagation on the performance of the model we implemented a single-gene propagation methodology. Using single-gene propagation also comes with the additional benefit of reducing the computational cost of 3D-SearchE. In this scenario, the expression profile of a single gene of interest was propagated within the network (**Figure 5.10a**). This component of our method was initially devised as it has the advantage of requiring less computational power than the alternative, which is to propagate the RNA-seq values of all the genes simultaneously across the network.  However, results showed that the propagation of a single gene reduced the performance of the model to that of a random classifier AUPRC = 0.19 (**Figure 5.10b**). This would suggest that the global expression of genes and the topology of the network are interlinked and likely influence one another. Together, these results indicate that the global properties of the network and gene expression are important features of enhancers and can be used to classify them.

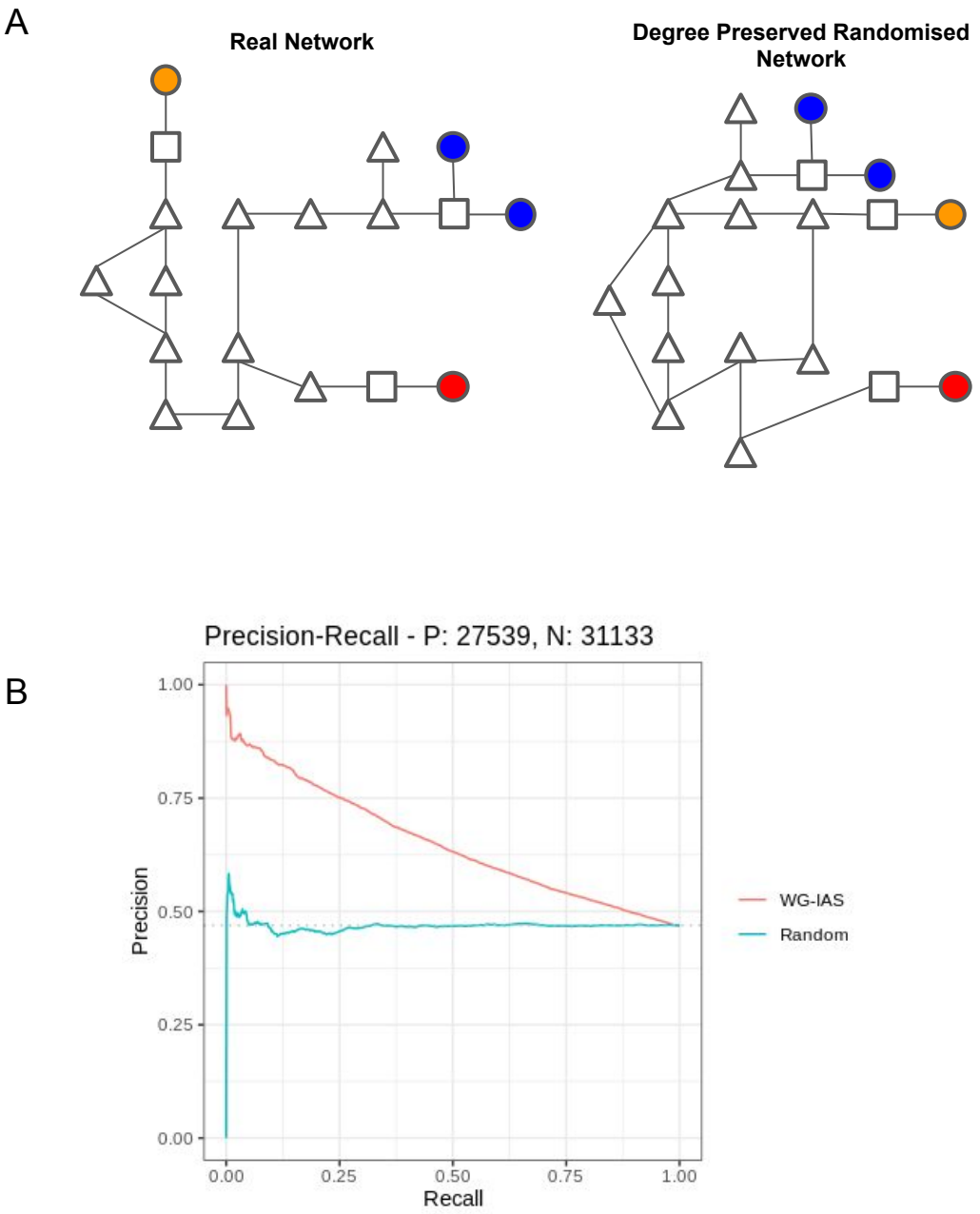



***Figure 5.9. WG IAS vs a randomised model*** A) A schematic of how the degree preserving randomisation rewires the network while preserving the degree of each node. Precision-recall curve for the randomised ChIN vs the non-randomised mESC DNaseI derived ChIN. For the non-randomised network the area under the precision recall curve is 0.65. The randomised ChIN AUPRC is 0.498. The baseline is 0.493.

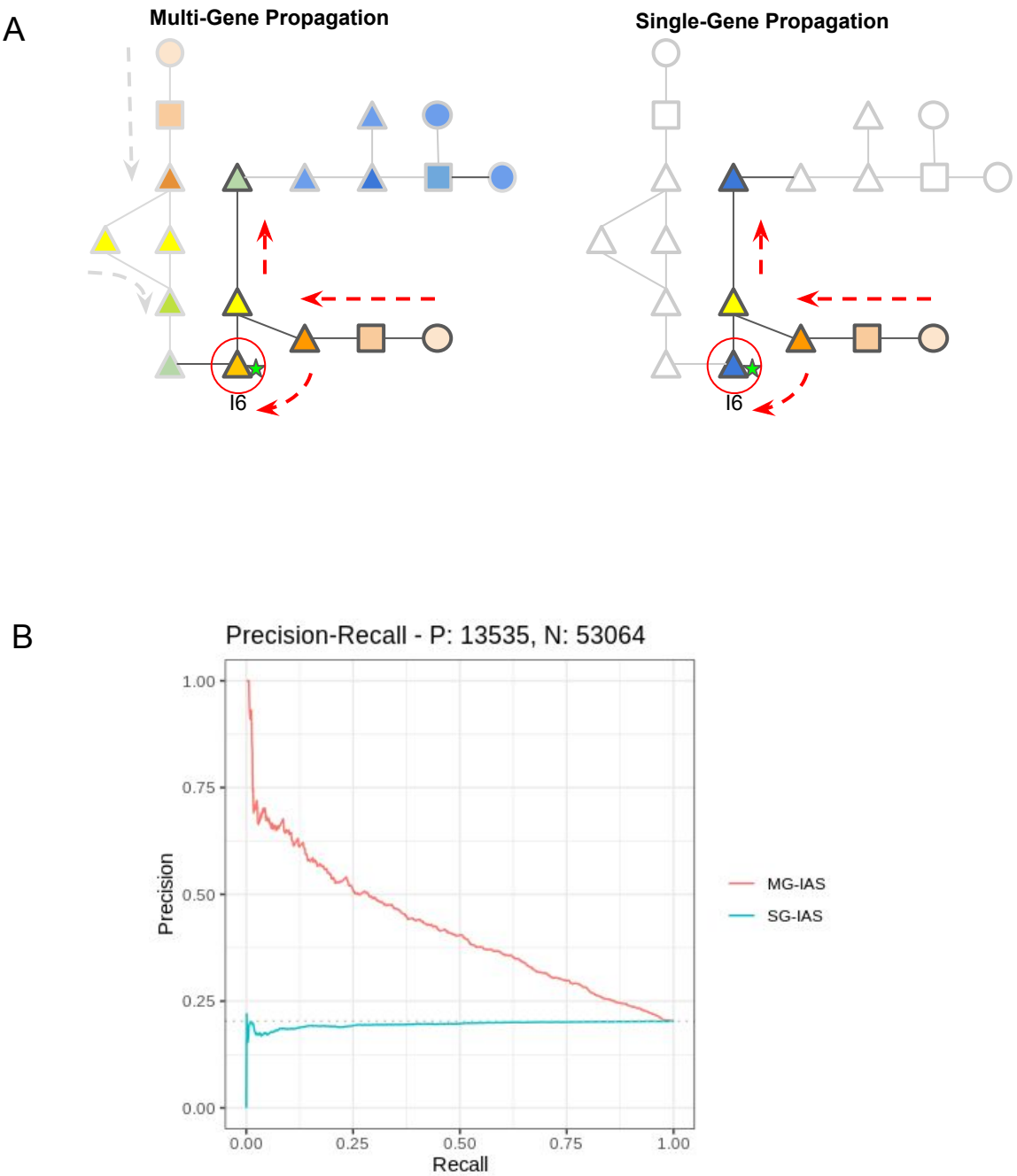



***Figure 5.10. A schematic of how multi-gene and single-gene propagation differ in the relative imputed activity scores.*** Multi-gene propagation highlights I6, an enhancer labelled node, with a higher IAS than single-gene propagation. Precision recall curves for the multi-gene propagation vs the single gene propagation. The multi-gene propagation on the same set of nodes performs with an AUPRC of while the single gene propagation AUPRC is 0.19 over a baseline of 0.20.

## 5.3.6 Gene neighbourhoods can be used to link putative enhancers and genes

After identifying enhancers the second challenge is then identifying the gene or genes they regulate. We have outlined a top-down approach that leverages the global expression of genes and the total structure of the network to classify enhancers. The underlying notion being that the regulation of genes is influenced at a global level. More specifically, enhancers are able to regulate the expression of genes as determined by the topology of the network. To identify which enhancers directly regulate which gene(s) we developed a bottom-up approach from the perspective of the genes. Using this approach we were also able to normalise the differences in relative expression across the network. For example, enhancers proximal to a low expression gene will receive a lower IAS and will therefore be ranked lower. By isolating these enhancers in the context of their proximal genes we are also able to reduce the effects of expression bias in the network and improve the classification of enhancers. At a biological level enhancers of all varieties will be in close proximity to the genes they regulate. Therefore, within the network the enhancer that regulates a particular gene will be within a reasonable distance. We used a network distance measure to reduce the number of enhancer candidates for a gene.

Following multi-gene propagation we tested the classification performance for a set of genes. We identified a set of the most highly expressed genes (115) in the mESC DNase I network. We then segmented the network around each gene to generate a local neighbourhood with a distance of 3 (**Figure 5.11**). On average each neighbourhood contained ~141 intergenic nodes. The set of intergenic nodes for the 115 genes contained 8,864 enhancer nodes, representing 55% of the total. IAS was then used to classify the enhancer nodes. The average performance across all gene neighbourhoods resulted in an AUPRC of 0.69 over a baseline of 0.40 (**Figure 5.12a.**).

To ascertain the performance in lower expression regions of the network we performed the same analysis on two more subsets of genes with medium and low expression. The average performance across the 100 gene neighbourhoods in the medium expression set resulted in an AUPRC of 0.63 over a baseline of 0.36 (**Figure 5.12b**) while the low expression set had an AUPRC of 0.79 over a baseline of 0.49 (**Figure 5.12c**). Despite the lower levels of expression in these regions there is no performance penalty and the number of enhancer nodes in each set also do not appear to have a significant effect on performance (**Figure 5.12d**).

These results also show that the performance of IAS is better than the AUPRC determined when assessing all intergenic nodes at a global level. For example, an intergenic node A in a low expression region of the network will receive a lower IAS following propagation than node A in a high expression region. In the context of all intergenic nodes in the network when calculating the AUPRC node B is considered more likely to be an enhancer node than node A. When ranking the nodes at a local level we are able to remove this bias and determine the nodes with the highest enhancer potential in the context of local gene expression levels.

