Investigation of transcription factor binding at distal regulatory elements

Joanna Louise Mitchelmore

The Babraham Institute

Jesus College

University of Cambridge

This dissertation is submitted for the degree of

Doctor of Philosophy

September 2017

Investigation of transcription factor binding at distal regulatory elements

Joanna Louise Mitchelmore

Summary

Cellular development and function necessitate precise patterns of gene expression. Control of gene expression is in part orchestrated by a class of remote regulatory elements, termed enhancers, which are brought into contact with promoters via DNA looping. Enhancers typically contain clusters of transcription factor binding sites, and TF recruitment to them is thought to play a key role in transcriptional control.

In this thesis I have addressed two issues regarding gene regulation by enhancers. First, with recent genome-wide enhancer mapping, it is becoming increasingly apparent that genes are commonly regulated by multiple enhancers in the same cell type. How a gene's regulatory information is encoded across multiple enhancers, however, is still not fully understood. Second, numerous recent studies have found that enhancers are enriched for expression-modulating and disease-associated genetic variants. However, understanding and predicting the effects of enhancer variants remains a major challenge.

I focussed on a human lymphoblastoid cell line (LCL), GM12878, for which ChIP-Seq data are available for 52 different TFs from the ENCODE project. Significantly, Promoter Capture Hi-C data for the same LCL are available, making it possible to link enhancers to target genes globally. In the first part of the thesis, I investigated how gene regulatory information is encoded across enhancers. Specifically, I asked whether a gene tends to use multiple enhancers to bring the same or distinct regulatory information. I found that there was a general trend towards a "shadow" enhancer architecture, whereby similar combinations of TFs were recruited to multiple enhancers. However, numerous examples of "integrating" enhancers were observed, where the same gene showed large variation in TF binding across enhancers. Distinct groups of TFs were associated with these contrasting models of TF enhancer binding.

To investigate the functional effects of variation at enhancers, I additionally took advantage of a panel of LCLs derived from 359 individuals, which have been genotyped by the 1000 Genomes Project, and for which RNA-Seq data are publically available. I used TF binding models to computationally predict variants impacting TF binding, and tested the association of these variants with the expression of the target genes they contact based on Promoter Capture Hi-C. Compared

to the standard eQTL calling approach, this offers increased sensitivity as only variants physically contacting the promoter and predicted to impact TF binding are tested. Using this approach, I discovered a set of predicted TF-binding affinity variants at distal regions that associate with gene expression. Interestingly, a large proportion of these binding variants fall at the promoters of other genes. This finding suggests that some promoters may be able to act in an enhancer-like manner via long-range interactions, consistent with very recent findings from alternative approaches.

Declaration

This dissertation is the results of my own work and includes nothing, which is the outcome of work done in collaboration except where specifically indicated.

This thesis is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text.

This thesis does not exceed 60,000 words, excluding bibliography, figures and appendices.

Cambridge, UK, September 2017

Joanna Louise Mitchelmore

Acknowledgements

I would like thank Mikhail, for believing in me and giving me the opportunity to enter into the computational biology world. Throughout the process his constant support and encouragement, openness to new ideas, infectious enthusiasm and ability to turn any setback into a positive have helped make this PhD so enjoyable. I could not have wished for a better supervisor. I would like to thank all other members of the lab for making it such a fun environment to discuss ideas and problems. In particular, I would like to thank my mentor Paula, who helped me hugely in getting to grips with everything computational and was always happy to listen to my endless questions and give her time to help. She also made the office a really fun place to be and became a good friend. A special thanks to Jonathan for all his help in everything statistics related – I am not sure what I would have done without you! Also thanks to Lina, Michiel and Pawel – I could not have wished for better lab mates to share this experience with. Thanks as well for the constant encouragement and support during the writing process.

Thanks also go to all my other friends and colleagues at Babraham, they are definitely a huge part of what has made my time here so enjoyable. In particular thanks to Joana, Jack, Azad, Steph, Stevie-bevie, Joerg, and Tom, whom I now count as my close friends. Also thanks go to my friends outside of Cambridge, especially Ellie and Alice, who put up with numerous stressed during the writing process and always encouraged me. Thanks to my parents, for their support in everything I do and my sister for always being there. Finally, a huge thanks to Marco for his constant support, always making me laugh, and putting with my slightly crazed dysfunctional-self during the writing process. This whole thing would have been much harder without you.

Acknowledgement of assistance

1) Initial training in techniques and subsequent mentoring:

Dr Paula Friere Pritchett–Training and help in R; Mentoring

Dr Jonathan Cairns-Help and advice in statistical analysis

Dr Chris Wallace-Advice in statistical analysis

Dr Simon Andrews, Dr Felix Krueger – Computational advice

Dr Olivia Casanueva- Assessor

Dr.Mikhail Spivakov– Training and advice in computational analysis; Mentoring

1 Table of Contents

1 Introduction

The cell types in multi-cellular organisms differ dramatically in both structure and function, yet nearly all contain the same DNA. How then do the differences between cell types arise? The answer lies in gene regulation; each cell type "turns on" (expresses) a unique subset of genes encoding the proteins necessary for its function. The ability to switch on specific sets of genes is also critical in enabling cells to respond to changes in their environment. This is important both for multicellular organisms, where it enables cellular homeostasis to be maintained, and unicellular organisms, where it facilitates efficient use of cellular resources in response to the environment. There are many stages of gene expression –transcription, RNA processing, RNA transport and localisation, translation and mRNA degradation –all of which can be regulated. However transcription, as the first step where DNA is transcribed to mRNA, is perhaps one of the most important points of regulation for most genes.

1.1 Distal gene regulation

1.1.1 Transcription

Transcription is catalysed by the enzyme RNA polymerase (RNA pol). While in bacteria there is a single RNA polymerase, in eukaryotes there are three different polymerases - RNA pol I, RNA pol II and RNA pol III (Ebright, 2000; Roeder & Rutter, 1969, 1970). RNA pol I and III transcribe a limited number of genes, encoding transfer RNAs, ribosomal RNAs and small nuclear RNAs (Warner, 1999; Weinmann & Roeder, 1974). In contrast RNA Pol II transcribes all protein-coding genes, as well as the majority of non-coding RNAs (Hahn, 2004). Transcription is initiated at the core promoter, a ~100bp region immediately upstream of the transcription start site (TSS). None of the RNA pols are able to recognise and bind to the core promoter DNA by themselves, instead requiring accessory factors. While bacterial RNA pol requires only a single additional protein (σ factor) for transcription initiation, eukaryotic RNA pol II requires the coordinated action of at least six different proteins, termed general transcription factors (GTFs, i.e., TFIIA, -B, -D, -E, -F and –H) (Burgess et al., 1969; Orphanides, Lagrange, & Reinberg, 1996; Roeder, 1996). The GTFs assemble into a transcription preinitiation complex (PIC), which through a series of GTF-DNA interactions anchors RNA pol II to the double stranded promoter DNA (Hampsey, 1998; Lee & Young, 2000). Following assembly of the PIC, TFIIH unwinds 10-15bp of DNA surrounding the TSS in order to position the single-stranded

template DNA in the active cleft of RNA pol II (termed the open complex) to initiate RNA synthesis (Grünberg & Hahn, 2013). After the synthesis of ~30bp of RNA, RNA pol II releases its contacts with the core promoter and transitions into the elongation stage (Grünberg & Hahn, 2013). In metazoans, this transition often involves the pausing of RNA pol II just downstream of the promoter, where it remains stably associated with the nascent RNA, and is capable of resuming elongation upon further signals (Adelman & Lis, 2012). As such there are several steps at which transcription initiation can be influenced in eukaryotes.

1.1.2 Transcription in the context of chromatin

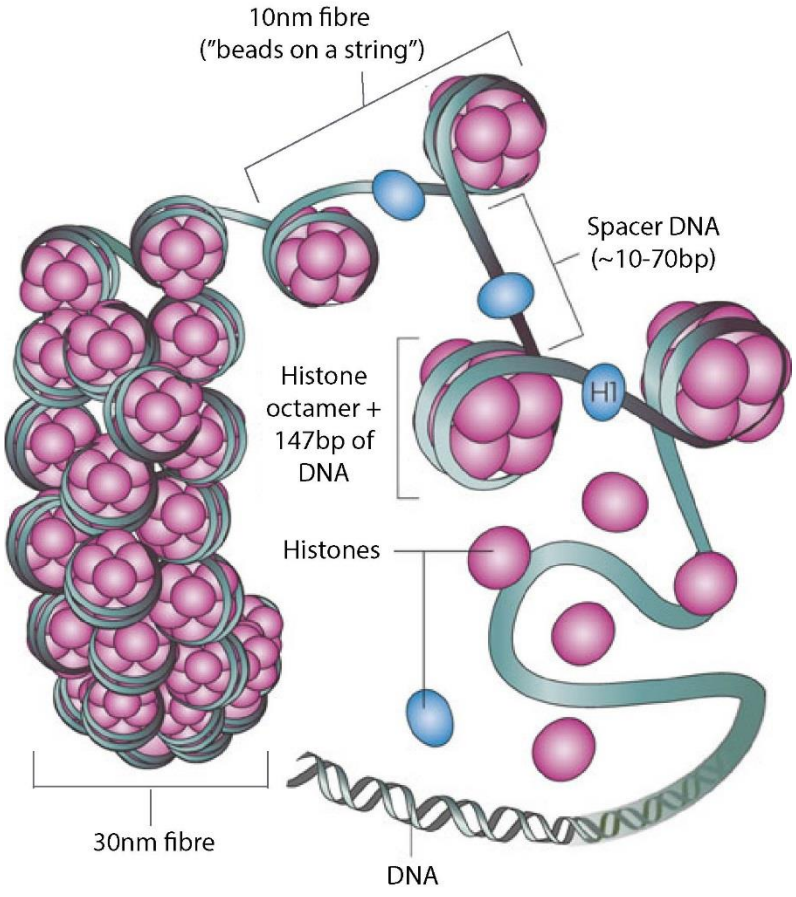


Figure 1.1. Chromatin structure.

147bp of DNA is wrapped around the histone octamer core, which consists of two copies each of histone H2A, H2B, H3 and H4, to form the nucleosome particle. Nucleosomes are connected by ~10-70bp of linker DNA. A fifth histone, H1, binds to the linker DNA at the site where it enters and exits the nucleosome. Repeating nucleosomes form the "beads on a string" 10nm fibre, which can further condense and shorten to form the 30nm fibre. Figure adapted from Figueiredo, Cross, & Janzen (2009).

In eukaryotes genomic DNA is not “naked” but instead is wrapped around histone proteins, resulting in a DNA-protein complex termed chromatin (Felsenfeld & Groudine, 2003) (Figure 1.1). The basic repeating structural unit of chromatin is the nucleosome, which consists of a histone octamer and ~200 base pairs of DNA (Noll, 1974). The histone octamer is made up of two molecules each of H2A, H2B, H3 and H4 (the core histones); 147bp of DNA is wrapped 1.65 times around this histone octamer to form the core nucleosome particle (Arents et al., 1991; Luger et al., 1997). The nucleosome cores are connected by 10-70bp of linker DNA, which associates with linker histone H1, to form the nucleosome (Hergeth & Schneider, 2015). Repeating nucleosomes form a 10nm diameter fibre, which under a microscope resembles “beads on a string” (with the nucleosome core particles as the beads, and the linker DNA as the string) (Olins & Olins, 1974; Woodcock, Safer, & Stanchfield, 1976). In vitro this 10nm fibre was shown to form a shorter, thicker helical fibre approximately 30nm in diameter, termed the “30 nanometre fibre” (Finch & Klug, 1976; Gerchman & Ramakrishnan, 1987). However whether this 30nm fibre exists in vivo remains unclear (Fussner, Ching, & Bazett-Jones, 2011; Maeshima, Hihara, & Eltsov, 2010; Nishino et al., 2012). The chromatin fibre can then be further condensed – either via the 30nm fibre, alternative secondary structure or directly from the 10nm fibre- along with scaffold proteins into higher order structures (Luger, Dechassa, & Tremethick, 2012).

For transcription initiation to occur RNA pol II and the other components of the PIC need to be able to access the DNA; thus chromatin structure can pose a significant barrier to this process (Knezetic & Luse, 1986; Lorch, LaPointe, & Kornberg, 1987). As such, while the packaging of DNA into chromatin is critical in enabling the DNA to fit into the nucleus, it also provides an additional opportunity to exert control over transcription initiation (Li, Carey, & Workman, 2007). Nucleosomes form the main point at which this control can be exerted, either through their precise positioning on the DNA, blocking/providing access to the core promoter, or through their ability to influence the degree of chromatin packaging (Bannister & Kouzarides, 2011; Perner & Chung, 2013).

The histones proteins making up nucleosome core particle have flexible N-terminal tails that project from the nucleosome; many residues in these tails can be post-translationally modified (Allfrey, Faulkner, & Mirsky, 1964). These modifications include methylation of arginine residues; methylation, acetylation and ubiquitination of lysines; and phosphorylation of serines and threonines (Kouzarides, 2007). With the exception of methylation, all the modifications result in a net reduction in positive charge of the histone octamer, which reduces the affinity of the histone octamer for the negatively charged DNA as well as possibly for other nucleosomes (Bowman & Poirier, 2015; Kouzarides, 2007). This “loosening” of DNA-histone and nucleosome-nucleosome

interactions, makes nucleosomes easier to displace from the DNA, and may lead to a more open chromatin structure (Bowman & Poirier, 2015; Kouzarides, 2007; Lawrence, Daujat, & Schneider, 2016). Some evidence to support this idea comes from the finding that acetylated histones are easier to displace in vivo (Reinke & Hörz, 2003; Zhao, Herrera-Diaz, & Gross, 2005). Increased histone acetylation at promoter regions has also been linked to active transcription, while histones in inaccessible heterochromatin are generally unacetylated (Liang et al., 2004; Pokholok et al., 2005; Roh, Cuddapah, & Zhao, 2005). Histone modifications can be added or removed by histone modifying enzymes, thus providing a mechanism through which chromatin structure can be altered to influence transcription (Butler et al. , 2012; Li et al., 2007; Pennisi, 1997). One example of such is p300, which as a histone acetyl transferase (HAT), adds acetyl groups to histones (Bannister & Kouzarides, 1996). The HAT function of this enzyme was showed to be required for transcriptional activation in at least some instances (Kraus, Manning, & Kadonaga, 1999).

As well as increasing accessibility to the DNA through "loosening" histone-DNA interactions within and between nucleosomes, DNA accessibility can also be controlled by changing the precise positioning of nucleosomes on the DNA. This is achieved by a class of chromatin regulators termed chromatin remodelling complexes, which use ATP to either slide the nucleosome along the DNA, or to transiently eject the nucleosome from the DNA (Becker & Workman, 2013; Clapier et al., 2017). In general promoter regions, in particular the region immediately upstream of the TSS, tend to be depleted of nucleosomes relative to transcribed regions. These findings are consistent across a range of lower and higher eukaryotes, including yeast and humans (Lee et al., 2007; Ozsolak et al., 2007; Yuan et al., 2005). Exactly how nucleosome re-positioning/eviction influences transcription initiation and gene activity though remains unclear. Perhaps the most intuitive mechanism is that the nucleosome blocks binding of the PIC components, and thus eviction of the nucleosome enables PIC assembly and the consequent initiation of transcription (Kornberg & Lorch, 1999). Indeed at the yeast genes *PHO5* and *HSP82,* histones are evicted from the promoter region upon gene activation and reassembled when the gene is turned off (Adkins & Tyler, 2006; Boeger et al., 2004; Reinke & Hörz, 2003; Zhao et al., 2005). Also it was shown that a nucleosome at the core promoter of human interferon-beta (IFN-beta) slides in vivo in response to viral infection, and that this is necessary for transcriptional activation (Lomvardas & Thanos, 2001). However a genome-wide study in yeast found that strikingly many partial PICs were assembled in the presence of nucleosomes (Zanton & Pugh, 2006). These partial PICs lacked TFIIH and RNA pol II, implying that nucleosome displacement might only be necessary when the template DNA is engaged in the RNA pol II active site (Zanton & Pugh, 2006). It is thus likely that the role of nucleosomes in regulating

transcription initiation is more complex than initially thought, with the existence of multiple different mechanisms which can be employed within a given organism.

1.1.3 Transcription factors and regulatory elements

It is evident that there are many points at which the rate of transcription initiation can be influenced, including recruitment of RNA pol and the GTFs, stabilisation of the PIC, release of proximal paused RNA pol II in metazoans, and the modification of chromatin– how are these processes influenced in a gene specific manner? The answer to this question lies in a set of regulatory proteins, termed transcription factors (TFs). TFs recognise and bind to short specific DNA sequences (referred to as TF binding motifs), and once bound can either activate or repress transcription (Spitz & Furlong, 2012a; Vaquerizas et al., 2009). TFs are able to recognise and bind to their motif due to extensive complementarity between the surface of the protein (termed the DNA binding domain) and surface features of the DNA double helix in the region of the specific nucleotide sequence of the motif (Todeschini, Georges, & Veitia, 2014). This complementarity results in a series of contacts between the TF and the DNA, most often involving ionic bonds, hydrogen bonds and hydrophobic interactions, ensuring a highly specific and strong interaction.