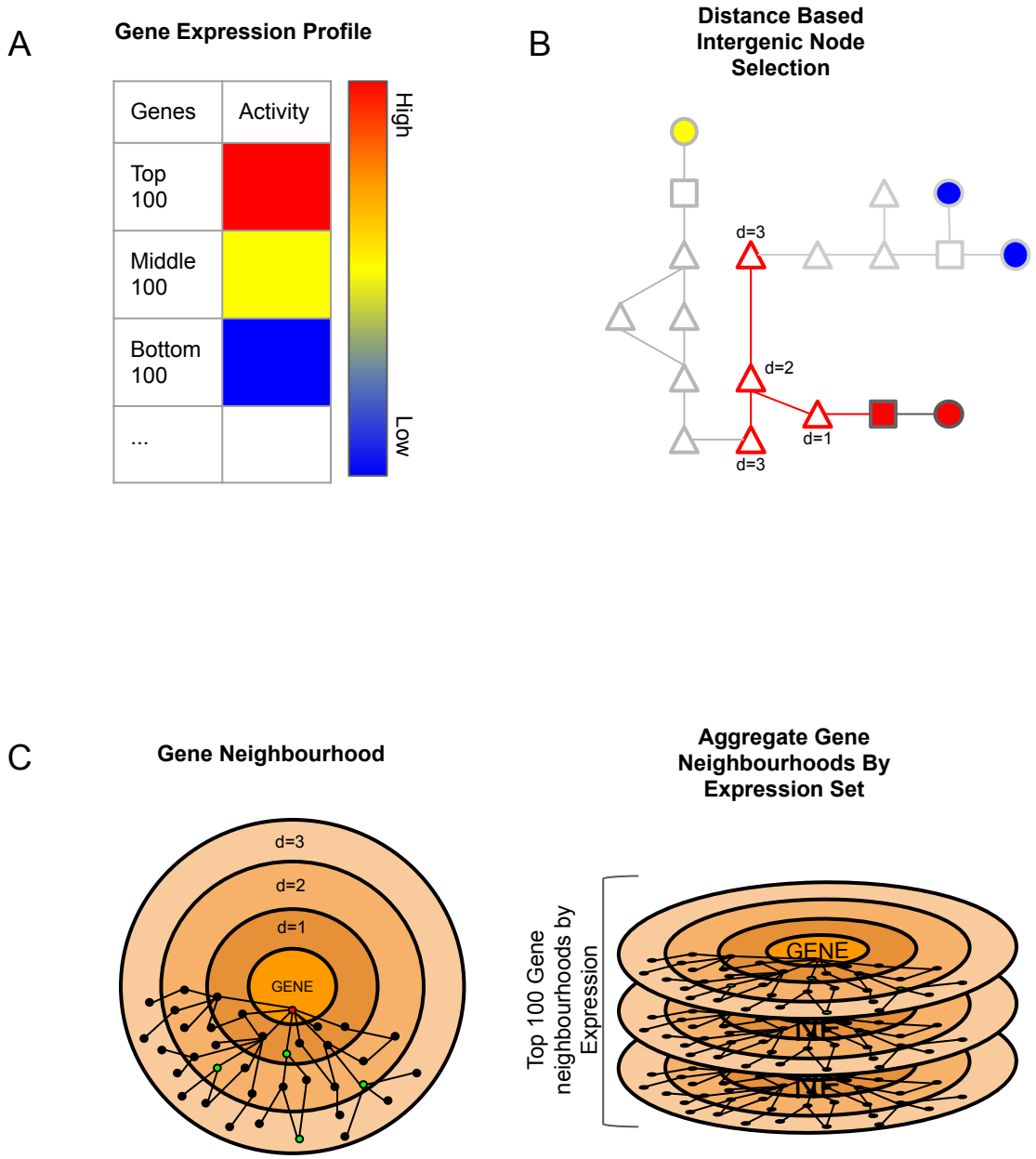



***Figure 5.11. Gene neighbourhoods*** A) i) Genes were grouped into expression sets based on relative expression levels and the top, middle and bottom 100 genes were selected. B) For each gene in each expression set, a group of intergenic nodes were selected based on a distance of 3 from the gene node. C) Each gene has a defined neighbourhood consisting of all intergenic nodes within a distance of three. Each neighbourhood is then aggregated into three groups defined by the expression sets.

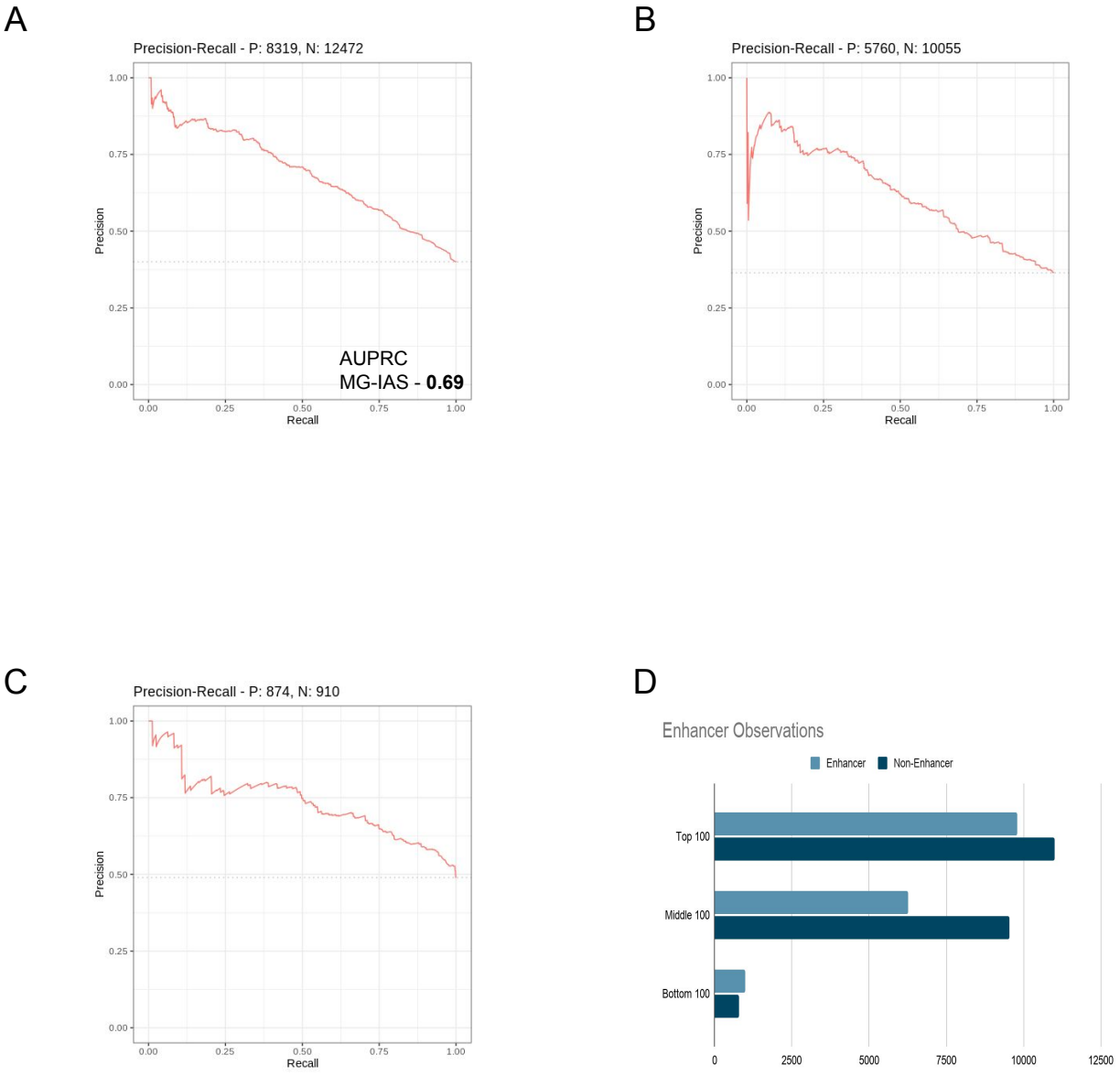



***Figure 5.12. AUPRC for high, medium and low expressed regions.*** The AUPRC for each of the three expression sets to identify if the predictive performance is consistent across high, medium and low expressed regions of the ChIN. A) Top 100 AUPRC = 0.69 over a baseline of 0.40, B) middle 100 AUPRC = 0.62 over a baseline of 0.36, C) bottom AUPRC = 0.73 over a baseline of 0.49. D) The total number of nodes in each expression set and the proportion of enhancer vs non-enhancer nodes in each group.

# 5.4 Discussion

Our results show that 3D-SearchE can be used to classify active enhancer nodes. The imputed activity score (IAS) calculated by 3D-SearchE was shown to be higher in nodes annotated with active enhancer features. Most strikingly, IAS showed a significant increase between enhancer groups with increasing numbers of independent enhancer annotations (**Figure 5.1**). These results show that 3D-SearchE ranks high confidence enhancer nodes higher on average than lower confidence and non-enhancer nodes. However, the wide range of IAS values in each group suggest a high level of heterogeneity between the nodes in each group in terms of their connectivity and proximity to genes. This results in many non-enhancer nodes receiving a high IAS and many enhancer nodes receiving a low IAS, both of which reduce the performance of the model when measured using the precision-recall metrics. As highlighted, there are several contributing factors towards the lower than expected performance, some of which we address and others that can be addressed in future iterations of 3D-SearchE.

Initially, the nodes were ranked globally using IAS and the precision and recall metrics calculated by comparing the global relative IAS. However, the expression of genes is non-uniform meaning that there are high and low regions of expression across the network. This results in enhancer nodes in low expression regions of the network being ranked lower than non-enhancer nodes in high expression regions of the network. By measuring the precision and recall of 3D-SearchE globally the relative differences in gene expression in different regions of the network is not accounted for. This is addressed with the use of a network segmentation approach that isolates intergenic nodes that are proximal to a gene, or genes, of interest. This achieves two things: One is that it removes the bias introduced by ranking nodes globally. Secondly, is that it produces a set of candidate enhancer regions that may be regulating any given gene of interest. Linking enhancers to their cognate genes is arguably more important and more challenging than simply classifying enhancers. The challenge of linking enhancers and genes is due to limited functional studies that conclusively show the effect of enhancer sequences on any given gene, thus the data available for validation is limited. This is compounded by the limited coverage of high resolution Hi-C in which the number of validated enhancer-promoter pairs is further reduced simply as they are not captured in the Hi-C dataset. Although we did not validate the list of candidate enhancers produced by the network segmentation, curating a database of functionally relevant enhancer-promoter association would represent the next important avenue of research to bolster our findings.

Indeed, the quantity and quality of the data used to validate our classification performance is a limiting factor, not only for the validation of functional enhancer-promoter contacts but also for the enhancer classification in isolation. The validation of our classifications are limited by the enhancer labels used as they do not constitute a perfect set of enhancer annotations. Some enhancers are likely not even identified by the annotations used as we do not include an exhaustive set of enhancer annotations. Therefore, a proportion of the false positive classifications made by 3D-SearchE may actually be true positives. This could be improved by including more enhancer annotations. Encouragingly, the classifications do improve as the number of independent annotations increase although there does remain some ambiguity around why the performance peaks when classifying five annotations rather than

seven. A purely speculative explanation that may be investigated further is that many nodes only contain one of the RNA Pol II variants or multiple RNA Pol II annotations of the same variant limiting the number of distinct enhancer annotations to 5. Further clarity of the classification performance could be provided by testing the performance for each individual state that includes active enhancers (H3K27ac and H3K4me1), primed enhancers (H3K4me1) and enhancer promoter (ePromoter) hybrids (H3K27ac, H3K4me1 and H3K4me3) (Dao et al. 2017). It would be expected that nodes annotated with the primed enhancers would receive a lower IAS than both the active and ePromoter nodes.