In prokaryotes TFs often exert their effect on transcription via a direct interaction with RNA pol. This can be by either providing an additional contact surface for RNA pol, helping it bind to the DNA, or by contacting RNA pol to facilitate its transition into an actively transcribing form to activate transcription (Seshasayee, Sivaraman, & Luscombe, 2011). Prokaryotic TFs can repress transcription by binding over the promoter region, blocking access of RNA pol and inhibiting transcription initiation (Marmorstein & Sigler, 1989; Rojo, 1999). In eukaryotes the step-wise assembly of the PIC provides many more points at which TFs can act to increase/decrease the rate of transcription initiation (Figure 1.2). Eukaryotic TFs can interact directly with the components of the PIC and RNA pol II, facilitating their recruitment and/or stability in the PIC, as well as aiding in the release of RNA pol II from proximal pausing (Adelman & Lis, 2012). However in addition to directly interacting with the PIC components, TFs can also interact indirectly via a complex termed Mediator (Flanagan et al., 1991; Kelleher, Flanagan, & Kornberg, 1990; Koleske & Young, 1994; Nonet & Young, 1989; Poss, Ebmeier, & Taatjes, 2013; Soutourina, 2017; Thompson et al., 1993). Mediator is a multi-subunit protein complex (consisting of 25 subunits in budding yeast and up to 30 subunits in humans) that acts as an interface between TFs and the PIC components in all eukaryotes, providing an increased contact surface area for TFs to act on (Poss et al., 2013;

Soutourina, 2017) Its main function is to transduce activating signals from TFs to the PIC. TF-recruited mediator establishes interactions with the PIC components, contributing to their recruitment and assembly, and thus enhancing transcription initiation (Poss et al., 2013; Soutourina, 2017). Along with

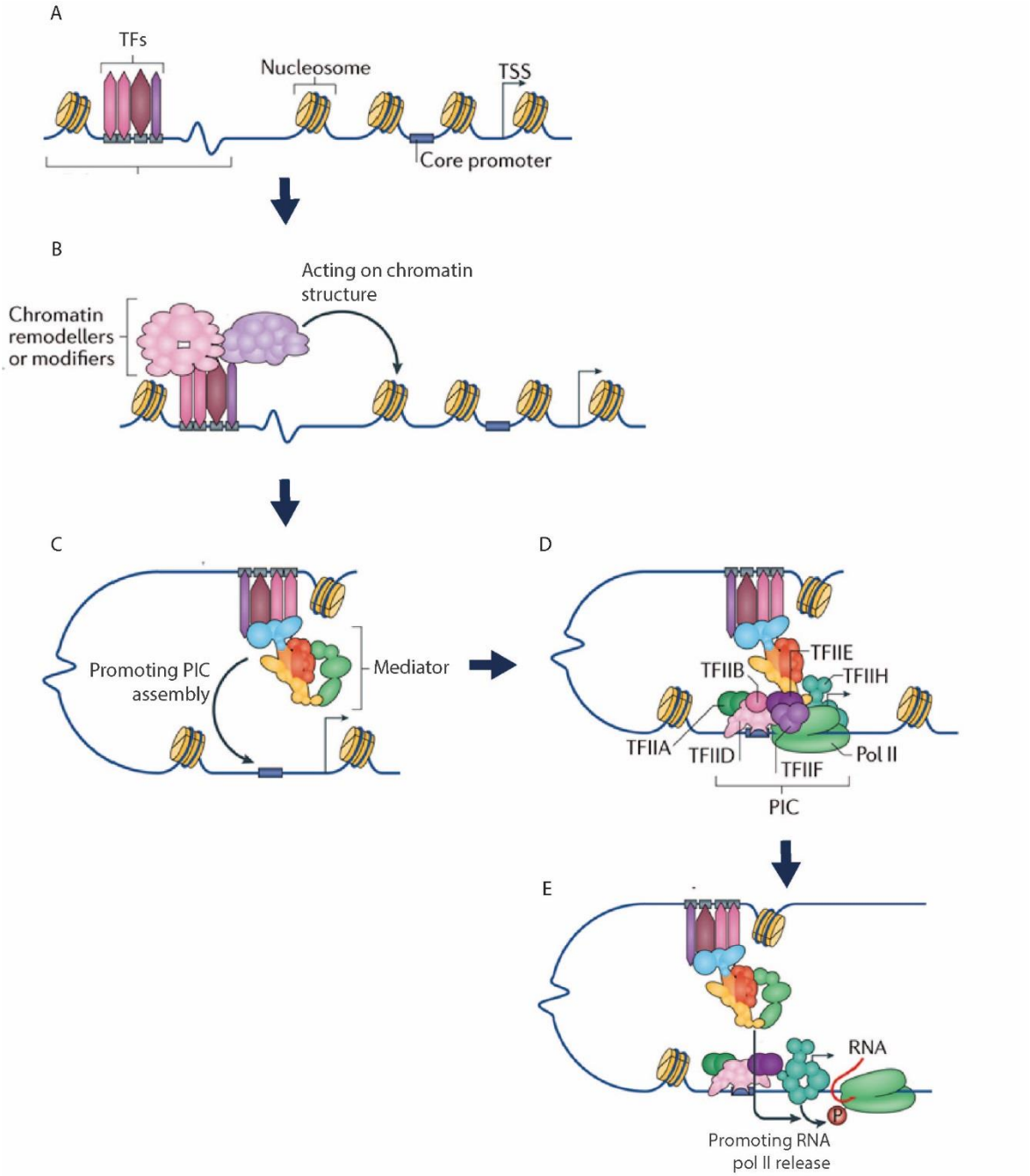


Figure 1.2. The role of TFs in transcriptional activation.

(A) TFs bind to their corresponding sequence motifs in the DNA. Upon binding they can recruit chromatin remodellers and/or chromatin modifying enzymes (B) which can increase DNA accessibility at the core promoter (C). TFs, either directly or via Mediator, can recruit the components of the PIC and facilitate its assembly (C and D). Finally, TFs can promote the release of RNA pol II into an active elongation state (E). Figure adapted from Soutourina (2017).

recruiting Mediator, there is also evidence that TFs may be able to influence transcription via Mediator by inducing large-scale conformational changes in Mediator upon binding, shifting the PIC from an inactive to active state (Meyer et al., 2010; Poss et al., 2013; Taatjes et al., 2002; Taatjes, Schneider-Poetsch, & Tjian, 2004). As well as interacting directly/indirectly with the PIC components, eukaryotic TFs can modulate transcription indirectly by recruiting chromatin modifying enzymes and remodellers (through direct interactions or via Mediator) to modify the chromatin architecture surrounding the promoter region (Agalioti et al., 2000; Berger, 2007; Lomvardas & Thanos, 2002). Modulating DNA accessibility at the promoter can influence not just the binding of the PIC components, but also the binding of other TFs which can further act to activate/repress transcription (Li et al., 2007). The histone modifications themselves can also serve as signals to recruit additional chromatin remodelling enzymes and/or TFs, further influencing transcription initiation (Kouzarides, 2007; Taverna et al., 2007).

TFs are recruited to genes by DNA sequences containing TF binding motifs, termed cis-regulatory elements. In prokaryotes and lower eukaryotes such as yeast, the majority of genes contain a single cis regulatory sequence that is located proximal (within ~100bp) to the core promoter (Bulger & Groudine, 2011; Venters & Pugh, 2009). In contrast a typical metazoan gene is regulated by several cis-regulatory elements, one of which is usually adjacent to the core promoter (often termed the promoter/promoter proximal region), and the other/s of which can be located strikingly several kilobases, either upstream or downstream, from the core promoter (Bulger & Groudine, 2011; Shlyueva, Stampfel, & Stark, 2014; Venters & Pugh, 2009). These distal cis-regulatory elements are termed enhancers (Long, Prescott, & Wysocka, 2016; Shlyueva et al., 2014).

1.1.4 Enhancers

The first enhancer, a 72bp sequence from the SV40 virus, was discovered over 35 years ago (Banerji, Rusconi, & Schaffner, 1981). It was observed that this sequence was able to drive ectopic expression of a cloned rabbit beta globin gene in HeLa cells, independent of its orientation and distance from the gene. Subsequently endogenous elements in the mouse immunoglobulin heavy chain which were also able to stimulate transcription independent of orientation and at distances of thousands of base pairs away from the gene, were identified (Banerji, Olson, & Schaffner, 1983). Interestingly, the Ig enhancer was only able to drive activity in lymphocyte-derived cell lines, providing the first evidence that that enhancer activity shows cell-type/tissue specificity (Banerji et al., 1983; Gillies et al., 1983). Since then, a large number of cell-type or developmental stage-

specific enhancers have been shown to regulate gene expression in multi-cellular organisms (Long, Prescott, & Wysocka, 2016). This has included enhancers that act over very large distances, for example the limb bud enhancer of the developmental gene *Shh* in mice and humans (Lettice et al., 2002, 2003). The *Shh* limb bud enhancer resides in an intron of another gene, *Lmbr1*, almost 900kb away from *Shh*. It initiates and controls the spatial expression pattern of *Shh* in the posterior margin of the limb bud, where *Shh* signalling is critical for establishing anterior-posterior patterning and determining digit identity (Lettice et al., 2002, 2003; Sagai et al., 2009) (Figure 1.3). In addition to the limb bud enhancer, *Shh* is regulated by at least ten other enhancer elements extending over a 900kb region of DNA (Anderson et al., 2014) (Figure 1.3). These enhancers direct expression of *Shh* in a modular fashion, with different enhancers directing expression of *Shh* in different spatial regions (Figure 1.3); this modular organisation is a key feature of enhancer regulation (Shlyueva et al., 2014a).

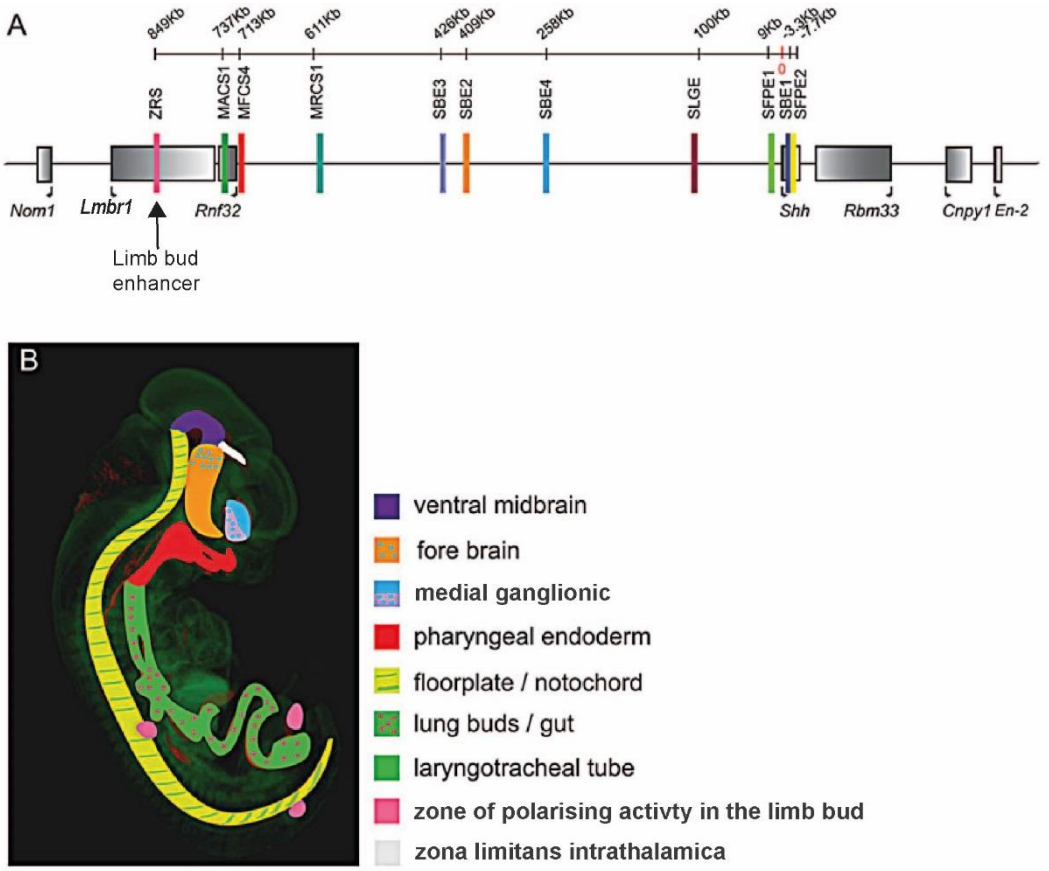


Figure 1.3. Enhancer regulation of Shh.

(A) The murine regulatory locus of *Shh*. Enhancers are depicted as coloured bars. Genes are indicated by grey rectangles, shaded from dark to light in the 5' to 3' orientation. (B) Schematic illustrating the sites of *Shh* expression in the E11.5 mouse embryo. The colours used to depict the regions of expression match the colour/s of the enhancer/s (in A) that direct expression in that particular region. As such a hatched/dot pattern indicates that multiple enhancers drive expression of *Shh* in that region. The enhancer/s responsible

for expression of *Shh* in the zona limitans intrathalamica have not yet been discovered. Figure adapted from Anderson et al. (2014).

Interestingly while enhancers were originally defined as being able to act independent of orientation, several studies have identified enhancers that appear to act in an orientation-dependent manner, suggesting that while enhancers are generally able to activate transcription independent of orientation, this may not be the case for all enhancers (Hozumi et al., 2013; Sauter et al., 2013; Swamynathan & Piatigorsky, 2002). The cell-type specific activity of enhancers is in contrast to promoter regions, which tend to more ubiquitously active (Heintzman et al., 2009; Thurman et al., 2012; Visel et al., 2009). Collectively these results have led to the realisation that enhancers play a key role in the control of spatiotemporal expression patterns in multicellular organisms.

1.1.5 Evolution of enhancers

A recent analysis of the closest unicellular relative of metazoans, Capsaspora, found that their regulatory elements largely lie proximal to genes, suggesting that distal regulation is indeed an animal evolutionary innovation (Sebé-Pedrós et al., 2016). The appearance of distal enhancers – given their key role in controlling cell-type specific expression in metazoans –is thus postulated to be one of the key features that enabled the emergence of animal multi-cellularity (Bulger & Groudine, 2011; Carroll, 2008; Sebé-Pedrós et al., 2016). As well as playing a key role in enabling the development of multi-cellularity, enhancers are also thought to have been critical in facilitating animal diversity (Levine, Cattoglio, & Tjian, 2014). Due to the modular nature of enhancers, mutations at a given enhancer may change expression of a gene in a particular region with no or very little effect on other regions (Carroll, 2008; Cho, 2012). This is in contrast to either mutations in the DNA coding for a TF, where a change in activity of the TF can affect expression of all target genes across all regions where the TF is active, or at the promoter of a gene, where expression of the gene across all regions can be affected (Cho, 2012). In fitting with this reasoning the mean lifetime of enhancers was found to be three times shorter than that of promoters across 20 mammalian species in liver (Villar et al., 2015). Strikingly almost half of enhancers in each species appeared to be recently evolved, suggestive of a role in generating species-specific differences (Villar et al., 2015).

One example highlighting the role of enhancers in morphological evolution is the pelvic enhancer in stickleback fish. Freshwater stickleback fish lack the pelvic bony spines that are present in

saltwater and ancestral fish (Bell, 1987). In marine sticklebacks development of the pelvic spines depends on expression of the TF Pitx1 in the pelvic region, which is under the control of a specific enhancer; freshwater sticklebacks have lost this enhancer (Chan et al., 2010; Shapiro et al., 2004). In addition to the loss or gain of enhancers (and consequent loss/gain of gene expression in a particular region), more subtle changes in enhancer activity can also lead to phenotypic variation. One example of such is the *spot* enhancer that controls expression of the *yellow* gene in the *Drosophila* wing. The *yellow* gene encodes an enzyme involved in the synthesis of a black pigment (Wittkopp, Vaccaro, & Carroll, 2002). In *D. melanogaster* the *spot* enhancer directs low levels of expression of the *yellow* gene throughout the wing, resulting in even pigmentation. In contrast, in *D. biarmipes* the *spot* enhancer directs high levels of *yellow* expression in the corner of the wing, resulting in a dark spot (Wittkopp et al., 2002). The *spot* enhancer in *D. biarmipes* was found to contain mutations which resulted in the creation of a binding site for both a transcriptional activator (Distalless) and a transcriptional repressor (Engrailed) (Arnoult et al., 2013; Gompel et al., 2005). The repressor, Engrailed, was found to be responsible for restricting *yellow* expression at the posterior area of the wing (Gompel et al., 2005).

1.1.6 Models of enhancer organisation

Investigations of enhancer organisation, including mapping TF binding in-vivo and TF motif identification, have revealed that enhancers are composed of clusters of TF binding sites (Shlyueva, Stampfel, & Stark, 2014). One reason why the binding of multiple TFs is thought to be important for enhancer activity is due to the high affinity of enhancer sequences for histone octamers (Tillo et al., 2010), which creates a barrier for the TFs in accessing the DNA. Cooperativity between TFs is thought to facilitate nucleosome eviction at enhancers and consequent TF occupancy (Calo & Wysocka, 2013).

Several distinct models have been proposed to explain how combinatorial binding of multiple TFs enables TF occupancy and enhancer activation (Figure 1.4). The “enhanceosome” model postulates that TFs cooperate directly with each other via physical interactions, resulting in the formation of a highly ordered nucleoprotein complex (Thanos & Maniatis, 1995). The mammalian IFN-beta *cis*-regulatory element is the best example of an enhanceosome. Eight TFs bind cooperatively to the interferon beta enhancer; disruption of the binding of individual TFs disables the enhancer, suggesting that composite nucleoprotein structure is critical for enhancer activity (Thanos & Maniatis, 1995) . As well as facilitating the binding of the TFs to the DNA, the nucleoprotein complex

may also provide a surface through which other co-factors, which play a role in stimulating transcription at the target promoter, can be recruited. As such the activity of the enhancer will be greater than the sum of the individual TF contributions (Spitz & Furlong, 2012). The enhanceosome model requires a highly specific arrangement of TF binding sites so that the TFs can physically interact with each other when bound. However, most developmental enhancers do not

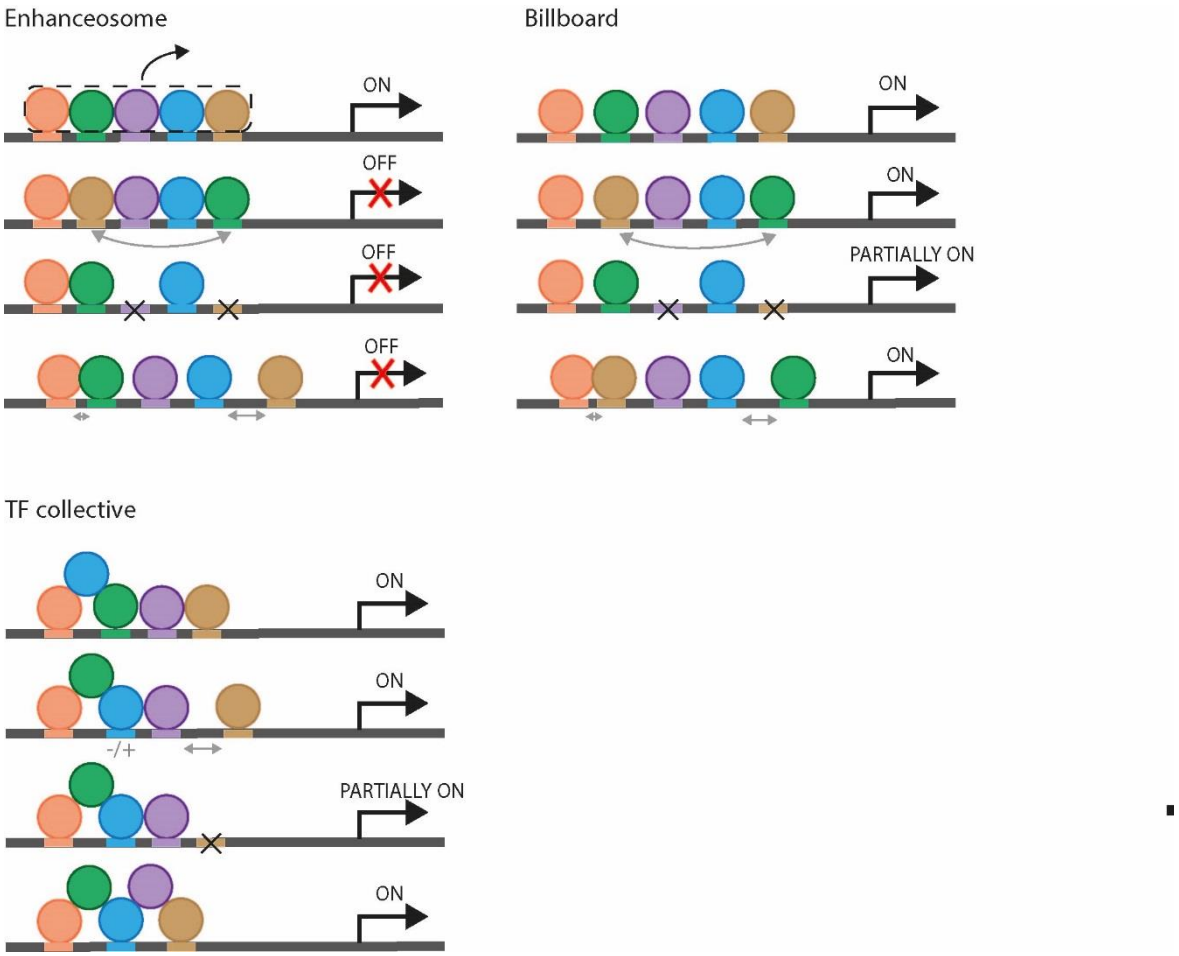


Figure 1.4. Models of enhancer organisation.

In the enhanceosome model TFs cooperate to form a nucleoprotein complex that integrates information from all the TF binding sites to activate transcription. In this model the exact order, spacing, identity and number of binding sites are critical for its function; changing the order, spacing or removing any of the binding sites will result in a complete loss of ability of the enhancer to activate transcription. In the billboard model the TFs deliver their "doses" of activation independently to the target promoter. The spacing, identity and order of the TF binding sites is flexible. Consequently changing the order or spacing between TF binding sites will not result in any change in level of activation conferred by the enhancer, while the loss of TF binding sites will result only in a reduction in the level of activation that the enhancer confers. Finally, in the TF collective TFs form cooperative complexes at the enhancer though both TF-TF and DNA-TF interactions. As a result, the spacing, order and identity of the TF binding sites are very flexible; changes in the spacing and identity of the binding sites will not result in any changes in the level of activation conferred by the enhancer. Given that TFs can be recruited by other TFs in the absence of their binding site, loss of TF binding sites may result in no change in the level of activation conferred (if the TF can be recruited by another TF bound). Alternatively, if

the TF cannot be recruited by another TF already bound, then the loss of its binding site will result in a decrease in the level of activation conferred, comparable to that of the billboard model.

appear to have form such ordered nucleoprotein complexes. In addition, large-scale cross-species comparisons of enhancers and synthetic enhancer reporter studies largely support a more flexible organisational models enhancer organisation (Smith et al., 2013; Taher et al., 2011). Two more flexible models of enhancer organisation have been proposed: the "billboard" and the "TF collective". In the billboard model there are no direct interactions between TFs and each TF independently interacts with the target promoter (Kulkarni & Arnosti, 2003). As such, TF motif spacing and orientation is flexible. TFs may still bind cooperatively (indirectly) via collectively competing for DNA access with the same histone octamer (Miller & Widom, 2003). The "TF collective" proposes that TFs form cooperative complexes at the enhancer through both DNA-TF and TF-TF interactions (Junion et al., 2012). Unlike the "enhanceosome" model, where the TF-TF interactions are necessary to produce a highly ordered nucleoprotein complex, in the "collective" they enable TFs without motifs to be recruited to the enhancer. Motif organisation at the collective is thus the most flexible; both the arrangement and composition of motifs can be variable. It is likely that most enhancers fall within the spectrum of the model extremes, with some motifs showing flexible organisation and others with defined spacing and orientations (Long, Prescott, Wysocka, et al., 2016).

1.1.7 Chromatin signatures of enhancers

Active enhancers tend to show distinct chromatin signatures. The binding of TFs displaces nucleosomes, resulting in increased chromatin accessibility ("open" chromatin) and hypersensitivity of the DNA to enzymes such as DNase I or Tn5 transposase as measured by ATAC-Seq (Buenrostro et al., 2013; ENCODE Project Consortium, 2012). Furthermore nucleosomes in the vicinity of active enhancers tend to contain certain histone modifications, including H3K4me1 and H3K27ac (ENCODE Project Consortium, 2004; Heintzman et al., 2009). In the genomic era these enhancer features have enabled the systematic annotation of putative enhancers genome-wide (Maston et al., 2012). For example, chromatin immunoprecipitation (ChIP) of histone marks and TFs, coupled with next generation sequencing, enable the identification of regions and TF binding genome-wide respectively. As such enhancers can be predicted based on clusters of bound TFs and presence of H3K27ac and an enrichment of H3K4me1 over H3K4me3 (ENCODE Project Consortium, 2012).

1.1.8 Molecular mechanisms of enhancer activation

We have already seen in Section 1.1.3 the general mechanisms through which TFs are theorised to exert their effect on promoters (e.g. recruiting components of the PIC and facilitating its assembly, increasing accessibility of the promoter region, releasing RNA pol II from proximal pausing) – are any of these mechanisms particularly favoured by enhancers or are there any other mechanisms that might be unique to enhancers? There is some evidence to support a model whereby enhancers recruit chromatin modifying enzymes via TFs to open up the chromatin at the promoter regions (termed the "hit and run" model). Several studies have shown that tethering a chromatin modifying enzyme to enhancers can modulate target gene transcription, as predicted by the "hit and run" model (Hilton et al., 2015; Kearns et al., 2015; Mendenhall et al., 2013). However for two of the studies it was unclear whether this was a direct result of chromatin modification at the promoter by the long-range activity of the enhancer-tethered enzyme (Kearns et al., 2015; Mendenhall et al., 2013). As well as targeting activating marks to the promoter, enzymes at enhancers may also remove repressive marks (Vernimmen et al., 2011). For example the α-globin promoter, which in normal development loses its repressive Polycomb-associated chromatin marks prior to activation, shows increased binding levels of Polycomb proteins and associated marks upon deletion of a key enhancer (Vernimmen et al., 2011).

The presence of RNA-poll II at enhancers led to the proposal of a model where RNA poll II is delivered to the core promoter via enhancers, either directly or through a tracking model where RNA poll II travels to the promoter via the intervening DNA (De Santa et al., 2010; Ptashne & Gann, 1997; Vieira et al., 2004). Finally the recent discovery that many enhancers produce bi-directional RNA, termed enhancer RNA (eRNA), has prompted speculation that these may also play a role in transcriptional activation (Kim et al., 2010). Support for the various mechanisms tends to come from single loci studies, and thus it is hard to tell which of the models will generalise. It is possible that different mechanisms are adopted by different enhancers and/or different TFs recruited to them (Beagrie & Pombo, 2016).

1.1.9 Activation over a distance

Given that one of the characterising features of enhancers is their ability to stimulate transcription over long distances, a key question is how they exert their long-range effects. Two major models have been proposed to explain how enhancers can act on promoters tens of kilobases away (Figure

1.5): scanning, whereby proteins recruited to the enhancer slide along the DNA until they reach a promoter sequence, or looping, which involves the formation of DNA loops between the enhancer and promoter, bringing the enhancer into close proximity with the promoter (Blackwood & Kadonaga, 1998; Bulger & Groudine, 1999). Over the past ~15 years a series of experiments have

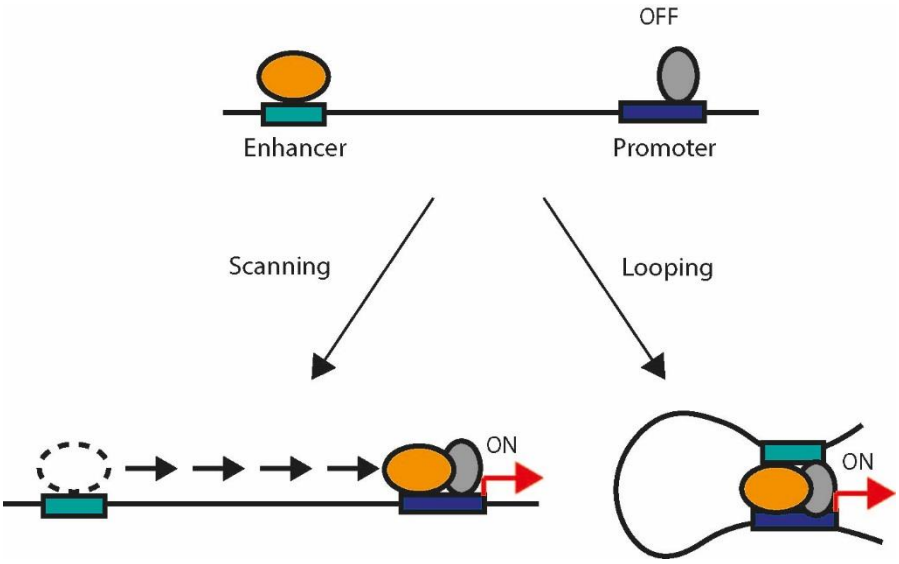


Figure 1.5. Mechanisms of transcriptional activation by enhancers.

In the scanning model RNA pol II and the associated transcriptional machinery track through the intervening DNA from the enhancer to the promoter. The looping model proposes that the enhancer comes into physical contact with the promoter, via the looping out of the intervening DNA.

provided support in favour of a looping model. The first direct evidence that an enhancer is in close physical proximity to the gene it regulates and thus supporting a looping model, came from two studies investigating the B-globin locus (Carter et al. , 2002; Tolhuis et al. , 2002). Carter et al. used RNA TRAP, a technique which enables chromatin in the immediate vicinity of actively transcribing genes to be tagged and recovered, to identify DNA sequences in close proximity to the actively transcribing B-globin gene. They found multiple regions in a known long-range B-globin enhancer located over 50kb away were in physical proximity to the B-globin gene (Carter et al., 2002). The second study, by Tolhuis et al. (2002) utilised the then-recently developed chromosome conformation capture (3C) technology, which is a biochemical approach that enables the detection of sequences in close physical proximity, to confirm that known enhancer regions were in close physical proximity to the B-globin gene (Tolhuis et al., 2002).

3C-based methods (discussed in the next section) have since been used to probe the distal interactions of thousands of promoters. These have revealed that promoters are frequently involved in multiple long-range interactions (Mifsud, et al., 2015; Rao et al., 2014; Schoenfelder et

al., 2015; Y. Zhang et al., 2013). The distal interacting regions are enriched for regulatory marks such as enhancer-associated histone marks and TF binding sites, indicating that many promoter interactions might be of regulatory nature (Mifsud et al., 2015; Rao et al., 2014; Schoenfelder et al., 2015; Y. Zhang et al., 2013). Taken together these findings support the idea that enhancers contact their target promoter via DNA looping. Additional support for a looping model comes from studies using imaging techniques, such as DNA-FISH, where loci of interest (e.g. a promoter and enhancer) are probed with DNA fragments labelled with different fluorophores on fixed cells (Giorgetti & Heard, 2016; Matharu & Ahituv, 2015). The distance between the two signals within the nucleus are measured, and compared to the cells where the loci are not predicted to interact. Numerous chromatin looping interactions have been observed between enhancers and their promoters using DNA-FISH, for example the *SHH* limb bud enhancer was shown to physically interact with its promoter in the developing limb bud (Amano et al., 2009). Direct evidence for the functional importance of DNA looping came from a study that used zinc finger proteins to induce looping between the B-globin gene and enhancer (Deng et al., 2012). Looping was found to induce strong transcriptional activation of the B-globin gene (Deng et al., 2012). Despite the strong evidence in support of a looping model, it is possible that not all enhancers act via looping. For example, a recent study found evidence of large scale PARP-mediated chromatin decompaction between neural enhancers and *Shh* (Benabdallah et al., 2017). The authors suggested this may be more compatible with a tracking model of activation (Benabdallah et al., 2017).

1.1.10 3C-based methods to identify promoter-enhancer interactions

In 3C-based methods, cells are cross-linked with formaldehyde to covalently link DNA segments that are in close spatial proximity (Job Dekker, Marti-Renom, & Mirny, 2013; Schmitt, Hu, & Ren, 2016) (Figure 1.6) . Next chromatin is fragmented by restriction enzyme digestion, followed by ligation such that cross-linked fragments are ligated together to form hybrid DNA molecules. The ligation frequency between chromatin fragments (and hence cross-linking probability) is taken as a proxy for their interaction frequency, and thus spatial proximity. In conventional 3C, single ligation products are detected by PCR using locus-specific primers (Dekker et al., 2002) (Figure 1.6). However, this is laborious and only enables interactions between two pre-chosen loci can be investigated. Several variants of 3C that differ in the way hybrid DNA molecules are detected have since been developed, overcoming some of these limitations. Notably one variant, Hi-C, enables the genome-wide detection of interactions between all loci (Lieberman-Aiden et al., 2009). Hi-C includes an additional step after fragmentation, where the DNA ends are filled in with biotinylated nucleotides before ligation (Figure 1.6). Biotinylated ligation junctions, each of which corresponds

to a ligation event between a pair of loci, can then be purified using streptavidin beads and sequenced using massively parallel sequencing. While Hi-C enables interaction between all loci to be detected, the resolution is limited by sequencing depth. For example, with several hundred million reads pairs (as is often routine), interactions can be detected at 100kb resolution in the human genome. Unless a far higher sequencing depth is achieved, which is usually prohibitive due to cost, Hi-C is not suitable for detecting specific regulatory interactions. Several 3C variants have

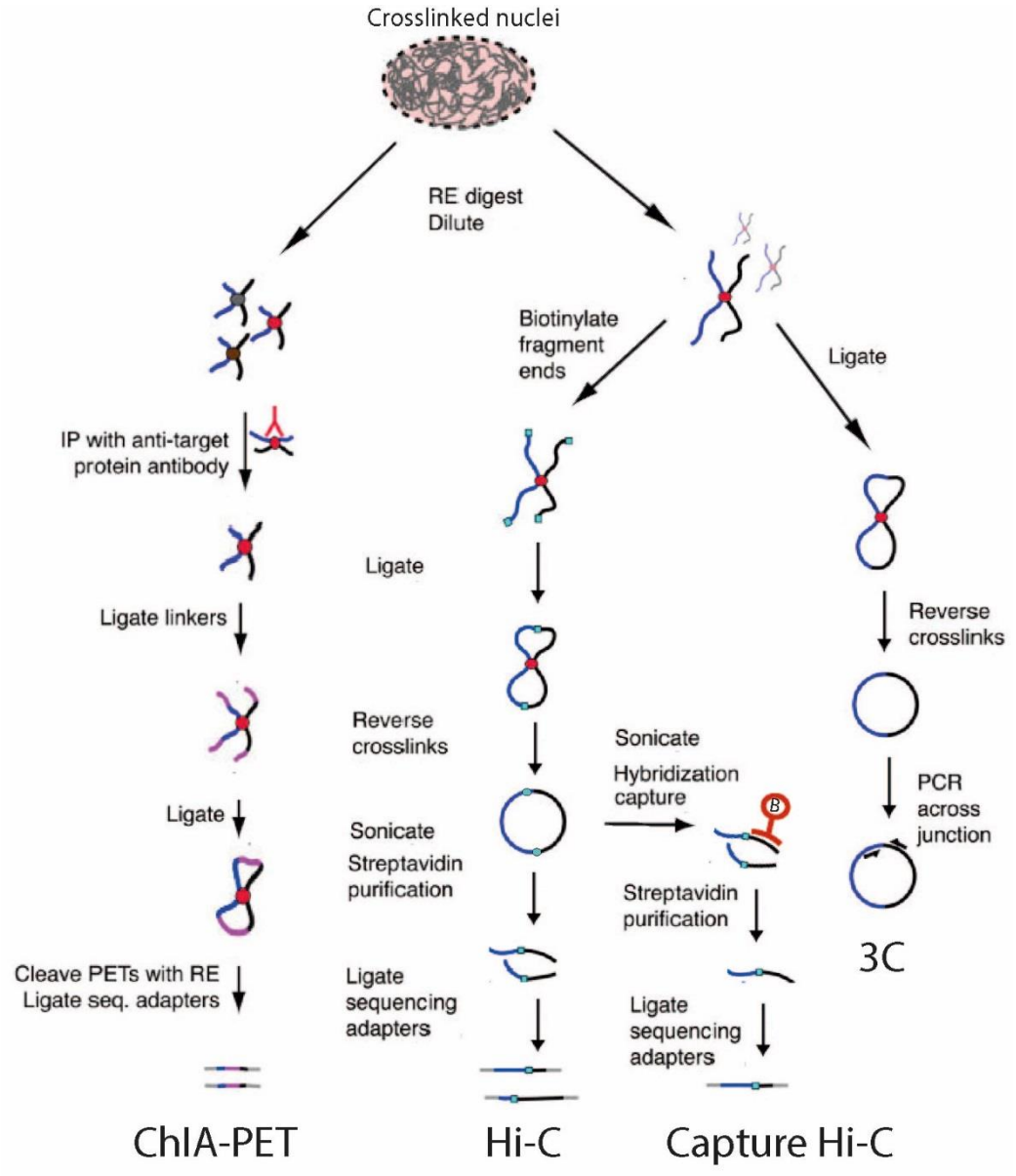


Figure 1.6. Chromosome Conformation Capture techniques.