Finally, we show that whole gene (WG) mapping, as opposed to TSS mapping, marginally increases the classifying performance of 3D-SearchE. We show that for certain network motifs, the use of WG mapping helps to resolve the topological differences between nodes. We show that by using the coordinates of the TSS we obtain different scores than if we were to use the whole gene coordinates. To understand why we would use either we must first revisit the main aim of this method, which is to identify enhancers that regulate gene expression. Given that enhancers initiate transcription at the TSS and transcriptional machinery are localised at this location we thought that it may be more relevant to set the TSS as the seed nodes. Alternatively, we also know that enhancers have a broader function than just initiating transcription. In modelling both mapping procedures we were able to better understand how the algorithm would react and capture intergenic nodes. The most important result is seen when a gene is split across two nodes and independently interacts with an intergenic node. In this network we can see that the TSS and WG mapping results differ and WG mapping is later shown to be superior in the classification of genic nodes. This is likely because the WG mapping captures the topology of the network better than TSS mapping. This is because the TSS mapping results in a loss of *a priori* information that the algorithm can use as the node that does not contain the TSS is effectively considered non-genic by 3D-SearchE.

Because 3C can have experimental and post-processing variability it is likely that many genes that map to more than one node each of those nodes have independent interactions with intergenic nodes This highlights the complexity of the networks and the need to include more information in order to account for the local patterns of connectivity. This also highlights the need to refine the algorithm not only in the design but also in the parameterisation. In **chapter 4** we used rudimentary parameterisation to limit the propagation and enrich the maximum value of IAS at a distance of 3 nodes from the seed node. As shown in **chapter 3** the networks both between cell-type and between 3C methods can influence the architecture of the networks, particularly in the density. This will mean that the same propagation settings can be optimised for one network and not for another. It would therefore be beneficial to parameterise the propagation settings for each network in order to maximise the classification performance of 3D-SearchE. To avoid the problem of overfitting, the network measures identified in **chapter 3** could be used as a reference point to understand which propagation settings are optimal for each type of network.

While there is scope for improving 3D-SearchE, it is demonstrably a useful feature for classifying enhancers when representing 3C data as ChINs. It performs comparably with other network measures that have previously been shown to be useful in the identification of

other genomic features (Pancaldi et al. 2016; Thibodeau et al. 2017). Most importantly, IAS is the only method that integrates gene expression data to achieve this and provides a new perspective on the relationship between gene expression and the 3D chromatin architecture.

# Chapter 6 - Concluding Remarks and Future Perspectives

---

The challenges that exist in demystifying enhancers stems from a fundamental gap in our knowledge of their composition and mode of action. It is becoming increasingly apparent that the chromatin architecture plays an important role in gene regulation by localising enhancers to gene promoters in specific patterns. These effects can be obvious as shown by the loss of interaction between the ZRS enhancer and the *SHH* promoter following the deletion of a key CTCF site (Ushiki et al. 2021). Additionally, subtle changes in the chromatin architecture may also be a contributing factor in the emergence of complex traits (Boyle, Li, and Pritchard 2017).

The research described in this thesis describes a new feature of enhancers by examining the distinct connectivity patterns of enhancers and genes within chromatin interaction networks. By representing 3C data as networks and utilising network theory approaches we have shown that these distinct connectivity patterns can be used to predict enhancers. This work provides further evidence of the role of chromatin organisation as a communication network between genes and regulatory elements such as enhancers to regulate gene expression. As a result of these findings we have developed a novel method that reverse engineers the transfer of information from enhancers to genes via chromatin. This is achieved using network propagation to transmit information from genic nodes to putative enhancer nodes. We are not aware of any current methods that utilise 3C and gene expression data in this manner. Here, we will discuss the implications of this work in the context of the wider literature and provide future perspectives on the use of our method and other network theory approaches in the investigation of gene regulation.

## 6.1 IAS as a predictive feature of enhancers

Results show that the chromatin architecture can be used to predict enhancers using their distinct connectivity within the network as shown by the centrality measures in **chapter 3**. IAS is able to predict enhancer nodes by measuring both the connectivity of intergenic nodes and their proximity to all other genic nodes within the chromatin interaction networks (**see chapters 3 and 4**). Importantly this is achieved independently of any typical enhancer features such as histone marks, P300 and RNA pol II occupancy. Both IAS and the centrality measures represent a new set of features of enhancers that describe their topological characteristics. This work provides additional evidence of enhancer activity that can be used as a standalone measure to assess the probability of identifying an enhancer in a given segment of DNA. Alternatively, it could be incorporated into existing and future predictive models as an input feature. Topological characteristics have been used previously as input features to a SVM model to predict broad domains and super enhancers using centrality scores and graphlets with reasonable success (Thibodeau et al. 2017). Indeed, we are

currently in the process of using the centrality scores and IAS as input features in a random forest classifier to further validate their use as a feature to predict enhancers as well as to understand which measures contribute the most towards the predictive performance of the classifier. The ability to incorporate network measures into a classifier will also be included in the final package for 3D-SearchE.

### 6.1.1 Limitations and ongoing & future improvements

In order to improve the use of network measures and IAS there are several limitations that need to be addressed. One of the major limitations of these approaches is the availability and quality of the data. 3C experiments are expensive and therefore are only available in a limited number of cell-types. We also show that the type of 3C experiment can affect the number of captured enhancers and the topology of the network (**see chapter 3**). Additionally, we show that there is a tradeoff between identifying more enhancers and the resolution of the fragments to more precisely locate the enhancers. The accuracy of using topological measures to characterise and predict enhancers will be aided by improvements in the resolution and coverage of 3C experiments along with falling costs. Some of these issues may be alleviated by combining multiple 3C datasets. By incorporating, for example, the mESC PCHi-C and DNaseI ChINs the network would maintain the superior coverage provided by the DNaseI ChIN while improving the resolution of the nodes where the two networks overlap. This may not be limited to combining 3C data from the cell type either. We show in **chapter 4** that by combining the PIC networks of monocytes, neutrophils and T-cells into a single consensus network we were not only able to identify cell-type specific active nodes, but that the predictive performance of the model was actually higher. This work may therefore have future applications in predicting enhancer nodes, or active nodes, in cell-types in which obtaining 3C data is not possible or feasible but data for a related cell-type is available.

As well as the 3C data, we also use gene expression data in the form of bulk RNA-seq. Given that we are using gene activity as an input the use of GRO-seq may improve the results obtained using our method (Lopes, Agami, and Korkmaz 2017). GRO-seq works by quantifying the production of nascent RNA and is therefore less susceptible to the low half life of some RNA products.GRO-seq therefore provides a more accurate and less noisy picture of the rate of transcription of genes. Similarly, bulk RNA-seq also captures a lot of noise by averaging the expression of genes across cell populations.

The features used to validate the enhancer predictions are also limited in terms of availability across all cell-types. This is shown by the different validation features used between the mouse ESC and PIC networks as well as the quality and reliability of the data in identifying true enhancers. In **chapter 3** we also show that the enhancer features do not perfectly overlap that demonstrates the ambiguity in the processes used to identify enhancers. To reduce the uncertainty of identifying enhancers more features can be added as more data becomes available. For example, the transcription factor Yin Yang 1 (YY1) has been implicated in mediating enhancer-promoter interactions (Weintraub et al. 2017). The addition of additional enhancer datasets annotated across chromatin interaction networks for many cell-types would be of obvious benefit to the type of analyses outlined in this thesis.

Higher centrality scores and IAS have both been shown to indicate a higher likelihood of finding an enhancer in any given 3C fragment. However, we did not test to see if higher scores are indicative of more active enhancers. This could be achieved by making better use of the features that are already in use. Histone marks, for example, can denote different strengths of enhancers. The use of more chromatin states could be used to differentiate between weak and strong enhancers. We could then better understand if centrality scores and IAS tend to be higher in more active enhancers and can be used to delineate enhancers in this way. H3K27ac, Starr-seq and CAGE-seq data also contain information about the activity of the enhancers and could also be used in this way. Interestingly, these enhancer activity values could theoretically be propagated in reverse, from enhancers to genes. This would add further supporting evidence to the use of propagation to investigate the enhancer-chromatin-promoter relationship. Finally, as well as data that supports the identification of enhancers it would also be beneficial to include negative data in order to identify true negative predictions and more reliably measure the predictive performance of the centrality scores and IAS. Negative annotations such as methylation and heterochromatin marks could be used to label nodes that are unlikely to harbour active enhancers.

Following this work, curating a database of enhancer features within the framework of chromatin interaction networks would be a powerful resource for the study of gene regulation. The use of graph databases, such as Neo4j (“Graph Database Platform” 2020), have been widely adopted in recent years due to their ability to infer indirect relationships between data points. Importantly, the use of graph databases makes it relatively easy to integrate additional data. This can be achieved as additional annotations on the nodes, in this case the 3C fragments, with additional genomic information as discussed previously with YY1. It may also be achieved by aggregating the multiple 3C networks as previously discussed. The most interesting approach involves integrating other biological networks such as protein-protein interaction networks as a separate network layer such as with multiplexed networks (Didier, Brun, and Baudot 2015). Each of these integration approaches, coupled with the rich set of network theory tools would open the door to a wide range of potential discoveries. Done in the correct manner, these databases can also be made more accessible to non-computational biologists through interactive user interfaces to aid in the interpretability and discoverability of these data sets.

The ideal network would be built from a high resolution full Hi-C dataset. Bonev et al. have published the most complete map of chromatin interactions to date in mice with the use of ultra deep Hi-C (Bonev et al. 2017). However, a consequence of higher resolution and coverage is the increase in the size of the network. This inevitably will require more computational resources and should be considered for future experiments using these networks.

## 6.3 Network approaches in understanding the role of chromatin organisation in gene regulation

These findings add further evidence of the role of chromatin architecture in the regulation of gene expression. It is well established that the hierarchical structure of chromatin is

organised in such a way that distinct regions of chromatin, such as with heterochromatin and euchromatin, can regulate gene expression. However, the mechanisms of how this organisation occurs and the nuances of whether chromatin organisation drives gene expression or vice versa is poorly understood. This thesis provides evidence that the use of chromatin interaction networks and gene expression can be used to predict enhancers. It also provides observations that may help to explain the relationship between chromatin organisation and gene expression.

### 6.3.1 Chromatin organisation and its consequences

The chromatin interaction networks contain information about the organisation of chromatin and can be investigated in a manner that is not possible using traditional 3C analytical methodology. Current models such as phase separation (Y. Zhang and Kutateladze 2019) and transcription factories (Sutherland and Bickmore 2009) suggest that gene expression is influenced by changes in the concentrations of chromatin and proteins such as TFs that form distinct membraneless compartments. This tacitly implies that enhancers are enriched in regions of chromatin that are proximal to expressed genes in 3D-space. Indeed, we find that intergenic nodes that are proximal and well connected with genic nodes, as shown by a higher IAS, are more frequently labeled with enhancer features.