Chromosome conformation capture is based on a restriction enzyme (RE) digest of cross-linked chromatin, followed by proximity ligation. This results in fragments being ligated together that were close together in 3D space, but potentially far apart on the linear genome. Several variants of chromosome conformation capture exist, differing in their strategies for fragmentation, enrichment and detection of the ligation junctions. In ChIA-PET, a ChIP step is included before the proximity ligation, to enrich for DNA interactions involving a protein of interest. In Hi-C an additional step is included after fragmentation, where fragment

ends are filled in with biotin. This facilitates the enrichment of successful ligation junctions in the final sequencing library. As a result Hi-C is a very high throughput chromosome conformation capture method that enables all interactions between all genomic loci to be assayed. Capture Hi-C includes a capture step after ligation, using baits designed to capture sequence regions of interest (for example promoter regions). This enables a subset of interactions of interest to be detected at higher resolution. In 3C, single ligation products are detected using PCR with primers against two loci of interest. Figure adapted from Risca & Greenleaf (2015).

now been developed that enable a large number of interactions to be interrogated at an increased resolution. One such variant is Chromatin immunoprecipitation interaction assay with paired end tagging (ChIA-PET), as well as the recent HiChIP, which adds a chromatin immunoprecipitation step to enrich for interactions involving certain proteins (Fullwood et al., 2009; Mumbach et al., 2016). The use of an RNA-poll II or cohesin antibody, for example, enables a subset of promoter-enhancer interactions to be detected at high resolution (DeMare et al., 2013; G. Li et al., 2012) (Figure 1.6). A variant of Hi-C, Capture Hi-C (CHiC), uses sequence capture technologies to enrich for Hi-C interactions that involve a specific region of interest (Hughes et al., 2014; Mifsud et al., 2015; Schoenfelder et al., 2015) (Figure 1.6). One version of this, Promoter Capture Hi-C (PCHiC), uses sequence capture to pull down all fragments containing nearly all annotated promoters, enriching for all interactions involving promoters (Mifsud et al., 2015; Schoenfelder et al., 2015). Significantly, this enables the global detection of promoter interactions independent of the proteins bound.

3C-based techniques have been key in providing support for the role of chromatin looping in long-range enhancer regulation, as well as advancing our understanding of the role of 3D genome architecture in gene regulation (Dekker et al., 2013). Given the finding that many enhancers do not regulate the closest gene, linking enhancers to target genes is a non-trivial challenge (Daniel, Nagy, & Nagy, 2014). The fact that most enhancers appear to form chromatin loops with their target promoter, also enables 3C-based techniques to be used to link enhancers to target genes. The ability to assign enhancers to target genes based on physical proximity thus provides a breakthrough in enhancer assignment. Capture Hi-C, as well as deeply sequenced Hi-C, are particularly significant as they enable enhancers to be assigned to target genes globally (Mifsud et al., 2015; Rao et al., 2014; Schoenfelder et al., 2015).

1.2 Multi-enhancer logic

1.2.1 Concurrent regulation by multiple enhancers

It is becoming clear that many genes rely on the action of not one, but multiple enhancers, to generate the correct spatial-temporal expression patterns. In some cases each enhancer directs expression in a particular cell type, and as such the gene is under the control of a single enhancer in a given tissue/time (Spitz & Furlong, 2012). However there are many developmental genes that appear to be regulated by multiple enhancers driving partially or completely overlapping expression patterns, suggestive of concurrent regulation by enhancers (Barolo, 2012; Hong, Hendrix, & Levine, 2008). Enhancers driving similar patterns of expression have been termed "shadow enhancers" by Hong et al. (2008), after observing that many early patterning genes in *Drosophila melanogaster* have a secondary element with very similar TF occupancy and the same activity as a previously characterised enhancer (Figure 1.7). Shadow enhancers are not limited to Drosophila. For example the expression of both *Sox10* and *Shh* in mice was found to be driven by enhancers with highly similar spatial activities (Jeong et al., 2006; Werner et al. , 2007) (Figure 1.7).

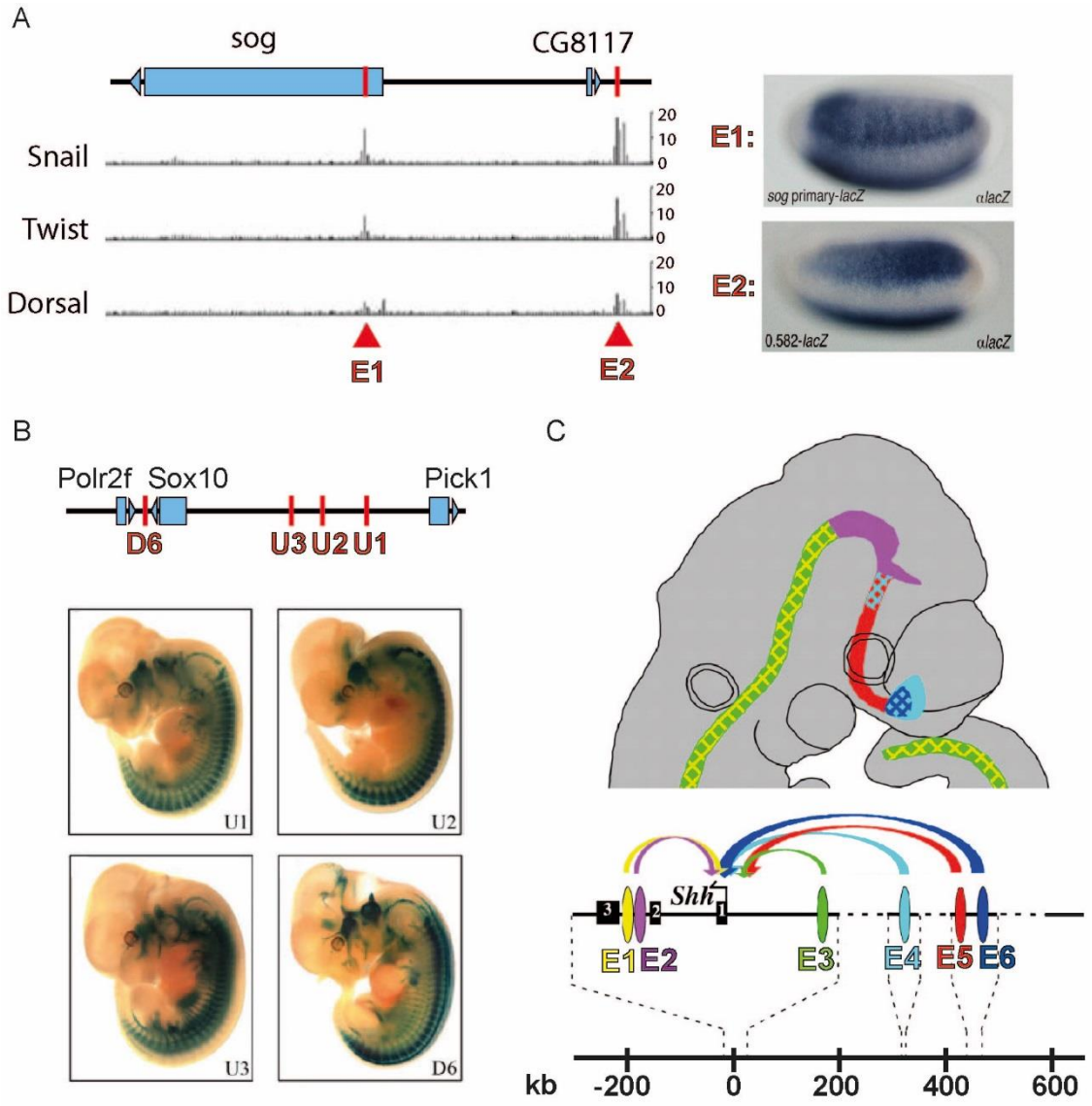


Figure 1.7. Shadow enhancers.

(A) Clusters of Snail, Dorsal and Twist binding were identified based on whole genome ChIP-chip assays by Hong et al., (2018). The leftmost cluster (E1) coincided with a previously identified enhancer belonging to the *Sog* gene. A second cluster (E2) was detected in the intron of the neighbouring gene, CG8117. A ~1kb DNA fragment encompassing the second cluster (E2) was able to direct lateral stripes of gene expression (pictures on the right-hand side), in the same way as the original enhancer, suggesting it functions as an authentic enhancer. They termed these enhancers, shadow enhancers. (B) Werner et al. (2007) identified four shadow enhancers (U1,U2,U3 and D6) that were able to direct very similar patterns of expression of β-galactosidase in transgenic mice to the neural crest where *Sox10* is active. (C) Jeong et al. (2006) identified six enhancers (E1 to E6) that were able to target reporter gene expression to sites of *Shh* transcription in the central nervous system of mouse embryos. In the schematic view of the mouse neural tube, regions are colour coded to indicate the distinct enhancer element driving their expression; solid colours indicate sites of *Shh* expression controlled by a single enhancer, whereas hatched patterns indicate that expression at that site is driven by more than one enhancer. As such, due to driving highly similar patterns of expression E1 + E3 and E4 +E6 could be classed as shadow enhancers. Figure are adapted from Hong et al. (2008), Werner et al. (2007) and Jeong et al. (2006).

1.2.2 Multiple enhancers may confer robustness

The apparent redundant activities of shadow enhancers have led to the hypothesis that they may confer phenotypic robustness. Recent studies on the Drosophila genes *shavenbaby* and *snail* have provided support for this hypothesis. *Shavenbaby* is regulated by five enhancers with extensively overlapping activities (Frankel et al., 2010). Removal of two of these enhancers does not impact trichome patterning at optimal temperatures, but at both low and high temperature extremes results in extensive trichome loss (Frankel et al., 2010). In addition, embryos heterozygous for the *wingless* gene that encodes a protein involved in trichome formation, only show trichome defects upon removal of the two shadow enhancers (Frankel et al., 2010). A similar picture emerged for *snail. Snail* is regulated by a proximal, and recently identified, distal enhancer located in an intron of a neighbouring gene (Ip et al., 1992). Quantitative imaging assays and genetic complementation experiments have shown a role for the *snail* enhancers in maintaining reliable expression at high temperatures, and upon a reduction in the concentration of a key activator (Dunipace, Ozdemir, & Stathopoulos, 2011; Perry et al., 2010). Removal of either one of the *snail* enhancers, in particular the distal enhancer, at either high temperatures or with a reduced concentration of a key activator, resulted in erratic patterns of gastrulation (Dunipace et al., 2011; Perry et al., 2010).

Whether shadow enhancers may play a similar role in buffering expression against environmental and genetic perturbations in mammals remains an open question. The finding that shadow enhancers of several mammalian genes are functionally redundantly under normal conditions, suggests it is plausible that they may play a role in conferring robustness. For example, deletion of

either of the two *TCRγ* enhancers had little effect on *TCRγ* transcription, but deletion of both enhancers caused a large reduction in transcription and defects in γδ thymocyte development in mice (Xiong, Kang, & Raulet, 2002). Similarly, one of the two enhancers driving expression *Pax3* in the neural crest, whilst sufficient to rescue neural crest cell development in mice lacking endogenous *Pax3,* is not necessary for development (Degenhardt et al., 2010).

How might shadow enhancers confer robustness to environmental and intrinsic fluctuations? A simple mechanism proposed by Perry and colleagues is that multiple enhancers decrease the overall failure of enhancer-mediated transcriptional activation (Perry, Boettiger, & Levine, 2011; Perry et al., 2010) (Figure 1.8). If enhancers act independently, then the combined probability that a cell will fail to express a gene in a given timeframe is the product of their individual failure rates. For example, if two enhancers each have a 10% failure rate, their combined failure rate is 1% i.e. transcription will fail in 1% of cells. This model was based on the observation that Drosophila embryos lacking either one of the two *hunchback* shadow enhancers had a greater number of inactive nuclei (Perry et al., 2011). This idea is consistent with the recent finding that at least some enhancers increase the frequency of transcriptional bursting, where several transcripts are produced in rapid succession followed by a period of little activity, rather than the size of bursts (Bartman et al., 2016; Fukaya, Lim, & Levine, 2016). If enhancers modulate burst frequency it might be expected that multiple enhancers would produce more consistent transcription, and thus cells would show less variation in transcript levels as observed with hunchback (Perry et al., 2011).

1.2.3 Additive action of enhancers boosts expression

Shadow enhancers may also act additively to ensure high levels of expression (Figure 1.8). The additive action of enhancers has been observed in both Drosophila and mammals. For example a quantitative live imaging study of Drosophila pre-cellular embryos found that two enhancers of a patterning gene *kni*, which drive near identical patterns of expression, showed additive activity (Bothma et al., 2015). Similarly, it was demonstrated through targeted deletions that the two enhancers driving neuronal expression of *Pomc* act additively in adult mice (Lam et al., 2015). By measuring mRNA levels for each enhancer deletion Lim et al. were able to quantify the contribution of each enhancer to *Pomc* expression, finding that the more distal enhancer is responsible for ~80% of *Pomc* expression and the more proximal enhancer ~20% (Lam et al., 2015). Two shadow enhancers at the α-globin locus also appear to act additively. Deletion of either of them resulted in a reduction in nascent α:β globin ratio in a manner consistent with the independent and additive

action of the enhancers (Hay et al., 2016). What might be the benefit of using multiple additive enhancers as opposed to a single enhancer of greater strength? It is possible that each enhancer has a maximum rate, at which it can activate the promoter, and so the only way to increase the expression level of a gene is to increase the number of enhancers used (Barolo, 2012).

Despite reductions in gene expression upon removal of either of the shadow enhancers at the *Pomc* or α globin gene, only very modest effects on metabolic phenotype and peripheral blood haemoglobin levels were observed (Hay et al., 2016; Lam et al., 2015). In the case of the α globin gene, this was thought to be due to stress erythropoiesis which compensated for the reduced α:β globin ratio (Hay et al., 2016). The lack of physiological effects suggests that the expression levels of these genes may be super-thresholded, perhaps to confer robustness. As such the additive, as well as the redundant action, of enhancers may confer robustness but via slightly differing mechanisms. Additive enhancers can increase robustness when expression is super-thresholded, such that the threshold can be reached with just one enhancer. Whereas redundant enhancers may ensure reliability of expression, through decreasing the overall failure rate of activation. It is of course possible that shadow enhancers may confer robustness through both mechanisms simultaneously.

1.2.4 Synergistic action of enhancers

In many cases, enhancers that appear to act redundantly due to overlapping spatio-temporal activities, are actually both necessary to generate the correct expression pattern (Barolo, 2012) (Figure 1.8). One example of such is the Drosophila gap gene *hunchback,* which is involved in establishing the segmented body plan of the embryo along the anterior-posterior axis. *Hunchback* transcription is normally localised to the anterior part of the drosophila embryo, in response to an attenuating *bicoid* gradient. Transgenes containing the proximal and distal element together were found to recapitulate this endogenous expression pattern (Perry et al., 2011). However, strikingly the transgene showed ectopic expression when driven by the proximal element alone; this was not found to be the case for the distal element, which drove expression within the endogenous area. A similar logic, whereby similar but slightly distinct regulatory inputs are integrated in a manner that is different than the sum of their parts, was observed for other genes involved in Drosophila embryo segmentation , including additional gap genes and pair-rule genes (Dunipace, Ozdemir, & Stathopoulos, 2011; Perry et al., 2011). Several examples of enhancers acting in a synergistic manner have been also observed in mice. For example, in mouse myeloid cells two enhancers were found to act synergistically to maintain high levels of PU.1 expression (Leddin et al., 2011). One of

the enhancers was found to bind myeloid cell specific C/EBP-α, which was able to increase chromatin accessibility and permit PU.1 binding at the second enhancer (Leddin et al, 2011). Cross-talk between enhancers is one potential mechanism that may enable synergistic action. In the case of *hunchback,* repressors bound at the distal element may act in a dominant negative way to supress the proximal element (Perry et al., 2011). Indeed the distal element was found to contain binding sites for repressors downstream of the *torso* signalling pathway, which likely mediates the

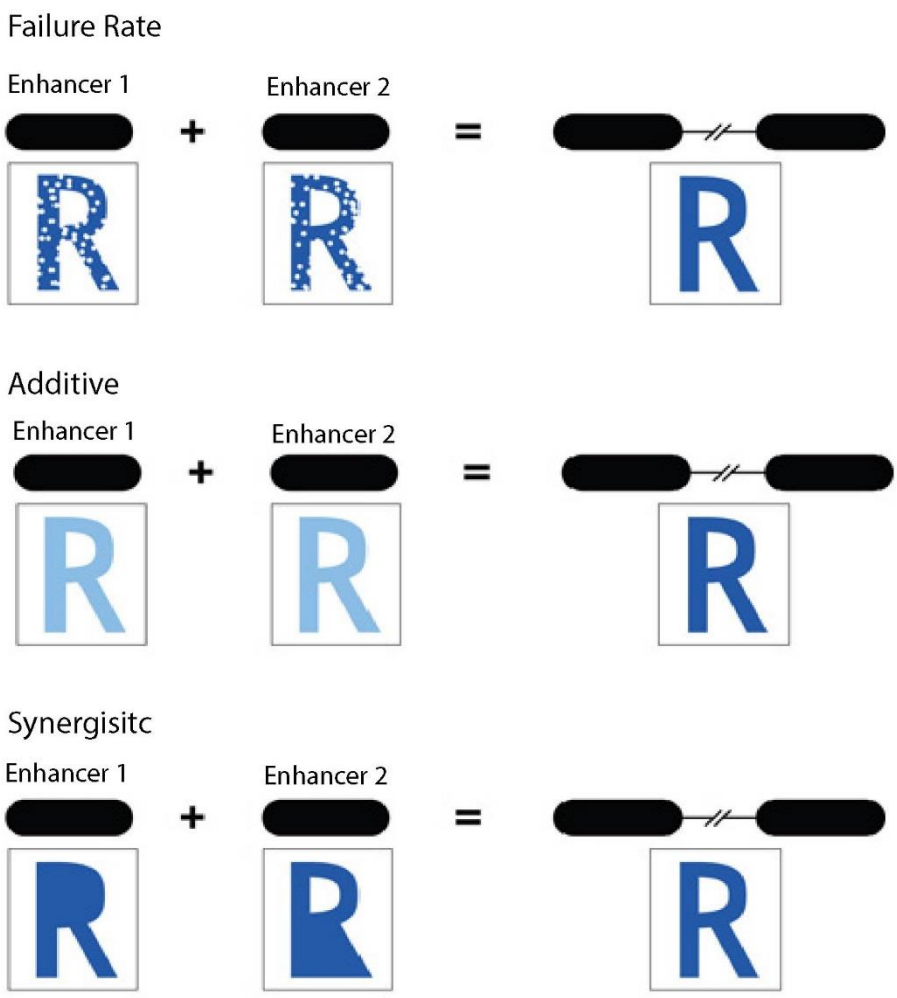


Figure 1.8. Possible mechanisms for the function of shadow enhancers.