We have already discussed the scale-free properties of the networks in **chapter 3**. The formation of scale-free networks are proposed to arise via two main models, preferential attachment and the fitness model. The preferential attachment model explains that nodes which already maintain edges with other nodes are more likely to establish new connections (Krapivsky and Krioukov 2008). The fitness model describes nodes as having an intrinsic property that makes them more able than other nodes to establish new connections (Caldarelli et al. 2002). Both models may be able to explain the emergence of scale-free properties in chromatin interaction networks. For example, enhancers and promoters are known to be able to recruit molecules that mediate enhancer-promoter contacts. This would support the fitness model in that enhancers and promoters contain intrinsic properties that allow them to establish more interactions than other genomic loci. Additionally, the preferential attachment model would support data indicating that younger genes with essential function tend to occur in TADs that contain established ancient genes(James, Trevisan-Herraz, and Rico 2021). Indeed, we have shown that both enhancer and genic nodes have significantly higher degree centrality scores than intergenic non-enhancer nodes (**see chapter 3**). These findings suggest that enhancer and genic nodes may be central to the scale-free properties observed in the chromatin interaction networks. This is in line with evidence that hub nodes (those with degree higher than 60) for example are enriched with features associated with essential cellular function (Sandhu et al. 2012). The emergence of scale-free properties in chromatin interaction networks has several notable benefits. One is that this specific spatial distribution of nodes within the chromatin interaction networks affords a level of redundancy, or fault tolerance, that is beneficial for the reduction of lethal or less-fit phenotypes (Colizza et al. 2006). This is achieved by reducing the probability of a single nucleotide polymorphisms or a copy-number variation perturbing the network structure at a functionally important node i.e. a gene or enhancer. This could be tested by a targeted approach to remove nodes across the network (node deletion). By measuring the importance of nodes through node deletion we could further characterise loci that may contain enhancers or other elements that are important in the regulation of gene expression.

Another benefit is related to the small-world properties that we also observe in the chromatin interaction networks (these two properties are often found together).

This has two benefits: one is that the networks can scale with an almost constant diameter, the other is this facilitates the efficient transmission of information across the network. The ability to scale the genome is tied into the presence of junk DNA and its potential to act as a buffer against mutational load while also harbouring potential future enhancers and genes (Graur 2017). The ability to transmit information efficiently across the network is related to the idea that genomic regulation acts at a systems level, genome wide. Indeed, ideas such as the omnigenic model support this theory. We also show from our findings that the entire chromatin interaction network is required to predict enhancers which suggests that genome wide chromatin organisation plays a role in gene regulation (**see 6.3.2 Global influences on gene expression and the omnigenic model**).

The use of polymer models could be used to create full *in silico* modelling of disease states to understand how SNPs and SVs result in pathogenic phenotypes. For example, SNPs and SVs can often become pathogenic by rewiring the regulatory circuit as shown by the somatic inversion of the GATA2 -110 super enhancer by inv(3) (q21;q26) or t(3;3)(21;q26) (Gröschel et al. 2014; Vinh et al. 2018)) and potentially by the tandem duplication found in KERNA (**See chapter 2**). Polymer model simulations have shown a remarkable ability to recapitulate Hi-C interaction maps as well as predict the reorganisation of chromatin under various states (Chiang et al. 2019; Chris A. Brackley et al. 2016). It may also be prudent to explore the differences in organisation between cell and 3C capture types in greater depth. For example, algorithms based on louvain clustering have been used to identify large scale folding of the genome including topologically associated domains (TADs) (Norton et al. 2018; Yan, Lou, and Gerstein 2017).

## 6.3.2 Global influences on gene expression

We show that using global gene expression profiles can be used to predict enhancers. The premise of our work is based on the idea that enhancers are localised to genes by the folding of chromatin. Typically, we model promoter-enhancer contacts at a local level whereby gene A is regulated by enhancer B through various mechanisms, and chiefly by chromatin loops. At a higher level it is known that TADs can insulate communities of genes and enhancers such that the frequency of contacts between the residents of the TADs are more frequent than those between a resident and an element outside of the TAD. Hypothetically, propagation of gene expression could have future applications in modelling the smaller scale organisation of chromatin such as phase separation, transcription factories and TADs. For example clusters of nodes with high IAS values represent well connected communities of nodes that could be validated using current TAD definitions.

Zoom out another level, and the chromatin is further organised into A and B compartments for open and closed chromatin, respectively. This suggests either a top-down organisation in which the larger structures such as A/B compartments and TADs influence the localisation of enhancers to promoters. Alternatively, it could be interpreted as a bottom-up organisation of chromatin in which the expression of genes dictates the folding of chromatin. These may not be mutually exclusive, the folding of the genome is complex and may employ a plethora of mechanisms and feedback loops to organise itself into a functional structure. In any case,

this reasoning suggests that gene regulation is a global event and should be investigated as such. If the organisation of chromatin into A/B compartments influences the localisation of an enhancer to a gene, or if the expression of genes influences the higher order structures the following remains true: Information about the location of enhancers is encoded within the chromatin network and the relative expression of all the genes. In fact, we show that the prediction of an enhancer node requires the expression of all genic nodes rather than one suggesting that there is a link between the global structure of chromatin and enhancer location. This idea is supported by the multi-loci nature of many complex diseases such as alzehimers that have been uncovered by genome wide association studies (GWAS) (Jansen et al. 2019). Furthermore, *in silico* polymer models of chromatin organisation supports the idea that gene expression is influenced, at least to some degree, by the global organisation of chromatin which has been termed the pan-genomic model (C. A. Brackley et al. 2020). Incidentally, we find that IAS is enriched greatly at loci containing expression QTLs; variants that are correlated with changes in gene expression. This would suggest that 3D-SearchE is a useful tool in investigating the emerging idea of the global regulation of gene expression.

# Bibliography


Alexander, Jeffrey M., Juan Guan, Bingkun Li, Lenka Maliskova, Michael Song, Yin Shen, Bo Huang, Stavros Lomvardas, and Orion D. Weiner. 2019. “Live-Cell Imaging Reveals Enhancer-Dependent Transcription in the Absence of Enhancer Proximity.” *eLife* 8 (May). https://doi.org/10.7554/eLife.41769.

Andersson, Robin, Claudia Gebhard, Irene Miguel-Escalada, Ilka Hoof, Jette Bornholdt, Mette Boyd, Yun Chen, et al. 2014. “An Atlas of Active Enhancers across Human Cell Types and Tissues.” *Nature* 507 (7493): 455–61.

Andersson, Robin, and Albin Sandelin. 2019. “Determinants of Enhancer and Promoter Activities of Regulatory Elements.” *Nature Reviews. Genetics* 21 (2): 71–87.

Anne-Laure Valton, Job Dekker. 2016. “TAD Disruption as Oncogenic Driver.” *Current Opinion in Genetics & Development* 36 (February): 34.

Arnold, C. D., D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark. 2013. “Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-Seq.” *Science* 339 (6123). https://doi.org/10.1126/science.1232542.

Babaei, Sepideh, Ahmed Mahfouz, Marc Hulsman, Boudewijn P. F. Lelieveldt, Jeroen de Ridder, and Marcel Reinders. 2015. “Hi-C Chromatin Interaction Networks Predict Co-Expression in the Mouse Cortex.” *PLoS Computational Biology* 11 (5): e1004221.

Baker, William K. 1968. “Position-Effect Variegation.” In *Advances in Genetics*, 133–69. Advances in Genetics. Elsevier.

Banerji, Julian, Laura Olson, and Walter Schaffner. 1983. “A Lymphocyte-Specific Cellular Enhancer Is Located Downstream of the Joining Region in Immunoglobulin Heavy Chain Genes.” *Cell* 33 (3): 729–40.

Barabasi, A. L., and R. Albert. 1999. “Emergence of Scaling in Random Networks.” *Science* 286 (5439): 509–12.

Barabási, Albert-László. 2009. “Scale-Free Networks: A Decade and beyond.” *Science* 325 (5939): 412–13.

Barabasi, Albert-Laszlo, and Reka Albert. 1999. “Emergence of Scaling in Random Networks. Science.” *Science* 2865439.

Barabási, Albert-László, and Zoltán N. Oltvai. 2004. “Network Biology: Understanding the Cell’s Functional Organization.” *Nature Reviews. Genetics* 5 (2): 101–13.

Beaulieu, Chandree L., Jacek Majewski, Jeremy Schwartzentruber, Mark E. Samuels, Bridget A. Fernandez, Francois P. Bernier, Michael Brudno, et al. 2014. “FORGE Canada Consortium: Outcomes of a 2-Year National Rare-Disease Gene-Discovery Project.” *American Journal of Human Genetics* 94 (6): 809–17.

Bejerano, Gill, Michael Pheasant, Igor Makunin, Stuart Stephen, W. James Kent, John S. Mattick, and David Haussler. 2004. “Ultraconserved Elements in the Human Genome.” *Science* 304 (5675): 1321–25.

Belton, Jon-Matthew, Rachel Patton McCord, Johan Harmen Gibcus, Natalia Naumova, Ye Zhan, and Job Dekker. 2012. “Hi-C: A Comprehensive Technique to Capture the Conformation of Genomes.” *Methods* 58 (3): 268–76.

Benabdallah, Nezha S., Iain Williamson, Robert S. Illingworth, Lauren Kane, Shelagh Boyle, Dipta Sengupta, Graeme R. Grimes, Pierre Therizols, and Wendy A. Bickmore. 2019. “Decreased Enhancer-Promoter Proximity Accompanying Enhancer Activation.” *Molecular Cell* 76 (3): 473–84.e7.

Bianconi, G., and A-L Barabási. 2001. “Competition and Multiscaling in Evolving Networks.” *EPL* 54 (4): 436.

Bickmore, Wendy A. 2013. “The Spatial Organization of the Human Genome.” *Annual Review of Genomics and Human Genetics* 14 (July): 67–84.