"Failure rate": if enhancers have an inherent failure rate (they fail to activate transcription in a set proportion of cells), the inclusion of an additional enhancer will reduce their combined failure rate and increase the proportion of cells where gene expression is activated. "Additive": the total activation output is a sum of the output of the two individual enhancers. If enhancers have an inherent "maximum" activation level, increasing the number of enhancers provides a way to increase the rate of transcription of a given gene. "Synergistic": enhancers with overlapping patterns combine to produce a novel expression pattern that is different to the sum of their parts.

repressive effect (Perry et al., 2011). Interestingly the *Pomc* neuron-specific shadow enhancers, that acted redundantly in adult mice, were found to act synergistically during the embryonic stage The removal of either enhancer at the embryonic stage drastically reduced expression of *Pomc* (Lam et al., 2015). A similar situation was observed for the TCRγ locus, while enhancers were found to act redundantly in γδ thymocytes, while in a different cellular context they were found to act non-redundantly (Xiong, Kang, & Raulet, 2002b). The synergistic action of enhancers at some time point in development may be a key in enabling redundant enhancers to be maintained over evolutionary time (Cannavò et al., 2016).

1.2.5 Regulation by multiple enhancers is widespread

To date most of the insights into multi-enhancer logic have come from single-locus studies of well-studied developmental genes, and it is thus unclear how widespread concurrent regulation by multiple enhancers is. Cannavò et al. recently attempted to gain a more global insight into the prevalence of "simultaneous" enhancer regulation (Cannavò et al., 2016). They used TF occupancy data from a range of Drosophila mesodermal tissues to predict pairs of enhancers active in the same tissue and associated with the same gene (shadow enhancers). They used two methods to do this; the first of which involved identifying correlated regions of TF binding within 50kb of each other and a gene active in the same tissue. For the second method, they utilised a machine learning approach, trained on enhancers with characterised activity, to predict tissue activity from TF occupancy. Shadow enhancers were defined as pairs of enhancers having either highly correlated TF occupancy or predicted activity in the same tissue, being within 50kb of each other, and within proximity of a gene expressed in the same tissue. They were able to identify ~1100 enhancers whose predicted activity overlapped with that of at least one other enhancer associated with the same gene (Cannavò et al., 2016). Out of the genes identified as having shadow enhancers, ~60% were associated with more than two enhancers, highlighting the potential complexity of multi-enhancer regulation (Cannavò et al., 2016) This suggests, at least for Drosophila developmental genes, that regulation by multiple enhancers in the same tissue and time point is relatively common.

Recent genome-wide enhancer mappings have provided evidence to suggest that concurrent regulation by multiple enhancers may also be widespread in vertebrates and not limited to developmental genes. First, the genome-wide annotation of enhancers in numerous mouse and human cell types has revealed that the number of active enhancers far exceeds the number of

expressed genes in a given cell type (ENCODE Project Consortium, 2004; Shen et al., 2012). Second, recent high-resolution chromosome conformation capture studies in mouse and human cells, which identify looping interactions of gene promoters, found that on average each promoter interacted with approximately four distal regions (Freire-Pritchett et al., 2017; Javierre et al., 2016; Ozsolak et al. 2007; Sanyal et al., 2012a). Collectively these findings suggest that regulation by multiple enhancers in a given cell type and time point might be a pervasive feature of metazoan gene regulation.

1.3 Natural variation and enhancer activity

1.3.1 GWAS and eQTLs - Population genomics approaches to assess the impact of natural variation

Given the importance of enhancers in the spatial-temporal control of gene expression, it might be expected that mutations, which alter enhancer activity carry phenotypic consequences. Indeed, several rare Mendelian disorders have been attributed to malfunctions of individual enhancers. Perhaps the most striking example is the dysregulation of *SHH* which has been found to cause limb malformations. *SHH* is normally expressed in the developing limb bud under the control of a long-range enhancer located >1 MB away from the gene, and is essential for limb patterning. Mutations at this long-range enhancer have been linked to a congenital abnormality in humans, known as pre-axial polydactyly, which results in the formation of extra digits. The mutations were found to alter the binding profiles of ETS factors at the long-range enhancer, causing ectopic expression of *SHH* in the limb bud (Lettice et al., 2002, 2003, 2012; Lettice, Hill, Devenney, & Hill, 2008).

Recent findings from genome-wide association studies (GWAS) have revealed that enhancer variants also play a role in many complex diseases and traits (Donnelly, Price, & Spencer, 2015.; Hirschhorn & Daly, 2005; McCarthy & Hirschhorn, 2008). GWAS involve the genotyping of hundreds of thousands or millions of single nucleotide polymorphisms (SNPs) in a large group of individuals, and testing each polymorphism for statistical association with a trait/disease of interest. Due to linkage disequilibrium (LD) this results in groups of correlated SNPs which show a significant association with the trait of interest. Over the past 10 years, hundreds of GWASs have been performed, identifying thousands of loci that may contribute to susceptibility of a diverse range of diseases (MacArthur et al. 2017). Strikingly the vast majority of these disease -associated SNPs (~93%) were found to fall outside of protein coding genes (Maurano et al., 2012a). Numerous studies have found that GWAS SNPs are enriched at regions of DNase hypersensitivity (DNase HS),

which is indicative of open chromatin, and at regions with enhancer-associated histone marks in the relevant cell type suggesting that many variants influence activity of non-coding regulatory elements (Ernst et al., 2011; Maurano et al., 2012; Schaub et al., 2012).

Expression quantitative trait loci (eQTL) studies, which identify genetic variants that influence gene expression, represent a population-based approach to evaluate the regulatory impact of non-coding variants (Albert & Kruglyak, 2015; Gilad, Rifkin, & Pritchard, 2008; Nica & Dermitzakis, 2013; Stranger & Raj, 2013). Standard eQTL studies involve testing for association between gene expression, as measured via microarrays or RNA sequencing, and genotypes of SNPs within a certain distance of the gene (often within 1MB) in tens to hundreds of individuals (Albert & Kruglyak, 2015a; Gilad, Rifkin, & Pritchard, 2008; Nica & Dermitzakis, 2013; Stranger & Raj, 2013).

A large number of eQTL studies have now been carried out in both human cell lines and primary tissues (e.g., Albert & Kruglyak, 2015; Battle et al., 2014; Blauwendraat et al., 2016; Grundberg et al., 2012; GTEx Consortium, 2015; Lappalainen et al., 2013; Montgomery et al., 2010; Pai, Pritchard, & Gilad, 2015; Yang et al., 2012), collectively reinforcing the notion that expression-modulating variation is widespread. The most recent studies which include ~1000 individuals found that the majority of genes contained at least one eQTL; given that power to detect significant variant-expression associations is strongly affected by the number of individuals included in the study, additional eQTLs are likely to be revealed as ever greater numbers of individuals are used (Battle et al., 2014). eQTLs were found to be highly enriched at DNase1 hypersensitivity sites and in chromatin states associated with active promoters and enhancers, suggesting that a large number of variants influencing expression through regulation. The majority of eQTLs lie close to the TSS, suggesting they fall at promoter regions (Stranger et al., 2007). Indeed as the distance between the eQTL and the TSS increases, the effect size tends to decrease (Westra & Franke, 2014).

As with GWAS studies, due to LD many correlated variants will show significant associations; the SNP with the most significant association (the “lead” SNP) is often considered as the most likely causal one. However, the “lead” eQTL SNP may not always be causal, for example, due to noise in the expression data, with estimates for the percentage of best eQTLs that are causal ranging from 30%-70% (Lappalainen et al., 2013). Several fine-mapping strategies have incorporated regulatory annotations to prioritise reduced sets of putative causal variants (Kumasaka, Knights, & Gaffney, 2015; Spain & Barrett, 2015; Wen, Luca, & Pique-Regi, 2015).

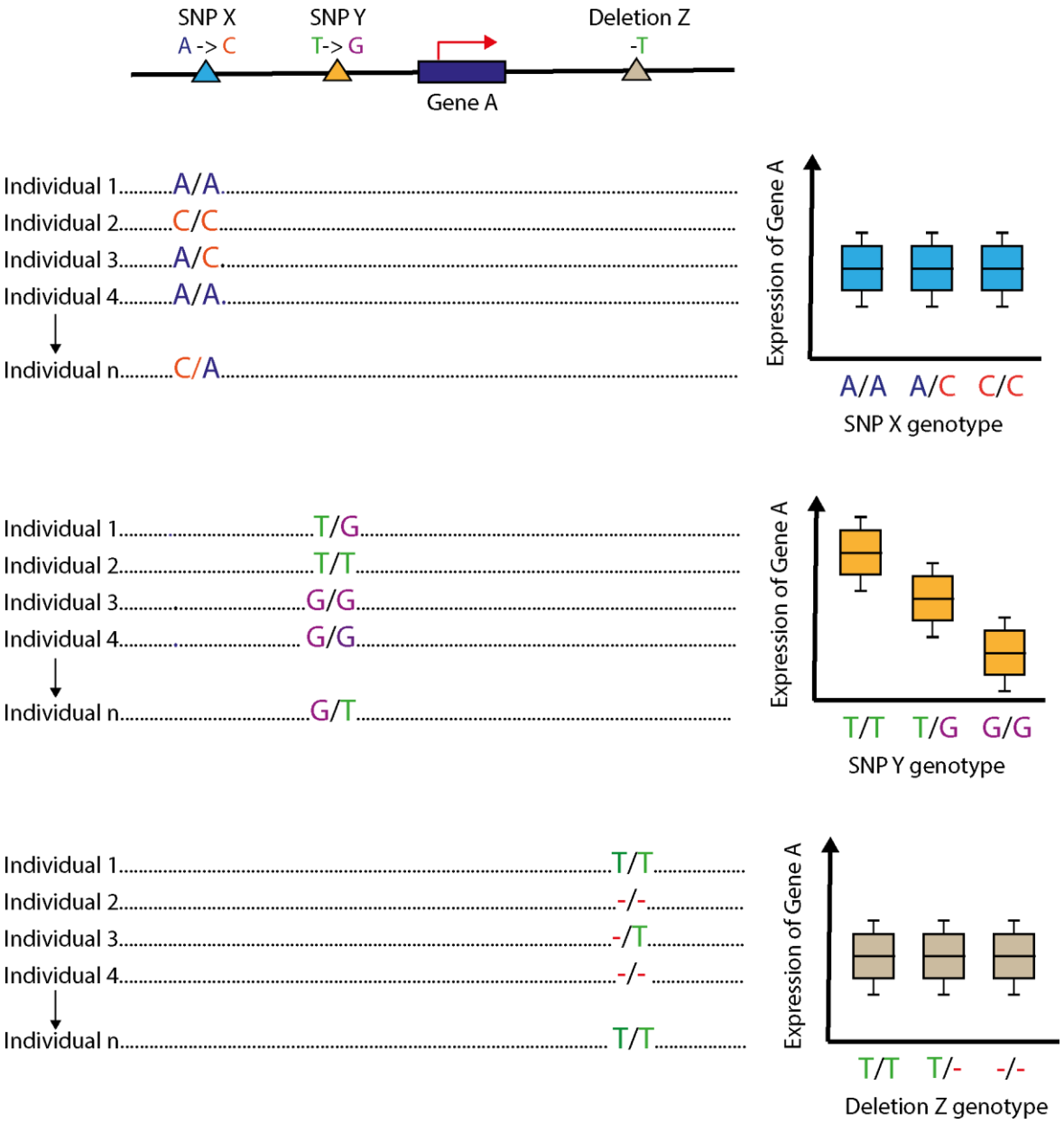


Figure 1.8. Schematic illustrating the principles of an eQTL analysis.

Two hypothetical SNPs (X and Y) and one indel (Z) are tested for association with expression of gene A. For each variant (X, Y and Z) an association test is carried out between the variant genotype and expression levels of gene A in hundreds of individuals. Here only for SNP Y does expression associate with genotype, and as such SNP Y can be termed an eQTL. For each variant the genotypes of a subset of the individuals tested are shown (for example in the first set of individuals, A/A indicates that the individual is homozygous for nucleotide A at SNP X). The boxplots on the right show the expression levels of the individuals, split by genotype at the respective SNP/indel.

1.3.2 How do non-coding variants influence regulatory activity?

1.3.2.1 Disrupting TF binding

Investigating the properties of eQTL variants provides the opportunity to obtain insights into the mechanisms through which non-coding variants act to modulate expression. Numerous studies have found eQTLs to be significantly enriched at TF peaks, consistent with the hypothesis that variants impact expression through TF binding (Gaffney et al., 2012). The availability of position weight matrices (PWMs), which model the sequence binding preferences of TFs, enable sequences to be scored according to their predicted affinity for a given TF (Berg & von Hippel, 1987, 1988; Fields et al., 1997; Stormo, 2000). Thus PWMs can be used to predict the effect on TF affinity of a particular variant (Andersen et al., 2008; Macintyre et al., 2010; Manke, Heinig, & Vingron, 2010; Moyerbrailean et al., 2016). Numerous studies have taken advantage of such computational binding predictions to test whether eQTLs are enriched for SNPs predicted to impact TF binding. For example one such study found that eQTLs in a lymphoblastoid cell line (LCL) showed a significant enrichment for variants predicted to alter TF binding; this enrichment was larger than that for variants which fell at a TF binding site but were not predicted to impact binding (Wen, Luca, & Pique-Regi, 2015). This is further supported by the finding that SNPs associated with allele-specific TF binding (as assayed by TF ChIP-seq) show a significant enrichment for eQTLs (Cavalli et al., 2016). It was also estimated that for around 55% of eQTLs the SNP genotype correlates with DNase hypersensitivity levels (dsQTL); a likely mechanism behind this association between SNP and DNase HS is through the alteration of TF binding, which affects local nucleosome occupancy and thus DNase cut rates (Degner et al., 2012). Indeed dsQTLs are significantly enriched in predicted TF binding sites, and the allele with the highest predicted TF affinity tends to be associated with the higher chromatin accessibility (Degner et al., 2012).

A similar picture has also emerged from recent massively parallel reporter assays (MPRA), which directly assess the ability of thousands of regulatory elements to drive expression of a reporter gene (Inoue & Ahituv, 2015; Kwasnieski et al. , 2012; Patwardhan et al., 2012; Tewhey et al., 2016). One such study tested the regulatory impact of SNPs identified as eQTLs across a panel of LCLs; out of the identified expression-modulating SNPs (which overlapped a TF peak), 76% showed a significant difference in predicted TF affinity (Tewhey et al., 2016). Another such study directly tested the effect of targeted motif disruptions for selected TFs on ~2,000 enhancers predicted based on chromatin data from the ENCODE in HepG2 and K562 cells (Kwasnieski et al., 2012). They found that the disruption of the binding sites of activator TFs resulted in reduced activity in the

relevant cell type. MPRAs are emerging as a powerful complementary tool for investigating the effects of non-coding variants on expression; unlike eQTL analyses, where determining the causal SNP is hard due to LD, they are able to directly assess the impact of individual SNPs. The disadvantage of MPRAs is that they do not study variants in their native context, and so for example the effects of chromatin and long range interactions will not be evaluated. As such they are perhaps most powerful when used in parallel with another approach that takes into account the native in vivo environment, for example they can be used to assist in identifying the causal SNPs at previously identified eQTL loci (Tewhey et al., 2016).

1.3.2.2 Altering the chromatin state

In addition to expression QTLs, several studies have investigated the effect of DNA variation on chromatin state. These analyses have revealed widespread associations between sequence variants and histone modifications (including H3K4me1, H3K4me3 and H3K27ac), DNA methylation, DNase HS and ATAC- Seq signals (Alasoo et al., 2017; Banovich et al., 2014; Chen et al., 2016; Degner et al., 2012; Maya Kasowski et al., 2013; Kilpinen et al., 2013; McVicker et al., 2013). QTLs for many of these features were significantly enriched for loci associated with changes in gene expression (Banovich et al., 2014; Degner et al., 2012; Kilpinen et al., 2013; McVicker et al., 2013). Local differences in histone modifications were found to correlate with TF binding site polymorphisms, suggesting that changes in TF binding might underlie at least some histone modification variation (McVicker et al., 2013). A similar picture was found for methylation and DNase QTLs, with SNPs predicted to change TF binding affinity significantly enriched for association with DNA methylation at close CpG sites and DNase HS (Banovich et al., 2014; Degner et al., 2012). A large number of loci were associated with both differences in histone modifications and DNA methylation, suggesting that chromatin changes are often coordinated, perhaps via TF binding changes (Banovich et al., 2014a). Collectively these results suggest that sequence variants could initially impact TF binding, which can lead to concomitant changes in chromatin state and gene expression. If this is the case, changes in gene expression might occur either directly as a result of TF binding perturbation or as a consequence of alteration of chromatin state.

1.3.2.3 Chromatin looping to impacted gene

While the majority of eQTLs lie proximal to the TSS, many lie at greater distances suggesting that they may fall at long range regulatory elements. 3C-based methods including 3C, 4C and Capture Hi-C have demonstrated that variants are often in contact with the impacted gene, indicative of DNA looping between the variant and gene (e.g. Canver et al., 2015; Cowper-Sal·lari et al., 2012; Javierre et al., 2016b; Smemo et al., 2014). For example a SNP linked with breast cancer that associates with expression of TOX9, was demonstrated to physically interact with the TOX9 gene (Cowper-Sal·lari et al., 2012). Further investigation revealed that the risk allele showed increased binding of FOXA1 region in-vivo, suggesting the SNP modulates enhancer activity via impacting FOXA1 binding (Cowper-Sal·lari et al., 2012). Another study showed that obesity-associated variants located in an intron of *FTO* physically connect to the promoter of another gene, *IRX3* (Smemo et al., 2014). Enhancers at the risk associated locus were able to recapitulate parts of IRX3 expression, suggesting that they may act as long range enhancers to IRX3. Consistent with this variants were found to associate with expression of IRX3, and not FTO in human brains (Smemo et al., 2014). Bauer et al. show that SNPs associated with increases in fetal haemoglobin level are localised to an intron of BCL11A, which is decorated with enhancer-associated histone marks and in close physical proximity to the promoter of BCL11A (Bauer et al., 2013). Using allele-specific analysis, the authors further demonstrated that the enhancer variants impact TF binding and expression (Bauer et al., 2013).

In theory, chromatin loops between variants and target genes may play a “passive” role in modulating expression, whereby they facilitate communication between the perturbed enhancer and the promoter. Alternatively, given that many factors known to be involved in loop formation are DNA binding proteins (e.g. CTCF), some variants may influence expression by impacting binding of these factors and directly disrupting looping (Rao et al., 2014). Experimental perturbations of CTCF binding motifs at the anchor points of loops, have been shown to alter looping (Sanborn et al., 2015). A recent study that identified binding QTLs for five TFs found evidence to suggest that they may alter looping (Tehranchi et al., 2016). Using allele-specific Hi-C, they were able to show that for binding QTLs that were heterozygous in the sample, the high binding allele made significantly more distal contacts (Tehranchi et al., 2016). While one of the five TFs was CTCF, a factor known to be involved in the looping, the other TFs were not known to play a role in loop formation. In these instances whether the TF is directly important in anchoring the loop, or if it contributes to the overall loss of activity of the regulatory element which as a consequences results in loss of the loop, remains to be seen.