Boeva, Valentina, Tatiana Popova, Maxime Lienard, Sebastien Toffoli, Maud Kamal,

Christophe Le Tourneau, David Gentien, et al. 2014. “Multi-Factor Data Normalization Enables the Detection of Copy Number Aberrations in Amplicon Sequencing Data.” *Bioinformatics* 30 (24): 3443–50.
Boll, Werner, and Markus Noll. 2002. “The Drosophila Pox Neuro Gene: Control of Male Courtship Behavior and Fertility as Revealed by a Complete Dissection of All Enhancers.” *Development* 129 (24): 5667–81.
Bonev, Boyan, Netta Mendelson Cohen, Quentin Szabo, Lauriane Fritsch, Giorgio L. Papadopoulos, Yaniv Lubling, Xiaole Xu, et al. 2017. “Multiscale 3D Genome Rewiring during Mouse Neural Development.” *Cell* 171 (3): 557–72.e24.
Boyle, Evan A., Yang I. Li, and Jonathan K. Pritchard. 2017. “An Expanded View of Complex Traits: From Polygenic to Omnigenic.” *Cell* 169 (7): 1177–86.
Brackley, C. A., N. Gilbert, D. Michieletto, A. Papantonis, M. C. F. Pereira, P. R. Cook, and D. Marenduzzo. 2020. “Complex Small-World Regulatory Networks Emerge from the 3D Organisation of the Human Genome.” *bioRxiv*. https://doi.org/10.1101/2020.05.12.091041.
Brackley, Chris A., Jill M. Brown, Dominic Waithe, Christian Babbs, James Davies, Jim R. Hughes, Veronica J. Buckle, and Davide Marenduzzo. 2016. “Predicting the Three-Dimensional Folding of Cis-Regulatory Regions in Mammalian Genomes Using Bioinformatic Data and Polymer Models.” *Genome Biology* 17 (1): 1–16.
Brohée, Sylvain, and Jacques van Helden. 2006. “Evaluation of Clustering Algorithms for Protein-Protein Interaction Networks.” *BMC Bioinformatics* 7 (November): 488.
Broido, Anna D., and Aaron Clauset. 2019. “Scale-Free Networks Are Rare.” *Nature Communications* 10 (1): 1–10.
Buenrostro, Jason D., Beijing Wu, Howard Y. Chang, and William J. Greenleaf. 2015. “ATAC-Seq: A Method for Assaying Chromatin Accessibility Genome-Wide.” *Current Protocols in Molecular Biology / Edited by Frederick M. Ausubel ... [et Al.]* 109 (January): 21.29.1–21.29.9.
Cairns, Jonathan, Paula Freire-Pritchett, Steven W. Wingett, Csilla Várnai, Andrew Dimond, Vincent Plagnol, Daniel Zerbino, et al. 2016. “CHiCAGO: Robust Detection of DNA Looping Interactions in Capture Hi-C Data.” *Genome Biology* 17 (1): 1–17.
Caldarelli, G., A. Capocci, P. De Los Rios, and M. A. Muñoz. 2002. “Scale-Free Networks from Varying Vertex Intrinsic Fitness.” *Physical Review Letters* 89 (25): 258702.
Cameron, Christopher J. F., Josée Dostie, and Mathieu Blanchette. 2020. “HIFI: Estimating DNA-DNA Interaction Frequency from Hi-C Data at Restriction-Fragment Resolution.” *Genome Biology* 21 (1): 1–15.
Carrillo-de-Santa-Pau, Enrique, David Juan, Vera Pancaldi, Felipe Were, Ignacio Martin-Subero, Daniel Rico, Alfonso Valencia, and BLUEPRINT Consortium. 2017. “Automatic Identification of Informative Regions with Epigenomic Changes Associated to Hematopoiesis.” *Nucleic Acids Research* 45 (16): 9244–59.
Carter, David, Lyubomira Chakalova, Cameron S. Osborne, Yan-Feng Dai, and Peter Fraser. 2002. “Long-Range Chromatin Regulatory Interactions in Vivo.” *Nature Genetics* 32 (4): 623–26.
Caulfield, Mark, Jim Davies, Martin Dennys, Leila Elbahy, Tom Fowler, Sue Hill, Tim Hubbard, et al. 2017. “The 100,000 Genomes Project Protocol,” December. https://doi.org/10.6084/m9.figshare.4530893.v4.
Chen, Jihong, and Qiao Li. 2011. “Life and Death of Transcriptional Co-Activator p300.” *Epigenetics: Official Journal of the DNA Methylation Society* 6 (8): 957–61.
Chen, Lu, Bing Ge, Francesco Paolo Casale, Louella Vasquez, Tony Kwan, Diego Garrido-Martín, Stephen Watt, et al. 2016. “Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells.” *Cell* 167 (5): 1398–1414.e24.
Chiang, Michael, Davide Michieletto, Chris A. Brackley, Nattaphong Rattanavirotkul, Hisham Mohammed, Davide Marenduzzo, and Tamir Chandra. 2019. “Polymer Modeling Predicts Chromosome Reorganization in Senescence.” *Cell Reports* 28 (12): 3212–23.e6.
Cho, Eun-Jung, Toshimitsu Takagi, Christine R. Moore, and Stephen Buratowski. 1997.

"mRNA Capping Enzyme Is Recruited to the Transcription Complex by Phosphorylation of the RNA Polymerase II Carboxy-Terminal Domain." *Genes & Development* 11 (24): 3319.
Choromański, Krzysztof, Michał Matuszak, and Jacek Miękisz. 2013. "Scale-Free Graph with Preferential Attachment and Evolving Internal Vertex Structure." *Journal of Statistical Physics* 151 (6): 1175–83.
Colizza, V., A. Flammini, M. A. Serrano, and A. Vespignani. 2006. "Detecting Rich-Club Ordering in Complex Networks." *Nature Physics* 2 (2): 110–15.
Cowen, Lenore, Trey Ideker, Benjamin J. Raphael, and Roded Sharan. 2017. "Network Propagation: A Universal Amplifier of Genetic Associations." *Nature Reviews. Genetics* 18 (9): 551–62.
Crane, Emily, Qian Bian, Rachel Patton McCord, Bryan R. Lajoie, Bayly S. Wheeler, Edward J. Ralston, Satoru Uzawa, Job Dekker, and Barbara J. Meyer. 2015. "Condensin-Driven Remodelling of X Chromosome Topology during Dosage Compensation." *Nature* 523 (7559): 240–44.
Creyghton, Menno P., Albert W. Cheng, G. Grant Welstead, Tristan Kooistra, Bryce W. Carey, Eveline J. Steine, Jacob Hanna, et al. 2010. "Histone H3K27ac Separates Active from Poised Enhancers and Predicts Developmental State." *Proceedings of the National Academy of Sciences of the United States of America* 107 (50): 21931–36.
Crick, F. H. 1958. "On Protein Synthesis." *Symposia of the Society for Experimental Biology* 12: 138–63.
Csardi G, Nepusz T. 2006. "The Igraph Software Package for Complex Network Research." *InterJournal, Complex Systems*. https://igraph.org/.
Dao, Lan T. M., Ariel O. Galindo-Albarrán, Jaime A. Castro-Mondragon, Charlotte Andrieu-Soler, Alejandra Medina-Rivera, Charbel Souaid, Guillaume Charbonnier, et al. 2017. "Genome-Wide Characterization of Mammalian Promoters with Distal Enhancer Functions." *Nature Genetics* 49 (7): 1073–81.
Dekker, Job. 2008. "Gene Regulation in the Third Dimension." *Science* 319 (5871): 1793–94.
Dekker, Job, and Tom Misteli. 2015. "Long-Range Chromatin Interactions." *Cold Spring Harbor Perspectives in Biology* 7 (10): a019356.
Dekker, Job, Karsten Rippe, Martijn Dekker, and Nancy Kleckner. 2002a. "Capturing Chromosome Conformation." *Science* 295 (5558): 1306–11.
———. 2002b. "Capturing Chromosome Conformation." *Science* 295 (5558): 1306–11.
Deng, Wulan, Jongjoo Lee, Hongxin Wang, Jeff Miller, Andreas Reik, Philip D. Gregory, Ann Dean, and Gerd A. Blobel. 2012. "Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor." *Cell* 149 (6): 1233–44.
Despang, Alexandra, Robert Schöpflin, Martin Franke, Salaheddine Ali, Ivana Jerkovic, Christina Paliou, Wing-Lee Chan, et al. 2019. "Functional Dissection of TADs Reveals Non-Essential and Instructive Roles in Regulating Gene Expression." *Genetics*. bioRxiv.
Dickinson, Rachel E., Paul Milne, Laura Jardine, Sasan Zandi, Sabina I. Swierczek, Naomi McGovern, Sharon Cookson, et al. 2014. "The Evolution of Cellular Deficiency in GATA2 Mutation." *Blood*. https://doi.org/10.1182/blood-2013-07-517151.
Didier, Gilles, Christine Brun, and Anaïs Baudot. 2015. "Identifying Communities from Multiplex Biological Networks." *PeerJ* 3 (December): e1525.
Dixon, Jesse R., Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S. Liu, and Bing Ren. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature* 485 (7398): 376–80.
Donadieu, Jean, Marie Lamant, Claire Fieschi, Flore Sicre de Fontbrune, Aurélie Caye, Marie Ouachee, Blandine Beaupain, et al. 2018. "Natural History of GATA2 Deficiency in a Survey of 79 French and Belgian Patients." *Haematologica*. https://doi.org/10.3324/haematol.2017.181909.
Dorschner, Michael O., Michael Hawrylycz, Richard Humbert, James C. Wallace, Anthony Shafer, Janelle Kawamoto, Joshua Mack, et al. 2004. "High-Throughput Localization of Functional Elements by Quantitative Chromatin Profiling." *Nature Methods* 1 (3): 219–25.

Ehret, Charles F., and Gérard De Haller. 1963. “Origin, Development, and Maturation of Organelles and Organelle Systems of the Cell Surface in Paramecium.” *Journal of Ultrastructure Research* 9 (October): 1–42.
ENCODE Project Consortium. 2012. “An Integrated Encyclopedia of DNA Elements in the Human Genome.” *Nature* 489 (7414): 57–74.
Ercal, Gunes, and John Matta. 2013. “Resilience Notions for Scale-Free Networks.” *Procedia Computer Science* 20: 510–15.
Erdős, P., and A. Rényi. 1964. “On the Strength of Connectedness of a Random Graph.” *Acta Mathematica Academiae Scientiarum Hungaricae* 12 (1-2): 261–67.
Ernst, Jason, and Manolis Kellis. 2012. “ChromHMM: Automating Chromatin-State Discovery and Characterization.” *Nature Methods* 9 (3): 215–16.
———. 2017. “Chromatin-State Discovery and Genome Annotation with ChromHMM.” *Nature Protocols* 12 (12): 2478–92.
Fudenberg, Geoffrey, Maxim Imakaev, Carolyn Lu, Anton Goloborodko, Nezar Abdennur, and Leonid A. Mirny. 2016. “Formation of Chromosomal Domains by Loop Extrusion.” *Cell Reports* 15 (9): 2038–49.
Fukaya, Takashi, Bomyi Lim, and Michael Levine. 2016. “Enhancer Control of Transcriptional Bursting.” *Cell* 166 (2): 358–68.
Fulco, Charles P., Joseph Nasser, Thouis R. Jones, Glen Munson, Drew T. Bergman, Vidya Subramanian, Sharon R. Grossman, et al. 2019. “Activity-by-Contact Model of Enhancer-Promoter Regulation from Thousands of CRISPR Perturbations.” *Nature Genetics* 51 (12): 1664–69.
Gall, J. G., and M. L. Pardue. 1969. “Formation and Detection of RNA-DNA Hybrid Molecules in Cytological Preparations.” *Proceedings of the National Academy of Sciences of the United States of America* 63 (2): 378–83.
Ghavi-Helm, Yad. 2020. “Functional Consequences of Chromosomal Rearrangements on Gene Expression: Not So Deleterious After All?” *Journal of Molecular Biology* 432 (3): 665–75.
Ghavi-Helm, Yad, Aleksander Jankowski, Sascha Meiers, Rebecca R. Viales, Jan O. Korbel, and Eileen E. M. Furlong. 2019. “Highly Rearranged Chromosomes Reveal Uncoupling between Genome Topology and Gene Expression.” *Nature Genetics* 51 (8): 1272–82.
Ghavi-Helm, Yad, Felix A. Klein, Tibor Pakozdi, Lucia Ciglar, Daan Noordermeer, Wolfgang Huber, and Eileen E. M. Furlong. 2014. “Enhancer Loops Appear Stable during Development and Are Associated with Paused Polymerase.” *Nature* 512 (7512): 96–100.
Gibcus, Johan H., Kumiko Samejima, Anton Goloborodko, Itaru Samejima, Natalia Naumova, Johannes Nuebler, Masato T. Kanemaki, et al. 2018. “A Pathway for Mitotic Chromosome Formation.” *Science* 359 (6376). https://doi.org/10.1126/science.aao6135.
Gonzalez-Sandoval, Adriana, and Susan M. Gasser. 2016. “On TADs and LADs: Spatial Control Over Gene Expression.” *Trends in Genetics: TIG* 32 (8): 485–95.
“Graph Database Platform.” 2020. May 16, 2020. https://neo4j.com/.
Graur, Dan. 2017. “An Upper Limit on the Functional Fraction of the Human Genome.” *Genome Biology and Evolution* 9 (7): 1880–85.
Graur, Dan, Yichen Zheng, and Ricardo B. R. Azevedo. 2015. “An Evolutionary Classification of Genomic Function.” *Genome Biology and Evolution* 7 (3): 642–45.
Graur, Dan, Yichen Zheng, Nicholas Price, Ricardo B. R. Azevedo, Rebecca A. Zufall, and Eran Elhaik. 2013. “On the Immortality of Television Sets: ‘Function’ in the Human Genome according to the Evolution-Free Gospel of ENCODE.” *Genome Biology and Evolution* 5 (3): 578–90.
Gröschel, Stefan, Mathijs A. Sanders, Remco Hoogenboezem, Elzo de Wit, Britta A. M. Bouwman, Claudia Erpelinck, Vincent H. J. van der Velden, et al. 2014. “A Single Oncogenic Enhancer Rearrangement Causes Concomitant EVI1 and GATA2 Deregulation in Leukemia.” *Cell* 157 (2): 369–81.
Grosschedl, R., and M. L. Birnstiel. 1980. “Spacer DNA Sequences Upstream of the T-A-T-A-A-A-T-A Sequence Are Essential for Promotion of H2A Histone Gene