In summary natural variation may impact enhancer activity through a variety of mechanisms including altering TF binding, chromatin state and looping. Many of these key insights have come from eQTL studies and GWASs, through the identification and characterisation of large numbers of variants associated with changes in expression and phenotypic traits. 3C-based studied have highlighted the role of variants at long range regulatory elements in modulating expression.

1.3.3 Limitations of eQTL approaches

While eQTL studies have enabled the identification of thousands of non-coding expression modulating variants and investigations of their properties, they have several limitations. These limitations may be especially relevant for the discovery of novel expression-modulating enhancer variants. Typically in eQTL studies, variants are classified as either *cis* or *trans*, based on their distance from the gene and potentially reflecting the mechanism through which they act (Nica & Dermitzakis, 2013). To identify *cis*-variants associated with expression, generally only variants falling within a small window (often around 1MB) around the TSS are tested. This is in order to limit the multiple testing burden and consequently increase the power. Trans-variants are tested separately, and suffer from much lower power due to the huge number of tests performed to search the genome for all *trans*-eQTLs. Even within the small window in which cis-eQTLs are tested, there are still a huge number of variants, requiring a vast number of tests to be performed. As a result, they suffer from low statistical power and thus are only able to detect robust changes in gene expression. Although there are numerous examples of enhancer perturbations that have large effects on gene expression, individual enhancers are generally expected to have more modest effects on expression than promoters. This is in part due to the redundant nature of enhancers (e.g. shadow enhancers). Massively parallel reporter assays, which directly assess the ability of thousands of regulatory elements to drive expression of a reporter gene, found that most enhancer variants that affect expression induced a modest 1.3 - 2 fold change in transcriptional level (Patwardhan et al., 2012; Tewhey et al., 2016). eQTL studies might therefore be inherently biased towards the detection of promoter variants which have a larger effect, and that a greater proportion of distal regulatory variants may be missed. This may partly explain why the majority of eQTLs are localised very close to the TSS.

Due to the vast number of variants tested per gene, eQTL studies tend not to consider combinatorial effects of multiple SNPs at a given loci and instead test the effect of each SNP individually. They also often make the assumption that there is a single causal SNP per gene or in

some cases per LD block. However there is increasing evidence to show that genes may harbour multiple eQTLs across their regulatory regions, and these may be in LD with each other (Bauer et al., 2013; Corradin et al., 2014; Wen et al., 2015). As the "causal" eQTL is selected as the one with the most significant association, if both proximal and distal eQTLs are detected, due to generally larger effect sizes the proximal ones are likely to be the most significant. Furthermore recent work by Corradin et al. (2014) suggests that several variants distributed across multiple enhancers may co-ordinately affect expression of their target gene, terming this the "multi enhancer variant" (MEV) hypothesis (Corradin et al., 2014). They provided evidence for this in six human autoimmune traits, where they found GWAS associations were often the result of expression changes brought about by the coordinated action of several SNPs in LD distributed over multiple enhancers. When such SNPs are tested individually (as in the case in most eQTL studies), unless they are in perfect LD, their effects on expression may be missed (Corradin et al., 2014). This may either be due to the fact that the change expression is brought about by the coordinated action of multiple SNPs, or each SNP has a very small effect and thus it is only the larger combined effect that is above the detection threshold (Corradin et al. 2014).

In conclusion, while eQTL studies are a powerful way to investigate the effects of sequence variants, they are not without limitations. These limitations are especially likely to impact the discovery of enhancer variants. As such, there is still room for the development of novel population genomics approaches to identify and investigate the function of enhancer variants.

1.4 Aim

My thesis sets out to examine two key issues presented here regarding gene regulation by enhancers. First, how a gene's regulatory information is encoded across multiple enhancers. Second, understanding and predicting the effects of enhancer variants.

2 Methods

2.1 Promoter Capture Hi-C (PCHiC) data processing and interaction calling

Promoter Capture Hi-C data for GM12878 were obtained from Mifsud et al. (2015). Interactions were called at a HindIII restriction fragment level using the CHiCAGO pipeline (Cairns et al., 2016). The CHiCAGO pipeline uses a convolution background model to account for both random distance dependent collisions and technical noise from the assay and sequencing. CHiCAGO corrects for multiple testing using a p-value weighting procedure based on the expected true positive rate for the given interaction distance. As such, the scores represent soft-thresholded –log weighted p-values. The default threshold of 5 was used, which has previously been shown to maximise enrichment of promoter-interacting regions for regulatory chromatin marks (Cairns et al., 2016). Baits were annotated for transcriptional start sites (TSSs) using the bioMart package in R with Ensembl TSSs for GRCh37 (Smedley et al., 2015). Baits with TSSs for more than one gene (4,178 out of 22,076) were excluded from the analysis, as in these instances it is not being possible to tell which of the genes the detected interactions are formed with, which might confound analysis of gene-level properties.

2.2 Downloading and processing TF ChIP-Seq data

ChIP-seq data for 52 TFs in GM12878 were obtained from the ENCODE project (ENCODE Project Consortium, 2012). ChIP-seq narrow peak files for the 52 TFs (in GM12878) called with the SPP caller and thresholded the irreproducible discovery rate (IDR) were downloaded from the USCS ENCODE portal. Where multiple peak files existed for a single TF (due to either multiple ENCODE production groups performing ChIP-Seq for the same TF or different protocols being used) the intersect of the two files was taken for all TFs except for ERG1. One of the two ERG1 peak files (produced using a different protocol) had substantially fewer peaks than the other one and was identified by ENCODE as being of lower quality, likely missing many true positives. The union of the EGR1 peaks was taken instead to avoid losing peaks from the higher quality dataset. Histone and DNase peaks were also downloaded from the UCSC ENCODE portal (ENCODE Project Consortium, 2012a).

2.3 Definition of TF-bound regions/CRMs and integration with PCHiC data

For Chapter 3 analysis, the architectural factors (CTCF, RAD21 and SMC3) were excluded, resulting in a set of 49 TFs. For Chapter 3 analysis, all 52 TFs were used in annotating TF binding at CRMs– however due to the lack of availability of PWMs for 11 of the 52 TFs, binding affinities (and binding affinity variants) were only predicted for 41 of the 52 TFs.

The union of TF peaks for the 49/52 TFs was taken (minimum 1bp overlap) to produce a composite set of TF bound regions for Chapter 3/Chapter 4 analyses respectively. This took the form of a binary matrix, where for each region the presence/absence of each TF was indicated. For the analysis in Chapter 2, this set of TF bound regions was filtered to those containing at least three different TFs and then referred to as cis-regulatory modules (CRMs).

The CRMs/TF bound regions were overlaid onto promoter interacting regions detected in Promoter Capture Hi-C, requiring a minimum 1bp overlap. For Chapter 4, TF-bound regions falling (by at least 1bp) within 18kb window centred at of the midpoint of the bait were defined as proximal TF bound regions. An 18kb window was chosen as this represented the median length of three restriction fragments, whereby the median was calculated only on restriction fragments included in this analysis (i.e. those involved in interactions, either as the bait or PIR, and overlapping with at least one TF bound region). The fragment immediately up- and downstream of the bait are excluded from Promoter Capture Hi-C during the data processing steps, resulting in theory, in a three restriction fragment length "blind" window.

2.4 Analysis of multi-enhancer genes

All analysis was carried out in R statistical environment unless stated otherwise.

2.4.1 Enhancer similarity metrics

Similarity of a gene's enhancers based on the identity of the TFs bound (the initial metric, used in Section 3.22) was calculated as follows:

(1) $$\overline{H} = \frac{1}{T} \sum_{j=1}^{T} \frac{x_j}{e}$$

Where *T* is the total number of unique TFs at a given gene, *x* is the number of enhancers that a TF is bound at a given gene, *e* is the total number of enhancers for a given gene and *j* is the TF.

The modified enhancer similarity metric, which was used to identify genes with highly dissimilar enhancers (Section 3.24), was computed as follows. The number of enhancers bound per gene by a given TF for genes with N enhancers is assumed to have hypergeometric distribution. The theoretical mean number of enhancers bound per gene and variance were calculated using the hypergeometric function for each of the 49 TFs for genes with 2-11 enhancers, as follows:

(2) $$\bar{x}_{ej} = \frac{eK_j}{E}$$

(3) $$var_{x_{ej}} = \frac{eK_j(E-K_j)(E-e)}{E^2(E-1)}$$

Where the *e* is the number of enhancers belonging to a gene, j is the TF, K is the total number of enhancers bound by a given TF across all genes, and E is the total number of enhancers across all genes. The resulting means and variances were used to compute z-scores for all TFs bound at each gene, as follows:

(4) $$z_{wj} = \frac{x_{wj} - \bar{x}_{ej}}{\sqrt{var_{x_{ej}}}}$$

Where *w* is the gene of interest, *j* is the TF, *x* is the number of enhancers bound by the given TF at the given gene, and $\bar{x}$ is the mean number of enhancers bound by the given TF across all genes with the equivalent number of enhancers. For example for a TF bound at two out of five of a genes enhancers, a z –score was computed using the theoretical mean and standard deviation of number of enhancers bound at genes with five enhancers for that particular TF. For each gene, the mean of the z-scores for all TFs bound at at least one enhancer was computed (termed modified enhancer similarity) as follows:

(5) $$p_w = \frac{1}{T}\sum_{j=1}^{T} z_{wj}$$

Where T is the total number of unique TFs bound across a genes enhancers and z is the z-score for a given TF (j) at the given gene (w).

2.4.2 Enhancer similarity permutations

Observed enhancer similarity was calculated for all genes with 2-10 enhancers using the enhancer similarity metric. To compute the expected enhancer similarity, the gene IDs were permuted so that enhancers were grouped by random instead of by gene that they were linked to, with the distribution of number of enhancers per random group matching that of the number of enhancers for actual genes. In addition, the number of TFs per enhancer within each random groups were matched to that of the number of TFs per enhancer linked to the same gene. For example if fifty out of the total genes had three enhancers bound by 5, 7 and 10 TFs, then the permuted set of enhancers would also contain 50 randomly grouped enhancers, where each group consists of an enhancer bound by 5, 7 and 10 TFs. 1000 permutations were carried out, and for each permutation the mean enhancer similarity of the randomly grouped enhancers was computed. The mean enhancer similarities of the permutations were used to calculate 95% confidence intervals.

To test that CRMs on the same PIR were not influencing the results, observed enhancer similarity was computed limiting each gene to one CRM per PIR. For each gene with multiple CRMs mapping to a single PIR, one CRM was randomly selected per PIR. Enhancer similarity was then computed for these genes (across the one-per-PIR CRMs), as well as for genes which did not have multiple CRMs mapping to a single PIR. 1000 single-CRM-per-PIR draws were carried out, and the mean across the 1000 draws taken. To compute the expected enhancer similarity for genes with one CRM per PIR, for all PIRs a single CRM was randomly selected, and then the gene IDs permuted as described before. This was done in such a way that the distribution of number of CRMs per random group and number of TFs at each enhancer per group, matched that of the single-CRM-per-PIR genes. 1000 permutations were carried out. For each permutation the mean enhancer similarity of the randomly grouped enhancers was computed, and used to calculate 95% confidence interval.

2.4.3 Enrichment analysis of “integrating” genes

A set of genes with highly dissimilar enhancers were identified and a background set of genes, matched in distribution of both the number of enhancers per gene and number of unique TFs per gene, were generated. To identify genes with highly dissimilar enhancers, enhancer similarity (the modified metric described above) was calculated for all genes with up to a maximum of 11 enhancers. Genes were grouped by number of enhancers (ten groups, 2-11 enhancers), and the bottom 10^{th} percentile of genes from each group were selected based on enhancer similarity.

Genes with >11 enhancers were excluded due to very low numbers of genes with >11 enhancers (<100 genes). For each group of low enhancer similarity scoring genes, a background set of genes were selected from the remaining genes (with similarity scores greater than the 10th percentile) with equivalent numbers of enhancers. The background sets of genes were chosen so that the distribution of number of unique TFs per gene matched the equivalent group of the low scoring similarity genes. This was achieved by splitting all genes, for each number of enhancers, into 10 groups based on 10th-quantiles of number of unique TFs across enhancers for a gene. The number of genes sampled from each quantile in the background set for a given number of enhancers was then matched to the number of genes belonging to each quantile in the low enhancer similarity scoring set of genes (for the same given number of enhancers). The groups of genes with highly dissimilar enhancers were then pooled (across genes with different TFs), as were the background groups of genes, to create one set of genes with highly dissimilar enhancers and one set of background genes.

The g:Profiler package in R was used to carry out an enrichment analysis (Reimand, Kull, Peterson, Hansen, & Vilo, 2007) for Gene Ontology (GO) terms, regulatory pathways and human disease gene annotations. The default method was used for multiple testing correction, as this was shown to cope better with the complex structures of GO terms (Reimand et al., 2007).

2.4.4 Identification of "lone" and "homotypic" binding TFs

For each TF a binomial logistic regression was used to model the probability that the TF is bound at an enhancer as a function of the gene the enhancer belongs to (gene dependent model), as follows:

(6) gene dependent model: $$\log\left[\frac{p_{Bound}}{1-p_{Bound}}\right] = \beta_0 + \beta_1 x$$

Where P_{Bound} is the probability of that an enhancer is bound by a given TF, and x is the id of the gene that the enhancer belongs to. In addition, for each TF a null binomial logistic regression was fitted, where the predicted value was fixed to the proportion of the total bound enhancers across all genes (i.e. gene independent). This is shown below, where P_{Bound} and x are the same as for equation (6):

(7) null model: $$\log\left[\frac{p_{Bound}}{1-p_{Bound}}\right] = \beta_0$$

For each TF the akaike information criterion (AIC) (Akaike, 1974) of the null model was subtracted from the AIC of the gene-dependent model, as follows:

(8) $$\Delta AIC = (-2 \cdot LL(gene\ dependent\ model) + 2k) - (-2 \cdot LL(null\ model) + 2k)$$

Where *LL* is the log likelihood of the gene dependent/null logistic regression model respectively, and *k* are the total number of genes in the analysis. The AIC was calculated using the general linear model function in R. The difference in AIC between the fitted and null model for each TF ($\Delta\mathrm{AIC}$) was plotted against the total number of enhancers bound by the respective TF. A smooth curve was fitted using locally weighted polynomial regression (LOESS), and 95% confidence intervals calculated. Any TFs with an AIC difference greater than the upper confidence limit (95%) were taken as significantly "lone" binding, while any TFs with a difference in AIC less than the lower confidence limit were taken as "homotypic" binding.

2.4.5 Comparative feature enrichment analysis of "lone" versus "homotypic" binding TFs

Comparative feature enrichment analysis was performed using the ToppCluster webtool (Kaimal, Bardes, Tabar, Jegga, & Aronow, 2010). An FDR of 5% was used. Due to the two lists of genes being very comparable in number ("lone" TFs = 13, "homotypic" TFs= 12), FDR-corrected enrichment p-values could be compared between the two sets of genes.

2.5 Prediction of TF binding variants

2.5.1 Downloading, filtering and integration of 1000 Genomes variant calls

Variant calls from the 1000 Genomes Project were used. These represent a set of phased single nucleotide polymorphisms (SNPs), short insertions and deletions (INDELS) and structural variants called from a combination low-coverage sequencing, high coverage exome sequencing and high-density micro-array genotyping of 2,059 LCLs. Variant calls for 359 LCLs of European ancestry (CEU, TSI, FIN, GBR, IBS) that overlapped with proximal and distal TF-bound regions for were downloaded from the 20130502 1000 Genomes Project release (Phase 3) using tabix and VCFTools (Auton et al.,

2015a; Danecek et al., 2011). Multi-allelic variants and variants with a minor allele frequency <0.05 within the 359 LCLs were filtered out.

2.5.2 Variant sequences of TF bound regions obtained

The GRCh37 genomic sequence for each TF bound region was accessed using the Bioconductor BSGenome package. The genotypes of all variants within the TF bound region for all 359 LCLs were examined, and unique haplotypes identified. Variant sequences for each TF bound regions were obtained by injecting the variants of each unique haplotype into the reference sequence for the respective TF bound region.

2.5.3 PWMs for human TFs

A collection of PWMs from the ENCODE TF ChIP-Seq experiments were used (Kheradpour & Kellis, 2014). The PWMs were loaded into R using the motif library in the atSNP package (Zuo, Shin, & Keleş, 2015).

2.5.4 Computation of normalised binding affinities

Binding affinities of variant sequences for PWMs were predicted using TRAP, a biophysical model which calculates the total affinity of a sequence for a given TF (Roider, Kanhere, Manke, & Vingron, 2007). This was chosen over the classical "hit"-based approach, which scans and returns a separate score for each segment of a sequence which are then thresholded to identify binding sites, as it naturally incorporates the effects of multiple variants and also takes into account multiple low affinity sites. Variant sequence affinities were computed using the tRap R package, with the pseudocount parameter changed to zero, to allow for using frequency as opposed to count matrices.

Sequence binding affinities were normalised, so that changes in binding affinities could be compared between different PWMs, using a method proposed by Manke et al. (2008). In brief, they proposed using a statistical score (p-value), where the probability of observing a given affinity or higher in the background sequence is computed, with the aim of normalising the observed affinity in light of a random sequence model. They demonstrated that the affinity distribution (A) can be

parameterised by the general extreme value (gev) distribution, with the probability of observing an affinity score for a given TF as follows:

(9) $$logA \sim P(x|a,b,c) = \exp(-\left[1 + a\frac{x-c}{b}\right]^{-\frac{1}{a}})$$

Where *a* is the shape parameter, *b* is the scale parameter and *c* is the location parameter. To avoid estimating the gev distribution parameters (a, b and c) for a given PWM for all observed lengths of TF-bound regions individually (as affinity is length dependent), the parameters for a range of different lengths (*L*) were estimated and a regression approach used to predict the parameters for TF bound regions of other lengths, as suggested by Manke et al. (2008). The regression model used to estimate the scale parameter is shown (the other two parameters are estimated in an identical way):

(10) $$c(L) = c_0 + c_1 LogL$$

This was done using the fit.gev function in the tRap R package. Parameters were estimated for sequence lengths 40, 100, 200, 250, 300, 400, 500, 800, 1000, 2000, 3000. These were chosen as they encompass the lengths of all TF-bound regions included in the analysis (min=42, median= 262, max=2875). For genomic background sequences, TF-bound regions not bound by the TF of the respective PWM were used, and extended to the required length, to ensure that DNA context of regulatory region was matched.