Transcription in Vivo." *Proceedings of the National Academy of Sciences of the United States of America* 77 (12): 7102–6.
Halfon, Marc S. 2019. "Studying Transcriptional Enhancers: The Founder Fallacy, Validation Creep, and Other Biases." *Trends in Genetics: TIG* 35 (2): 93–103.
Hariprakash, Judith Mary, and Francesco Ferrari. 2019. "Computational Biology Solutions to Identify Enhancers-Target Gene Pairs." *Computational and Structural Biotechnology Journal* 17 (June): 821–31.
Harlen, Kevin M., and L. Stirling Churchman. 2017. "The Code and beyond: Transcription Regulation by the RNA Polymerase II Carboxy-Terminal Domain." *Nature Reviews. Molecular Cell Biology* 18 (4): 263–73.
Harmston, Nathan. 2020. "Regulation in Common: Sponge to Zebrafish." *Science* 370 (6517): 657–58.
Harmston, Nathan, Elizabeth Ing-Simmons, Ge Tan, Malcolm Perry, Matthias Merkenschlager, and Boris Lenhard. 2017. "Topologically Associating Domains Are Ancient Features That Coincide with Metazoan Clusters of Extreme Noncoding Conservation." *Nature Communications* 8 (1): 1–13.
Heintzman, Nathaniel D., Rhona K. Stuart, Gary Hon, Yutao Fu, Christina W. Ching, R. David Hawkins, Leah O. Barrera, et al. 2007. "Distinct and Predictive Chromatin Signatures of Transcriptional Promoters and Enhancers in the Human Genome." *Nature Genetics* 39 (3): 311–18.
Hnisz, Denes, Krishna Shrinivas, Richard A. Young, Arup K. Chakraborty, and Phillip A. Sharp. 2017. "A Phase Separation Model for Transcriptional Control." *Cell* 169 (1): 13–23.
Hoang, Stephen A., and Stefan Bekiranov. 2013. "The Network Architecture of the Saccharomyces Cerevisiae Genome." *PloS One* 8 (12): e81972.
Hofree, Matan, John P. Shen, Hannah Carter, Andrew Gross, and Trey Ideker. 2013. "Network-Based Stratification of Tumor Mutations." *Nature Methods* 10 (11): 1108–15.
Hsu, Amy P., Lisa J. McReynolds, and Steven M. Holland. 2015. "GATA2 Deficiency." *Current Opinion in Allergy and Clinical Immunology* 15 (1): 104.
Ibn-Salem, Jonas, Sebastian Köhler, Michael I. Love, Ho-Ryun Chung, Ni Huang, Matthew E. Hurles, Melissa Haendel, et al. 2014. "Deletions of Chromosomal Regulatory Boundaries Are Associated with Congenital Disease." *Genome Biology* 15 (9): 423.
Inoue, Fumitaka, and Nadav Ahituv. 2015. "Decoding Enhancers Using Massively Parallel Reporter Assays." *Genomics* 106 (3): 159–64.
James, Caelinn, Marco Trevisan-Herraz, and Daniel Rico. 2021. "Vertebrate Whole Genome Duplications Shaped the Current 3D Genome Architecture." *bioRxiv*. bioRxiv. https://doi.org/10.1101/2021.06.11.448047.
Jansen, Iris E., Jeanne E. Savage, Kyoko Watanabe, Julien Bryois, Dylan M. Williams, Stacy Steinberg, Julia Sealock, et al. 2019. "Genome-Wide Meta-Analysis Identifies New Loci and Functional Pathways Influencing Alzheimer's Disease Risk." *Nature Genetics* 51 (3): 404–13.
Javierre, Biola M., Oliver S. Burren, Steven P. Wilder, Roman Kreuzhuber, Steven M. Hill, Sven Sewitz, Jonathan Cairns, et al. 2016. "Lineage-Specific Genome Architecture Links Enhancers and Non-Coding Disease Variants to Target Gene Promoters." *Cell* 167 (5): 1369–84.e19.
Jeong, H., B. Tombor, R. Albert, Z. N. Oltvai, and A. L. Barabási. 2000. "The Large-Scale Organization of Metabolic Networks." *Nature* 407 (6804): 651–54.
Jiang, Qinghua, Shuilin Jin, Yongshuai Jiang, Mingzhi Liao, Rennan Feng, Liangcai Zhang, Guiyou Liu, and Junwei Hao. 2017. "Alzheimer's Disease Variants with the Genome-Wide Significance Are Significantly Enriched in Immune Pathways and Active in Immune Cells." *Molecular Neurobiology* 54 (1): 594–600.
Johanson, Timothy M., Hannah D. Coughlan, Aaron T. L. Lun, Naiara G. Bediaga, Gaetano Naselli, Alexandra L. Garnham, Leonard C. Harrison, Gordon K. Smyth, and Rhys S. Allan. 2018. "Genome-Wide Analysis Reveals No Evidence of Trans Chromosomal Regulation of Mammalian Immune Development." *PLoS Genetics* 14 (6): e1007431.

Joshi, Onkar, Shuang-Yin Wang, Tatyana Kuznetsova, Yaser Atlasi, Tianran Peng, Pierre J. Fabre, Ehsan Habibi, et al. 2015. “Dynamic Reorganization of Extremely Long-Range Promoter-Promoter Interactions between Two States of Pluripotency.” *Cell Stem Cell* 17 (6): 748–57.

Juan, David, Juliane Perner, Enrique Carrillo de Santa Pau, Simone Marsili, David Ochoa, Ho-Ryun Chung, Martin Vingron, Daniel Rico, and Alfonso Valencia. 2016. “Epigenomic Co-Localization and Co-Evolution Reveal a Key Role for 5hmC as a Communication Hub in the Chromatin Network of ESCs.” *Cell Reports* 14 (5): 1246–57.

Karpen, G. H. 1994. “Position-Effect Variegation and the New Biology of Heterochromatin.” *Current Opinion in Genetics & Development* 4 (2): 281–91.

Kellis, Manolis, Barbara Wold, Michael P. Snyder, Bradley E. Bernstein, Anshul Kundaje, Georgi K. Marinov, Lucas D. Ward, et al. 2014. “Defining Functional DNA Elements in the Human Genome.” *Proceedings of the National Academy of Sciences of the United States of America* 111 (17): 6131–38.

Kim, Tae-Kyung, Martin Hemberg, Jesse M. Gray, Allen M. Costa, Daniel M. Bear, Jing Wu, David A. Harmin, et al. 2010. “Widespread Transcription at Neuronal Activity-Regulated Enhancers.” *Nature* 465 (7295): 182–87.

Kleftogiannis, Dimitrios, Panos Kalnis, and Vladimir B. Bajic. 2015. “DEEP: A General Computational Framework for Predicting Enhancers.” *Nucleic Acids Research* 43 (1): e6.

Kolodziejczyk, Aleksandra A., Jong Kyoung Kim, Jason C. H. Tsang, Tomislav Ilicic, Johan Henriksson, Kedar N. Natarajan, Alex C. Tuck, et al. 2015. “Single Cell RNA-Sequencing of Pluripotent States Unlocks Modular Transcriptional Variation.” *Cell Stem Cell* 17 (4): 471–85.

Komarnitsky, P., E. J. Cho, and S. Buratowski. 2000. “Different Phosphorylated Forms of RNA Polymerase II and Associated mRNA Processing Factors during Transcription.” *Genes & Development* 14 (19): 2452–60.

Korkmaz, Gozde, Rui Lopes, Alejandro P. Ugalde, Ekaterina Nevedomskaya, Ruiqi Han, Ksenia Myacheva, Wilbert Zwart, Ran Elkon, and Reuven Agami. 2016. “Functional Genetic Screens for Enhancer Elements in the Human Genome Using CRISPR-Cas9.” *Nature Biotechnology* 34 (2): 192–98.

Kozyra, Emilia J., Victor B. Pastor, Stylianos Lefkopoulos, Sushree S. Sahoo, Hauke Busch, Rebecca K. Voss, Miriam Erlacher, et al. 2020. “Synonymous GATA2 Mutations Result in Selective Loss of Mutated RNA and Are Common in Patients with GATA2 Deficiency.” *Leukemia* 34 (10): 2673–87.

Krapivsky, Paul, and Dmitri Krioukov. 2008. “Scale-Free Networks as Preasymptotic Regimes of Superlinear Preferential Attachment.” *Physical Review. E, Statistical, Nonlinear, and Soft Matter Physics* 78 (2 Pt 2): 026114.

Krefting, Jan, Miguel A. Andrade-Navarro, and Jonas Ibn-Salem. 2018. “Evolutionary Stability of Topologically Associating Domains Is Associated with Conserved Gene Regulation.” *BMC Biology* 16 (1): 1–12.

Kruse, Kai, Sven Sewitz, and M. Madan Babu. 2013. “A Complex Network Framework for Unbiased Statistical Analyses of DNA-DNA Contact Maps.” *Nucleic Acids Research* 41 (2): 701–10.

Lambert, Samuel A., Arttu Jolma, Laura F. Campitelli, Pratyush K. Das, Yimeng Yin, Mihai Albu, Xiaoting Chen, Jussi Taipale, Timothy R. Hughes, and Matthew T. Weirauch. 2018. “The Human Transcription Factors.” *Cell* 172 (4): 650–65.

Laurenti, Elisa, Sergei Doulatov, Sasan Zandi, Ian Plumb, Jing Chen, Craig April, Jian-Bing Fan, and John E. Dick. 2013. “The Transcriptional Architecture of Early Human Hematopoiesis Identifies Multilevel Control of Lymphoid Commitment.” *Nature Immunology* 14 (7): 756–63.

Lee, Dongwon, Rachel Karchin, and Michael A. Beer. 2011. “Discriminative Prediction of Mammalian Enhancers from DNA Sequence.” *Genome Research* 21 (12): 2167–80.

Lek, Monkol, Konrad J. Karczewski, Eric V. Minikel, Kaitlin E. Samocha, Eric Banks, Timothy Fennell, Anne H. O’Donnell-Luria, et al. 2016. “Analysis of Protein-Coding Genetic

Variation in 60,706 Humans." *Nature* 536 (7616): 285–91.
Le, Tung B. K., Maxim V. Imakaev, Leonid A. Mirny, and Michael T. Laub. 2013. "High-Resolution Mapping of the Spatial Organization of a Bacterial Chromosome." *Science* 342 (6159): 731–34.
Lieberman-Aiden, Erez, Nynke L. van Berkum, Louise Williams, Maxim Imakaev, Tobias Ragoczy, Agnes Telling, Ido Amit, et al. 2009a. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.
———. 2009b. "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." *Science* 326 (5950): 289–93.
Li, Heng, and Richard Durbin. 2009. "Fast and Accurate Short Read Alignment with Burrows-Wheeler Transform." *Bioinformatics* 25 (14): 1754–60.
Li, Heng, Bob Handsaker, Alec Wysoker, Tim Fennell, Jue Ruan, Nils Homer, Gabor Marth, Goncalo Abecasis, Richard Durbin, and 1000 Genome Project Data Processing Subgroup. 2009. "The Sequence Alignment/Map Format and SAMtools." *Bioinformatics* 25 (16): 2078–79.
Lim, Leonard Whye Kit, Hung Hui Chung, Yee Ling Chong, and Nung Kion Lee. 2018. "A Survey of Recently Emerged Genome-Wide Computational Enhancer Predictor Tools." *Computational Biology and Chemistry* 74 (June): 132–41.
Lopes, R., R. Agami, and G. Korkmaz. 2017. "GRO-Seq, A Tool for Identification of Transcripts Regulating Gene Expression." *Methods in Molecular Biology* 1543. https://doi.org/10.1007/978-1-4939-6716-2_3.
Lundberg, Scott M., William B. Tu, Brian Raught, Linda Z. Penn, Michael M. Hoffman, and Su-In Lee. 2016. "ChromNet: Learning the Human Chromatin Network from All ENCODE ChIP-Seq Data." *Genome Biology* 17 (April): 82.
MacDonald, Jeffrey R., Robert Ziman, Ryan K. C. Yuen, Lars Feuk, and Stephen W. Scherer. 2014. "The Database of Genomic Variants: A Curated Collection of Structural Variation in the Human Genome." *Nucleic Acids Research* 42 (Database issue): D986–92.
Medina-Rivera, Alejandra, David Santiago-Algarra, Denis Puthier, and Salvatore Spicuglia. 2018. "Widespread Enhancer Activity from Core Promoters." *Trends in Biochemical Sciences* 43 (6): 452–68.
Mikecz, A. von, S. Zhang, M. Montminy, E. M. Tan, and P. Hemmerich. 2000. "CREB-Binding Protein (CBP)/p300 and RNA Polymerase II Colocalize in Transcriptionally Active Domains in the Nucleus." *The Journal of Cell Biology* 150 (1): 265–73.
Min, Xu, Wanwen Zeng, Shengquan Chen, Ning Chen, Ting Chen, and Rui Jiang. 2017. "Predicting Enhancers with Deep Convolutional Neural Networks." *BMC Bioinformatics* 18 (13): 35–46.
Mizuguchi, Takeshi, Geoffrey Fudenberg, Sameet Mehta, Jon-Matthew Belton, Nitika Taneja, Hernan Diego Folco, Peter FitzGerald, et al. 2014. "Cohesin-Dependent Globules and Heterochromatin Shape 3D Genome Architecture in S. Pombe." *Nature* 516 (7531): 432–35.
Moore, Jill E., Henry E. Pratt, Michael J. Purcaro, and Zhiping Weng. 2020. "A Curated Benchmark of Enhancer-Gene Interactions for Evaluating Enhancer-Target Gene Prediction Methods." *Genome Biology* 21 (1): 1–16.
Muench, David E., and H. Leighton Grimes. 2015. "Transcriptional Control of Stem and Progenitor Potential." *Current Stem Cell Reports* 1 (3): 139–50.
Nagano, Takashi, Yaniv Lubling, Csilla Várnai, Carmel Dudley, Wing Leung, Yael Baran, Netta Mendelson Cohen, Steven Wingett, Peter Fraser, and Amos Tanay. 2017. "Cell-Cycle Dynamics of Chromosomal Organization at Single-Cell Resolution." *Nature* 547 (7661): 61–67.
Narlikar, Geeta J., Hua-Ying Fan, and Robert E. Kingston. 2002. "Cooperation between Complexes That Regulate Chromatin Structure and Transcription." *Cell* 108 (4): 475–87.
Navlakha, Saket, and Carl Kingsford. 2010. "The Power of Protein Interaction Networks for Associating Genes with Diseases." *Bioinformatics* 26 (8): 1057–63.

Ng, Ashley P., and Warren S. Alexander. 2017. “Haematopoietic Stem Cells: Past, Present and Future.” *Cell Death Discovery* 3 (February): 17002.
Ni, Zuyao, Jonathan B. Olsen, Xinghua Guo, Guoqing Zhong, Eric Dongliang Ruan, Edyta Marcon, Peter Young, et al. 2011. “Control of the RNA Polymerase II Phosphorylation State in Promoter Regions by CTD Interaction Domain-Containing Proteins RPRD1A and RPRD1B.” *Transcription* 2 (5): 237–42.
Noguchi, Shuhei, Takahiro Arakawa, Shiro Fukuda, Masaaki Furuno, Akira Hasegawa, Fumi Hori, Sachi Ishikawa-Kato, et al. 2017. “FANTOM5 CAGE Profiles of Human and Mouse Samples.” *Scientific Data* 4 (1): 1–10.
Nora, Elphège P., Bryan R. Lajoie, Edda G. Schulz, Luca Giorgetti, Ikuhiro Okamoto, Nicolas Servant, Tristan Piolot, et al. 2012. “Spatial Partitioning of the Regulatory Landscape of the X-Inactivation Centre.” *Nature* 485 (7398): 381–85.
Norton, Heidi K., Daniel J. Emerson, Harvey Huang, Jesi Kim, Katelyn R. Titus, Shi Gu, Danielle S. Bassett, and Jennifer E. Phillips-Cremins. 2018. “Detecting Hierarchical Genome Folding with Network Modularity.” *Nature Methods* 15 (2): 119–22.
Ohno, S. 1972. “So Much ‘Junk’ DNA in Our Genome.” *Brookhaven Symposia in Biology* 23: 366–70.
Ono, Keiichiro, Tanja Muetze, Georgi Kolishovski, Paul Shannon, and Barry Demchak. 2015. “CyREST: Turbocharging Cytoscape Access for External Tools via a RESTful API.” *F1000Research* 4 (August): 478.
Orkin, S. H. 2000. “Diversification of Haematopoietic Stem Cells to Specific Lineages.” *Nature Reviews. Genetics* 1 (1): 57–64.
Osterwalder, Marco, Iros Barozzi, Virginie Tissières, Yoko Fukuda-Yuzawa, Brandon J. Mannion, Sarah Y. Afzal, Elizabeth A. Lee, et al. 2018. “Enhancer Redundancy Provides Phenotypic Robustness in Mammalian Development.” *Nature* 554 (7691): 239–43.
Ouyang, Z., Q. Zhou, and W. H. Wong. 2009. “ChIP-Seq of Transcription Factors Predicts Absolute and Differential Gene Expression in Embryonic Stem Cells.” *Proceedings of the National Academy of Sciences*. https://doi.org/10.1073/pnas.0904863106.
Page, Lawrence, Sergey Brin, Rajeev Motwani, and Terry Winograd. 1999. “The PageRank Citation Ranking: Bringing Order to the Web,” November. http://ilpubs.stanford.edu:8090/422.
Palikyras, Spiros, and Argyris Papantonis. 2019. “Modes of Phase Separation Affecting Chromatin Regulation.” *Open Biology* 9 (10): 190167.
Pancaldi, Vera, Enrique Carrillo-de-Santa-Pau, Biola Maria Javierre, David Juan, Peter Fraser, Mikhail Spivakov, Alfonso Valencia, and Daniel Rico. 2016. “Integrating Epigenomic Data and 3D Genomic Structure with a New Measure of Chromatin Assortativity.” *Genome Biology* 17 (1): 152.
Panigrahi, Anil, and Bert W. O’Malley. 2021. “Mechanisms of Enhancer Action: The Known and the Unknown.” *Genome Biology* 22 (1): 108.
Papayannopoulos, Venizelos. 2018. “Neutrophil Extracellular Traps in Immunity and Disease.” *Nature Reviews. Immunology* 18 (2): 134–47.
Peña-Castillo, Lourdes, Murat Tasan, Chad L. Myers, Hyunju Lee, Trupti Joshi, Chao Zhang, Yuanfang Guan, et al. 2008. “A Critical Assessment of Mus Musculus Gene Function Prediction Using Integrated Genomic Evidence.” *Genome Biology* 9 Suppl 1 (June): S2.
Pengelly, Ana Raquel, Ömer Copur, Herbert Jäckle, Alf Herzig, and Jürg Müller. 2013. “A Histone Mutant Reproduces the Phenotype Caused by Loss of Histone-Modifying Factor Polycomb.” *Science* 339 (6120): 698–99.
Peng, Tianran, Yanan Zhai, Yaser Atlasi, Menno Ter Huurne, Hendrik Marks, Hendrik G. Stunnenberg, and Wout Megchelenbrink. 2020. “STARR-Seq Identifies Active, Chromatin-Masked, and Dormant Enhancers in Pluripotent Mouse Embryonic Stem Cells.” *Genome Biology* 21 (1): 243.
Pennacchio, Len A., Nadav Ahituv, Alan M. Moses, Shyam Prabhakar, Marcelo A. Nobrega, Malak Shoukry, Simon Minovitsky, et al. 2006. “In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences.” *Nature* 444 (7118): 499–502.
Pope, Benjamin D., Tyrone Ryba, Vishnu Dileep, Feng Yue, Weisheng Wu, Olgert Denas,