2.5.5 Changes in TF affinity relative to GM12878 computed

Change in binding affinity compared to GM12878 were computed for all PWMs for all unique haplotypes (excluding that of GM12878) for a given TF bound region. This was done by subtracting –log10(normalised affinity of GM12878 haplotype) from –log10(normalised affinity of alternative haplotype); resulting in negative values where the affinity of the variant haplotype was reduced compared to GM12878 for a given PWM. In cases where GM12878 was heterozygote for haplotypes, the highest affinity haplotypes was used. The reasoning being that this is likely the one underlying the TF binding observed in the ChIP-Seq data. There were a small number of cases where the normalised affinity for a PWM was zero. To avoid filtering out these cases, the lowest non-zero normalised affinity observed for that PWM was used instead.

For each TF bound region, the median change in binding affinity (as described above) was computed across all PWMs for a given TF for each unique haplotype. PWMs which had a normalised affinity score >0.1 in GM12878 were excluded (i.e changes in binding of these PWMs were not used). The rationale behind this being that not all PWMs likely represent the sequence responsible for the binding seen in GM12878, especially when a TF has multiple distinct PWMs. Different haplotypes may still show differences in affinity for PWMs which do not describe the given TF binding, which may result in a region being incorrectly identified as showing binding variation.

Any regulatory regions for which at least one haplotype showed a median change in affinity relative to GM12878 >0.6 for a given TF, were taken as showing binding variation for the respective TF. This threshold was chosen based on dsQTL enrichment analysis. Haplotypes for a given regulatory region showing a change in binding relative to GM12878 >0.3 for a given TF were pooled and termed the low affinity allele. The GM12878 haplotype (highest affinity haplotype where GM12878 is heterozygous) was termed the high affinity allele. For the rest of analysis individual genotypes were annotated based on these pooled haplotypes, with individuals either being homozygous for the high affinity allele for a given TF (i.e <0.3 change in affinity relative to GM12878), heterozygous for the high affinity allele or homozygous for the low affinity allele (i.e >0.3 change in affinity relative to GM12878).

2.6 dsQTL enrichment analysis

2.6.1 Downloading and processing of dsQTL data

The dsQTL data were from Degner et al. (2012), available in Gene Expression Omnibus under accession number GSE31388 (Edgar, Domrachev, & Lash, 2002). The set of dsQTLs identified using a 2kb cis-candidate window were downloaded, and converted to GRCh37 using liftOver (Hinrichs et al., 2006).

2.6.2 Testing for overlap with dsQTLs

The variant/s underlying the affinity change at each regulatory region which showed an affinity change relative to GM12878 over a range of thresholds for any TF, were identified. The thresholds used were 0.00, 0.05, 0.11, 0.16, 0.21, 0.26, 0.32, 0.37, 0.42, 0.47, 0.53, 0.58, 0.63, 0.69, 0.74, 0.79, 0.84, 0.89, 0.95, 1.00. It was not possible to use a threshold >1 due to having too few binding

variants (<100) showing changes over this threshold. These sets of binding variants were filtered to SNPs (INDELS were excluded) which were singly responsible for causing the change in the affinity (i.e SNPs which jointly impacted affinity were excluded). They were further filtered to SNPs with a MAF >0.05 in the 70 LCLs in the Degner et al (2012) analysis. For the control threshold of zero, for each regulatory region that showed no change in binding affinity for any TFs relative to GM12878, a single SNP was randomly selected (also with an MAF >0.05).

To calculate the proportion of variant SNPs overlapping dsQTLs for each threshold, a 400bp window centred around each variant SNP was tested for overlap with dsQTL SNPs. For each set of variant SNPs identified at each of the thresholds, the same of number of SNPs were randomly drawn and the proportion of random SNPs that overlapped dsQTLs were computed. This was repeated 10,000 for each different threshold level. The proportions for each random draw were used to calculate 95% confidence intervals for each threshold.

2.7 Testing for association between variants and gene expression

2.7.1 Downloading and processing RNA-Seq data

The RNA-Seq data used was from the GEUVAIDS sequencing project, which carried out RNA-sequencing on LCLs across seven laboratories (Lappalainen, Sammeth, Friedländer, 't Hoen, et al., 2013b). PEER-factor normalised RPKMs, filtered to genes in the top 50th percentile, were downloaded Gene Expression Omnibus under accession number E-GEUV-1 for 356 LCLs (Edgar et al., 2002). PEER-factor normalisation was used by GEUVADIS to remove technical variation from the RNA-sequencing data; the PEER algorithm uses a factor-analysis based approach to infer hidden factors (PEER factors) that explain much of the transcriptome-wide expression variability, which are then removed from the data by regression. The PEER-factor normalised RPKMs were transformed to a standard normal distribution, as linear regression, which was used to test for association between variants and expression, is sensitive to outliers, which RNA-Seq data is prone to, and assumes a normal distribution.

2.7.2 Identifying variant-expression associations

As described in Section 4.55, individuals were pooled according to their haplotype for each regulatory region for a given TF. Individuals with two alleles showing a binding affinity change <0.3 relative to GM12878 were classified as homozygote for the high affinity allele for the given TF and

individuals with two alleles showing >0.3 change in affinity relative to GM12878 were classified as homozygous for the low affinity allele. Individuals with one high affinity and one low affinity allele (as just described) were classified as heterozygote. Where the same variant was predicted to impact binding of multiple different TFs, these were collapsed into one multi-TF binding variant and only tested once. Linear regressions were performed to test for association between binding type (homozygous for the high affinity allele, heterozygous, and homozygous for the low affinity allele) and expression of the target gene (as assigned using Promoter Capture Hi-C). For genes linked with multiple TF binding variants, a multiple regression was used, where each predicted binding variant was used as a predictor. An ANOVA was used to test the overall significance of each regression model, and multiple testing was performed via FDR estimation (10%) on the gene-level p-values. For genes with multiple variants that were significant at the gene-level (at 10% FDR), Wald tests were used to get individual p-values for the variants; variants with $p<0.05$ were taken as significant.

2.8 Analysis of expression-associated variants

2.8.1 Comparison with GEUVADIS eQTLs

All significant eQTLs called by GEUVADIS (Lappalainen et al., 2013) on European LCLs were downloaded from the Gene Expression Omnibus under accession number E-GEUV-1 (Edgar et al., 2002). GEUVADIS eQTLs were compared to significant predicted binding variant-expression associations identified in my approach (but using a multiple testing correction of 5% instead of 10% FDR), for which the change in binding affinity of the predicted binding variation was caused by a single variant (i.e. predicted binding variation that was driven by multiple variants were excluded).

For testing the increased sensitivity of my approach, I used a dataset of all GEUAVDIS eQTLs (i.e. included non-significant associations) which I received from the authors via personal communication. Multiple testing was performed at both 5% and 10% FDR across all variants for all genes.

To compare the proportion of significant associations between the prioritised variants and randomly selected ones, variant-gene pairings were drawn at random (the same number as the set of prioritised variants) and the number significant at 5% FDR level was computed. 1000 random draws were carried out, and the number of significant associations per draw used to compute a 95% confidence interval.

2.8.2 Investigating properties of distal variant-expression associations

To identify variants that fall at restriction fragments containing another gene TSS, restriction fragments containing variants linked via Promoter Capture Hi-C to a distal gene were annotated with TSS interacting fragments were annotated for TSSs using the bioMart package in R with Ensembl TSSs for GRCh37 (Smedley et al., 2015).

In calculating the distances between the variant and TSS on the same fragment, the closest Ensembl TSS to the variant was used. The genome segmentation data for GM12878 (broad HMM bed file) used for the analysis were downloaded from the ENCODE USCS portal ENCODE Project Consortium, 2012a; Ernst et al., 2011; Ernst & Kellis, 2010).

For the expression analysis genes were classified as unexpressed if they were in the lower 50^{th} percentile according to normalised PEER residuals. All variants lying on the same fragment as a TSS were within 9kb of the TSS, and as such had already been tested for association with expression of the gene as a "proximal" variant. These FDR adjusted gene level p-values from the previous regression were used (Section 4.72) to determine if they were significantly associated with expression of the gene on the same fragment.

3 Regulation by multiple enhancers

3.1 Introduction

Many enhancers that concurrently regulate genes bind highly similar sets of TFs, as demonstrated for hundreds of Drosophila developmental genes (Cannavò et al., 2016; Hong et al., 2008). However, it remains to be seen whether this is a common feature of multi-enhancer logic. The finding that enhancers with diverse patterns of TF occupancy can also give rise to highly similar spatiotemporal expression patterns suggests that genes could use enhancers with differing TF occupancies (Liberman & Stathopoulos, 2009a; Zinzen, Girardot, Gagneur, Braun, & Furlong, 2009). Indeed several genes regulated by enhancers that recruit different sets of TFs have been identified. For example, expression of *krüppel* was found to be driven by two enhancers that, although show highly similar patterns of expression, are activated by different sets of TFs (Wunderlich et al., 2015a). There are also examples of genes that use enhancers with differing activities, such as *hunchback* that is controlled by two enhancers that show distinct but overlapping patterns of expression (Perry, Boettiger, & Levine, 2011). This was thought to be due to the binding of different TFs at each of the enhancers (Perry et al., 2011).

Here I have investigated TF binding across enhancers targeting the same gene globally, with respect to whether they bind similar or diverse sets of TFs. To do this I used a human lymphoblastoid cell line (LCL) as a model system. LCLs were chosen due to the availability of both a Promoter Capture Hi-C dataset (Mifsud et al., 2015) as well as ChIP- Seq data for 49 different TFs from ENCODE (ENCODE Project Consortium, 2012) for the same LCL. LCLs are developed by infecting human B-lymphocytes with Epstein Barr virus in-vitro, which immortalises resting B-cells into an actively proliferating B-cell population (Hussain & Mulherkar, 2012; Sie, Loong, & Tan, 2009). The LCL on which the Promoter Capture and the series of TF ChIP-seq experiments were performed was derived from a healthy individual. Significantly, through the integration of the Capture Hi-C dataset with the ENCODE TF binding maps, a set of multi-enhancer genes in this LCL can be identified, and TF binding annotated at their enhancers. Throughout this section enhancers binding similar sets of TFs will be referred to as “shadow” enhancers, while enhancers binding diverse sets of TFs will be referred to as “integrating” enhancers.

3.2 Results

3.2.1 TF binding profiled at multi-enhancer genes

I first used Promoter Capture Hi-C data to identify promoter interacting regions (PIRs) for ~18,000 promoters; the other ~4000 baited fragments were discarded as they contained promoters for multiple genes. Promoter Capture Hi-C data normalisation and signal detection using the CHiCAGO pipeline (Cairns et al., 2016) resulted in the identification of 63,753 significant cis-interactions, involving 11,770 baited promoters. I asked whether PIRs tended to be occupied by chromatin marks predictive of regulatory activity. To this end, I compared the percent overlap of all PIRs with a given histone mark/TF to that of a control set of distance matched interactions. Both enhancer-associated (H3K4me1 and H3K27ac), active (H3K4me3), as well as repressive histone marks (H3K27me3 and H3K9me3) were significantly enriched at PIRs. PIRs were also significantly enriched for the binding of CTCF, Rad21 and Smc3 (Figure 3.1); proteins thought to play a role in enhancer-promoter loop formation (Dixon et al., 2012; Rao et al., 2014a). The enrichment of both regulatory-associated histone marks and architectural proteins suggests that the interactions identified are likely to be of a regulatory nature.

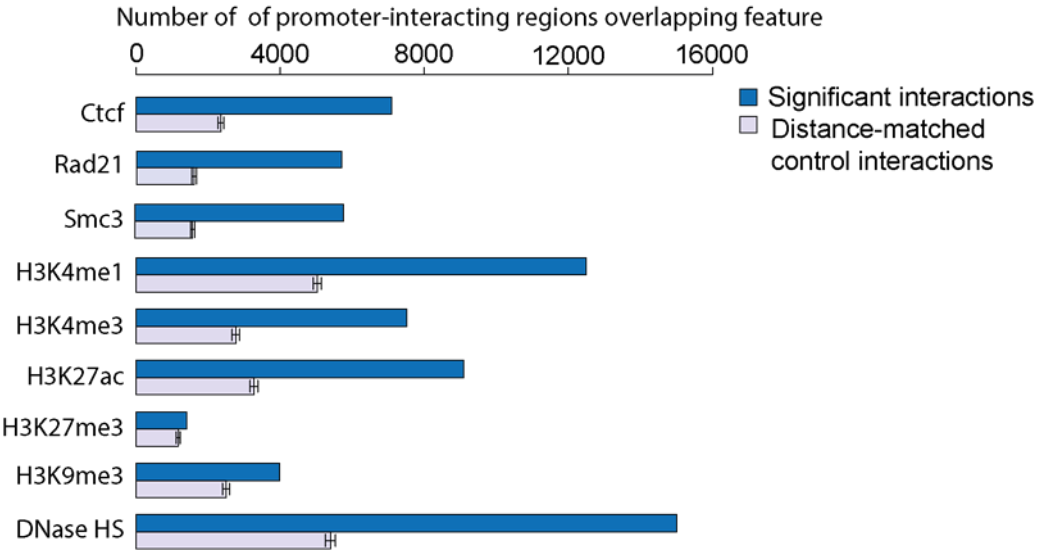


Figure 3.1 PIRs are significantly enriched for regulatory-associated histone marks and structural factors

Blue bars show the number of PIRs that overlap with regions containing the genomic feature (histone mark, structural factor or DNaseHS). Grey bars show the mean number of overlaps observed in distance-matched control regions over 100 permutations. Error bars show 95% confidence intervals across the permutations. PIRs were significantly enriched for all genomic features tested (permutation test p-value <0.01).

Next, I identified cis-regulatory modules (CRMs) at this set of PIRs. CRMs were defined by taking the union of all TF bound regions from ENCODE for 49 TFs; any composite region bound by at least three different TFs was taken as a CRM. This resulted in the identification of ~133,000 CRMs genome-wide. The CRMs were then overlapped with PIRs, and assigned as distal regulatory elements to the interacting gene. To illustrate this, two examples of genes for which I identified multiple distally interacting CRMs, *GADD45A* and *UBALD2*, are shown. For *GADD45A*, three CRMs were found to overlap with PIRs detected by Promoter Capture Hi-C, and were consequently assigned as distal regulatory elements (Figure 3.2a). The CRM lying at a PIR 486 kb downstream of *GADD45A* (in the gene body of *Il23R*) skips over two neighbouring genes to regulate *GADD45A*; I would have been unable to assign this to *GADD45A* using the traditional proximity-based method, demonstrating the benefits of using Promoter Capture Hi-C in assigning enhancers to target genes. For *UBALD2* I identified four CRMs overlapping with three PIRs (Figure 3.2b), illustrating how multiple CRMs can overlap with a PIR. In instances where multiple CRMs overlapped with a PIR, each CRM was taken as a discrete distal regulatory element. For both *GADD45A* and *UBALD2*, the distally interacting CRMs were defined as active enhancers in genome segmentation of GM12878 (Ernst et al., 2011; Kheradpour et al., 2013; Figure 3.2).

Out of the 11,770 genes for which PIRs were identified, 4,651 contained at least two CRMs at their respective PIRs. These genes were taken as "multi-enhancer" genes, and used in the following analysis. For these multi-enhancer genes the median number of distal CRMs per gene was four, with 50% of the genes having between two and seven distal CRMs (Figure 3.3a). The median number of PIRs per gene was three (Figure 3.3b). This is lower than the median number of PIRs due to multiple TF bound regions often mapping to a single PIR.

3.2.2 A metric to quantify the similarity of enhancer TF binding occupancies.

I first wanted to address whether in general, enhancers targeting the same gene in the same cell type tend to be bound by similar or diverse sets of TFs i.e. do genes tend to display a shadow or integrating enhancer architecture? In order to answer this question, I needed a method to quantify how similar a given gene's enhancers were in terms of TFs bound. To this end I devised a simple metric, the mean fraction of a gene's enhancers bound by each TF (present at the given gene), to quantify enhancer similarity in terms of TF occupancy (for formal definition see Section 2.4.1). To illustrate how the metric works I have quantified enhancer similarity of three hypothetical genes (X,Y and Z) that differ in terms of their TF binding across enhancers (Figure 3.4). Gene X has an

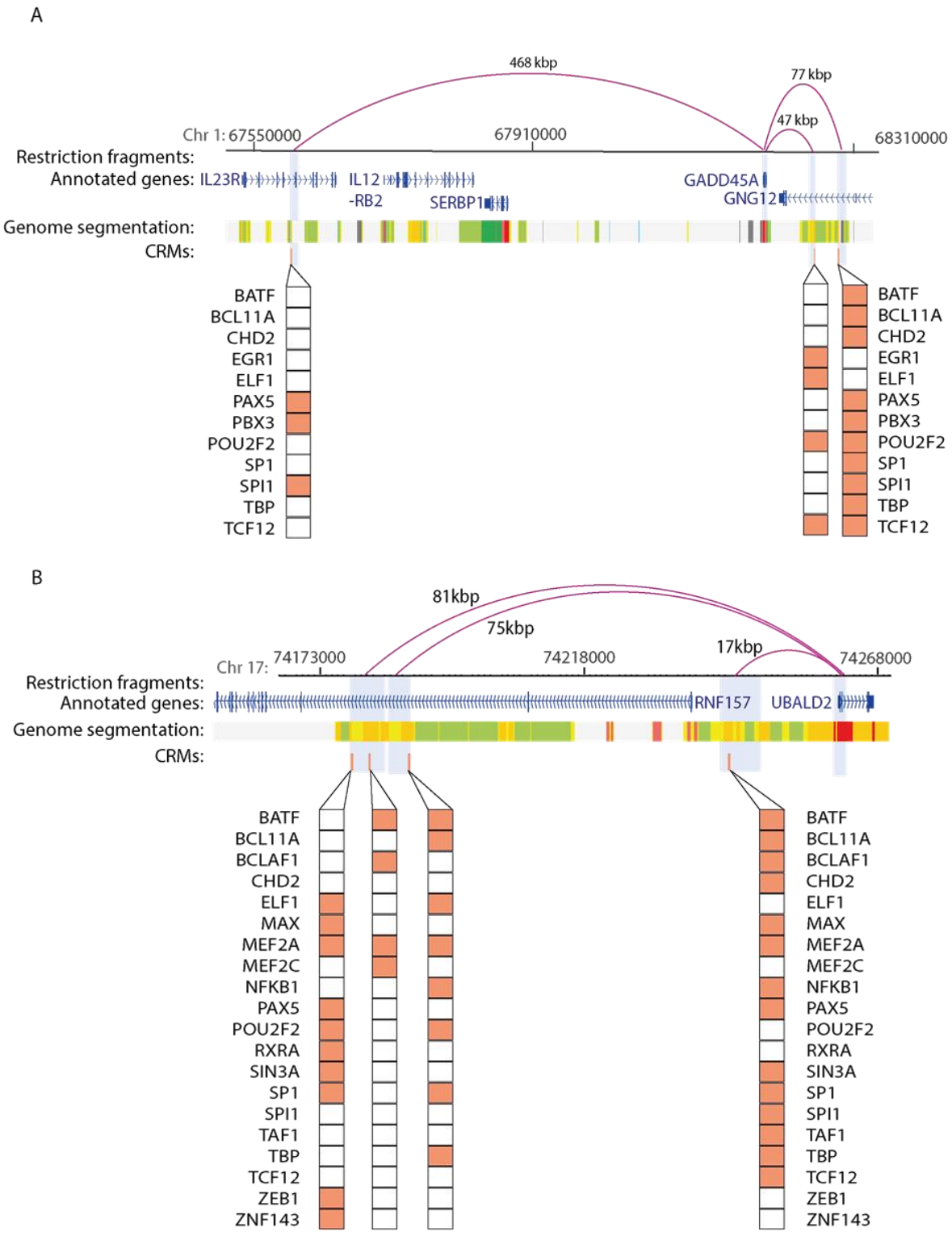


Figure 3.2 TF binding annotated at multi-enhancer genes

Genome browser representations of distal interactions for *GADD45A* and *UBALD2*S involving a fragment containing at least one cis-regulatory module (CRM). Significant interactions (as detected by Promoter Capture Hi-C) are shown as pink arches, with one end of the interactions containing the promoter and the other ends a promoter interacting region containing a CRM. CRMs were defined as regions bound by at least

3 different TFs according to ChIP- Seq data from ENCODE for 49 TFs (ENCODE Project Consortium, 2012a); the TFs bound at the respective regions are depicted beneath the browser representation (orange filled box indicates presence of TF). Genome segmentation tracks for GM12878 are shown (red, active promoter; orange, strong enhancer; yellow, weak enhancer; dark green/green, transcribed/weakly transcribed; grey, heterochromatin (Ernst et al., 2011; Kheradpour et al., 2013). (A) For *GADD45A* three CRMs were found to overlap with PIRs detected by Promoter Capture Hi-C. Each CRM was consequently assigned as a distal regulatory element, with *GADD45A* classified as having three enhancers in following analyses. (B) For *UBALD2*S four CRMs were found to overlap with three PIRs, with two CRMs mapping to the most distal PIR. Each of the four CRMs was taken as a discrete regulatory element, as such *UBALD2*S was classified as having four enhancers in following analyses.