Daniel L. Vera, et al. 2014. “Topologically Associating Domains Are Stable Units of Replication-Timing Regulation.” *Nature* 515 (7527): 402–5.
Raisner, Ryan, Samir Kharbanda, Lingyan Jin, Edwin Jeng, Emily Chan, Mark Merchant, Peter M. Haverty, et al. 2018. “Enhancer Activity Requires CBP/P300 Bromodomain-Dependent Histone H3K27 Acetylation.” *Cell Reports* 24 (7): 1722–29.
Rajagopal, Nisha, Wei Xie, Yan Li, Uli Wagner, Wei Wang, John Stamatoyannopoulos, Jason Ernst, Manolis Kellis, and Bing Ren. 2013. “RFECS: A Random-Forest Based Algorithm for Enhancer Identification from Chromatin State.” *PLoS Computational Biology* 9 (3). https://doi.org/10.1371/journal.pcbi.1002968.
Rao, Suhas S. P., Miriam H. Huntley, Neva C. Durand, Elena K. Stamenova, Ivan D. Bochkov, James T. Robinson, Adrian L. Sanborn, et al. 2014. “A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping.” *Cell* 159 (7): 1665–80.
Rico, Daniel, Joost H. A. Martens, Kate Downes, Enrique Carrillo-de-Santa-Pau, Vera Pancaldi, Alessandra Breschi, David Richardson, et al. 2017. “Comparative Analysis of Neutrophil and Monocyte Epigenomes.” *bioRxiv*. bioRxiv. https://doi.org/10.1101/237784.
Robinson, James T., Helga Thorvaldsdóttir, Wendy Winckler, Mitchell Guttman, Eric S. Lander, Gad Getz, and Jill P. Mesirov. 2011. “Integrative Genomics Viewer.” *Nature Biotechnology* 29 (1): 24–26.
Robson, Michael I., Alessa R. Ringel, and Stefan Mundlos. 2019. “Regulatory Landscaping: How Enhancer-Promoter Communication Is Sculpted in 3D.” *Molecular Cell* 74 (6): 1110–22.
Ron, Gil, Yuval Globerson, Dror Moran, and Tommy Kaplan. 2017. “Promoter-Enhancer Interactions Identified from Hi-C Data Using Probabilistic Models and Hierarchical Topological Domains.” *Nature Communications* 8 (1): 1–12.
Sadowski, Michal, Agnieszka Kraft, Przemyslaw Szalaj, Michal Wlasnowolski, Zhonghui Tang, Yijun Ruan, and Dariusz Plewczynski. 2019. “Spatial Chromatin Architecture Alteration by Structural Variations in Human Genomes at the Population Scale.” *Genome Biology* 20 (1): 148.
Sandhu, Kuljeet Singh, Guoliang Li, Huay Mei Poh, Yu Ling Kelly Quek, Yee Yen Sia, Su Qin Peh, Fabianus Hendriyan Mulawadi, et al. 2012. “Large-Scale Functional Organization of Long-Range Chromatin Interaction Networks.” *Cell Reports* 2 (5): 1207–19.
Schmidt, Florian, Nina Gasparoni, Gilles Gasparoni, Kathrin Gianmoena, Cristina Cadenas, Julia K. Polansky, Peter Ebert, et al. 2017. “Combining Transcription Factor Binding Affinities with Open-Chromatin Data for Accurate Gene Expression Prediction.” *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkw1061.
Schoenfelder, Stefan, and Peter Fraser. 2019. “Long-Range Enhancer–promoter Contacts in Gene Expression Control.” *Nature Reviews. Genetics* 20 (8): 437–55.
Schoenfelder, Stefan, Mayra Furlan-Magaril, Borbala Mifsud, Filipe Tavares-Cadete, Robert Sugar, Biola-Maria Javierre, Takashi Nagano, et al. 2015. “The Pluripotent Regulatory Circuitry Connecting Promoters to Their Long-Range Interacting Elements.” *Genome Research* 25 (4): 582–97.
Schoenfelder, Stefan, Biola-Maria Javierre, Mayra Furlan-Magaril, Steven W. Wingett, and Peter Fraser. 2018. “Promoter Capture Hi-C: High-Resolution, Genome-Wide Profiling of Promoter Interactions.” *Journal of Visualized Experiments: JoVE*, no. 136 (June). https://doi.org/10.3791/57320.
Sexton, Tom, Eitan Yaffe, Ephraim Kenigsberg, Frédéric Bantignies, Benjamin Leblanc, Michael Hoichman, Hugues Parrinello, Amos Tanay, and Giacomo Cavalli. 2012a. “Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.” *Cell* 148 (3): 458–72.
———. 2012b. “Three-Dimensional Folding and Functional Organization Principles of the Drosophila Genome.” *Cell* 148 (3): 458–72.
Shoaib, Muhammad, Nidhi Nair, and Claus Storgaard Sørensen. 2020. “Chromatin Landscaping At Mitotic Exit Orchestrates Genome Function.” *Frontiers in Genetics* 0.

https://doi.org/10.3389/fgene.2020.00103.
Sigal, Yaron M., Ruobo Zhou, and Xiaowei Zhuang. 2018. “Visualizing and Discovering Cellular Structures with Super-Resolution Microscopy.” *Science* 361 (6405): 880–87.
Sjoerd Johannes Bastiaan Holwerda, Wouter de Laat. 2013. “CTCF: The Protein, the Binding Partners, the Binding Sites and Their Chromatin Loops.” *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 368 (1620). https://doi.org/10.1098/rstb.2012.0369.
Song, Jimin, and Mona Singh. 2009. “How and When Should Interactome-Derived Clusters Be Used to Predict Functional Modules and Protein Function?” *Bioinformatics* 25 (23): 3143–50.
Song, Wei, Roded Sharan, and Ivan Ovcharenko. 2019. “The First Enhancer in an Enhancer Chain Safeguards Subsequent Enhancer-Promoter Contacts from a Distance.” *Genome Biology* 20 (1): 197.
Spielmann, Malte, and Stefan Mundlos. 2013. “Structural Variations, the Regulatory Landscape of the Genome and Their Alteration in Human Disease.” *BioEssays: News and Reviews in Molecular, Cellular and Developmental Biology* 35 (6): 533–43.
Spinner, Michael A., Lauren A. Sanchez, Amy P. Hsu, Pamela A. Shaw, Christa S. Zerbe, Katherine R. Calvo, Diane C. Arthur, et al. 2014. “GATA2 Deficiency: A Protean Disorder of Hematopoiesis, Lymphatics, and Immunity.” *Blood* 123 (6): 809–21.
Stumpf, Michael P. H., and Mason A. Porter. 2012. “Mathematics. Critical Truths about Power Laws.” *Science* 335 (6069): 665–66.
Sudmant, Peter H., Tobias Rausch, Eugene J. Gardner, Robert E. Handsaker, Alexej Abyzov, John Huddleston, Yan Zhang, et al. 2015. “An Integrated Map of Structural Variation in 2,504 Human Genomes.” *Nature* 526 (7571): 75–81.
Sutherland, Heidi, and Wendy A. Bickmore. 2009. “Transcription Factories: Gene Expression in Unions?” *Nature Reviews. Genetics* 10 (7): 457–66.
Takaya Saito, Marc Rehmsmeier. 2015. “The Precision-Recall Plot Is More Informative than the ROC Plot When Evaluating Binary Classifiers on Imbalanced Datasets.” *PloS One* 10 (3). https://doi.org/10.1371/journal.pone.0118432.
Thibodeau, Asa, Eladio J. Márquez, Dong-Guk Shin, Paola Vera-Licona, and Duygu Ucar. 2017. “Chromatin Interaction Networks Revealed Unique Connectivity Patterns of Broad H3K4me3 Domains and Super Enhancers in 3D Chromatin.” *Scientific Reports* 7 (1): 14466.
Thomas, C. A., Jr. 1971. “The Genetic Organization of Chromosomes.” *Annual Review of Genetics* 5 (1): 237–56.
Trapnell, Cole. 2015. “Defining Cell Types and States with Single-Cell Genomics.” *Genome Research* 25 (10): 1491–98.
Turnbull, Clare, Richard H. Scott, Ellen Thomas, Louise Jones, Nirupa Murugaesu, Freya Boardman Pretty, Dina Halai, et al. 2018. “The 100 000 Genomes Project: Bringing Whole Genome Sequencing to the NHS.” *BMJ* 361 (April): k1687.
Ushiki, Aki, Yichi Zhang, Chenling Xiong, Jingjing Zhao, Ilias Georgakopoulos-Soares, Lauren Kane, Kirsty Jamieson, et al. 2021. “Deletion of CTCF Sites in the SHH Locus Alters Enhancer-Promoter Interactions and Leads to Acheiropodia.” *Nature Communications* 12 (1): 2282.
Van der Auwera, Geraldine A., Mauricio O. Carneiro, Chris Hartl, Ryan Poplin, Guillermo Del Angel, Ami Levy-Moonshine, Tadeusz Jordan, et al. 2013. “From FastQ Data to High Confidence Variant Calls: The Genome Analysis Toolkit Best Practices Pipeline.” *Current Protocols in Bioinformatics / Editoral Board, Andreas D. Baxevanis ... [et Al.]* 43: 11.10.1–33.
Viksna, Juris, Gatis Melkus, Edgars Celms, Kārlis Čerāns, Karlis Freivalds, Paulis Kikusts, Lelde Lace, Mārtiņš Opmanis, Darta Rituma, and Peteris Rucevskis. 2019. “Topological Structure Analysis of Chromatin Interaction Networks.” *BMC Bioinformatics* 20 (23): 1–17.
Vinh, Donald C., Laura Palma, John Storring, and William D. Foulkes. 2018. “GATA2 Deficiency Due to de Novo Complete Monoallelic Deletion in an Adolescent With

Myelodysplasia." *Journal of Pediatric Hematology/Oncology*. https://doi.org/10.1097/mph.0000000000001136.
Visel, Axel, Matthew J. Blow, Zirong Li, Tao Zhang, Jennifer A. Akiyama, Amy Holt, Ingrid Plajzer-Frick, et al. 2009. "ChIP-Seq Accurately Predicts Tissue-Specific Activity of Enhancers." *Nature* 457 (7231): 854–58.
Wang, Chengqi, Michael Q. Zhang, and Zhihua Zhang. 2013. "Computational Identification of Active Enhancers in Model Organisms." *Genomics, Proteomics & Bioinformatics* 11 (3): 142–50.
Wang, Congmao, Chang Liu, Damian Roqueiro, Dominik Grimm, Rebecca Schwab, Claude Becker, Christa Lanz, and Detlef Weigel. 2015. "Genome-Wide Analysis of Local Chromatin Packing in Arabidopsis Thaliana." *Genome Research* 25 (2): 246–56.
Wang, Siyuan, Jun-Han Su, Brian J. Beliveau, Bogdan Bintu, Jeffrey R. Moffitt, Chao-Ting Wu, and Xiaowei Zhuang. 2016. "Spatial Organization of Chromatin Domains and Compartments in Single Chromosomes." *Science* 353 (6299): 598.
Wang, Zhibin, Chongzhi Zang, Kairong Cui, Dustin E. Schones, Artem Barski, Weiqun Peng, and Keji Zhao. 2009. "Genome-Wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes." *Cell* 138 (5): 1019–31.
Watts, Duncan J., and Steven H. Strogatz. 1998. "Collective Dynamics of 'small-World' Networks." *Nature* 393 (6684): 440–42.
Weintraub, Abraham S., Charles H. Li, Alicia V. Zamudio, Alla A. Sigova, Nancy M. Hannett, Daniel S. Day, Brian J. Abraham, et al. 2017. "YY1 Is a Structural Regulator of Enhancer-Promoter Loops." *Cell* 171 (7): 1573–88.e28.
Westra, Harm-Jan, Marjolein J. Peters, Tõnu Esko, Hanieh Yaghootkar, Claudia Schurmann, Johannes Kettunen, Mark W. Christiansen, et al. 2013. "Systematic Identification of Trans eQTLs as Putative Drivers of Known Disease Associations." *Nature Genetics* 45 (10): 1238–43.
Williamson, Iain, Laura A. Lettice, Robert E. Hill, and Wendy A. Bickmore. 2016. "Shh and ZRS Enhancer Colocalisation Is Specific to the Zone of Polarising Activity." *Development* 143 (16): 2994–3001.
Wlodarski, Marcin W., Shinsuke Hirabayashi, Victor Pastor, Jan Starý, Henrik Hasle, Riccardo Masetti, Michael Dworzak, et al. 2016. "Prevalence, Clinical Characteristics, and Prognosis of GATA2-Related Myelodysplastic Syndromes in Children and Adolescents." *Blood* 127 (11): 1387–97; quiz 1518.
Yáñez-Cuna, J. Omar, Evgeny Z. Kvon, and Alexander Stark. 2013. "Deciphering the Transcriptional Cis-Regulatory Code." *Trends in Genetics: TIG* 29 (1): 11–22.
Yang, Yaping, Donna M. Muzny, Jeffrey G. Reid, Matthew N. Bainbridge, Alecia Willis, Patricia A. Ward, Alicia Braxton, et al. 2013. "Clinical Whole-Exome Sequencing for the Diagnosis of Mendelian Disorders." *The New England Journal of Medicine* 369 (16): 1502–11.
Yan, Koon-Kiu, Shaoke Lou, and Mark Gerstein. 2017. "MrTADFinder: A Network Modularity Based Approach to Identify Topologically Associating Domains in Multiple Resolutions." *PLoS Computational Biology* 13 (7): e1005647.
Young, Robert S., Yatendra Kumar, Wendy A. Bickmore, and Martin S. Taylor. 2017. "Bidirectional Transcription Initiation Marks Accessible Chromatin and Is Not Specific