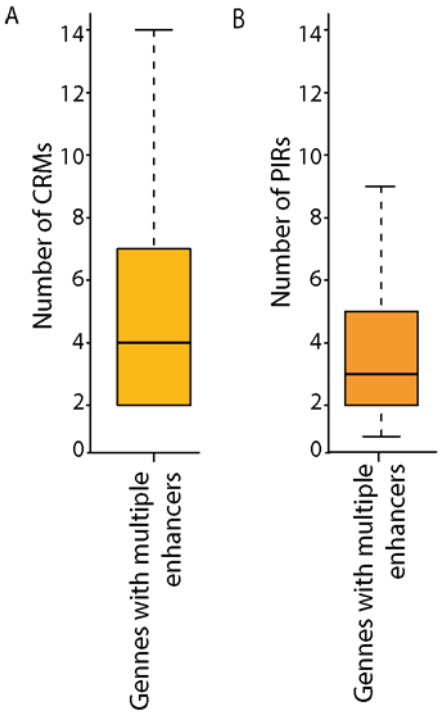


Figure 3.3 Multi-enhancer genes

Multi-enhancers genes were defined as those with at least two CRMs overlapping a PIR/s. (A) Boxplot shows the distribution of the number of CRMs assigned as distal regulatory elements for multi-enhancer gene. As multi-enhancer genes were defined as having at least two CRMs, the minimum number of observed CRMs is two. (B) Boxplot shows the distribution of number of PIRs for per multi-enhancer gene.

extreme shadow enhancer architecture, with exactly the same TFs bound at each of its enhancers (Figure 3.4a). In contrast, Gene Y has a different set of TFs bound at each of its enhancers, and as such is an example of a gene with an integrating enhancer architecture (Figure 3.4b). Gene Z falls between the two extremes in TF enhancer organisation, with some TFs bound at multiple enhancers and some TFs bound at a subset of enhancers (Figure 3.4c). I quantified enhancer similarity for gene X, resulting in a score of 1; each TF is bound at all three enhancers, giving a mean

number of enhancers bound of 3/3 = 1 (Figure 3.4a). I further quantified enhancer similarity for genes Y and Z, resulting in scores of 1/3 and 1/2 respectively (Figure 3.4b and c). The metric effectively distinguishes between the three hypothetical genes, with a higher score indicating a more "shadow" type architecture (Gene X) and a lower score indicating a more "integrating" architecture (Gene Y). An advantage of this metric is that it is easily interpretable, for example a similarity score of 1/3 means that on average a TF is bound at one third of a gene's enhancers.

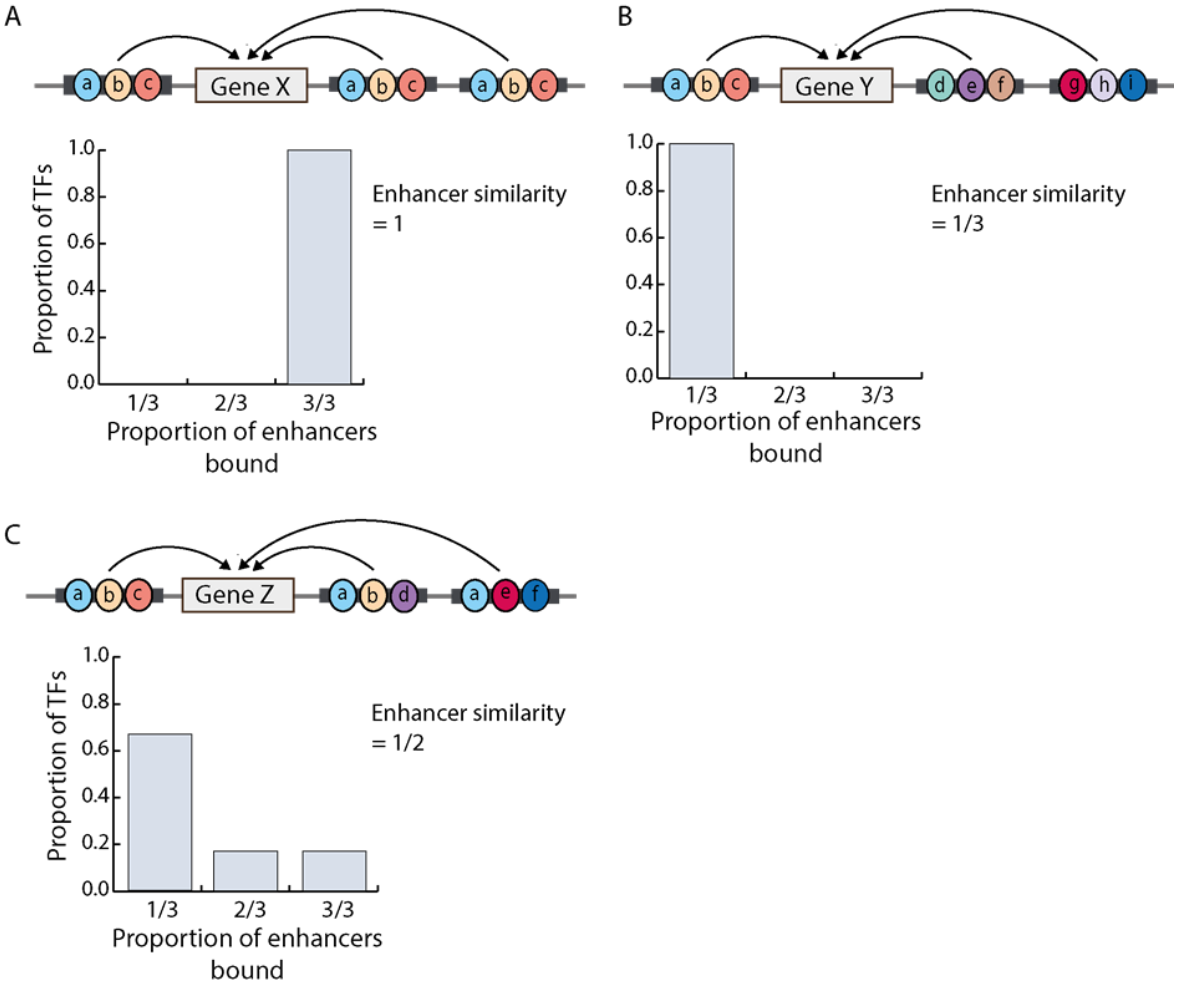
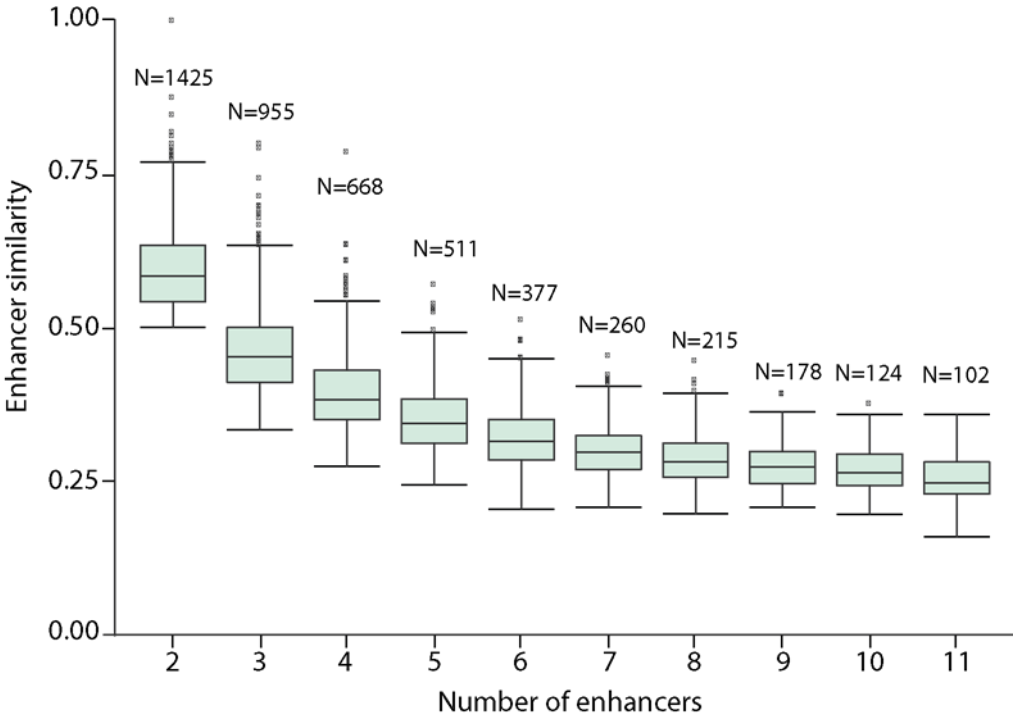


Figure 3.4 Quantifying similarity in TF occupancies across enhancers

Three hypothetical genes (X, Y and Z) are shown, each of which has three distal regulatory elements (black filled boxes) bound by three TFs (filled circles). For each gene the proportion of TFs bound across 1/3, 2/3 and 3/3 of the genes regulatory elements are displayed in a histogram. Enhancer similarity, calculated as the mean proportion of enhancers bound by a TF, is shown. (A) Gene X is bound by the same three TFs (A,B and C) at each of it's regulatory elements. As each TF is bound at 3/3 of the gene's enhancers, the mean proportion of enhancers bound by a TF is 3/3 = 1. (B) Gene Y is bound three different TFs at each of it's regulatory elements. Each of the nine TFs is bound across 1/3 of a genes enhancers, resulting in a mean proportion of enhancers bound by a TF of 1/3. (C) For Gene Z, TF A is bound across 3/3 of the gene's regulatory elements, TF B is bound across 2/3 of the gene's regulatory elements and the remaining four TFs (C, D, E and F) are bound across 1/3 of the gene's enhancers. This average fraction (across the six TFs) of enhancers bound by a TF is 1/2.

I next quantified enhancer similarity for the set of multi-enhancer genes that I identified in GM12878. Figure 3.5 shows the resulting distributions of enhancer similarities across genes, split by number of enhancers. It can be observed that genes with a greater number of enhancers tend to have lower similarity scores (Figure 3.5). This is in part a property of the metric, with the lowest bound equal to 1/N, where N is the number of enhancers per gene. To ensure the metric is able to distinguish between actual genes with differing enhancer similarities, I identified the genes with the lowest similarity scores (10th percentile) and highest similarity scores (90th percentile), and visualised their corresponding TF arrangements. Given that the metric depends on the number of enhancers (Figure 3.5), this was done separately for genes with different numbers of enhancers. Examples of the TF arrangements for genes with three, five and seven enhancers are shown in Figure 3.6. It can be seen that the genes with the most dissimilar enhancers (as defined by the enhancer similarity metric), have a much larger proportion of TFs bound at just one enhancer than genes with a high similarity score. Conversely, genes with a high similarity score have a much larger proportion of TFs bound across multiple regulatory elements. In conclusion the metric effectively enables the separation of genes based on similarity of enhancers, as well as the identification of groups of genes with both shadow (highly similar) and integrating (highly dissimilar) enhancers.



3.5 Distribution of enhancer similarities for genes with 2 -11 enhancers

Enhancer similarity, the mean proportion of enhancers bound by a TF, was computed for all genes with between two to eleven enhancers. Genes with >11 enhancers were excluded due to very low (<100) numbers

of genes in these categories. Boxplots show the distributions of enhancer similarities for genes split by number of enhancers. The number of genes in each category (i.e. with a given number of enhancers) are shown above the boxplot.

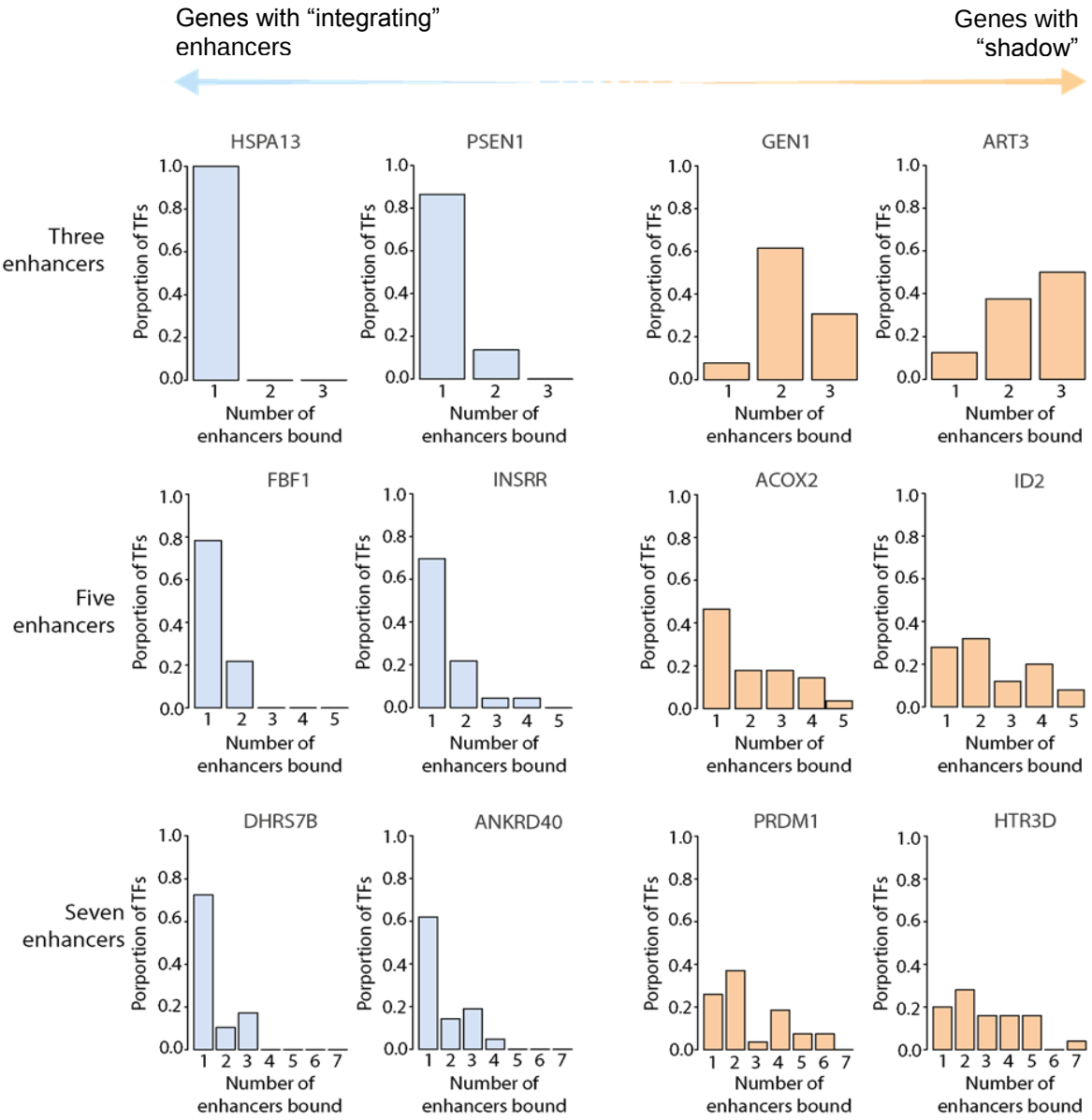


Figure 3.6 Enhancer similarity metric is able to identify genes with "shadow" and "integrating" enhancer architectures

Genes were split according to number of enhancers, and genes with the lowest similarity scores (10th percentile) and highest similarity scores (90th percentile) were identified. Two examples of the lowest scoring genes and highest scoring genes are shown for genes with three, five and seven enhancers. For each example gene the proportion of TFs bound at each number of a genes enhancers (1 to N, where N= total number of enhancer for a given gene) are shown in a histogram

3.2.3 Genes appear to favour a shadow enhancer architecture.

Having established an effective method to quantify enhancer similarity, I was able to ask whether in general enhancers targeting the same gene are bound by more or less similar TFs than would be expected by chance (ie. do genes tend to have a more “shadow” or “integrating” architecture than expected). To test this I grouped genes according to their number of enhancers and for each group permuted enhancers, resulting in enhancers grouped by random as opposed to by gene. This was repeated 1000 times and for each permutation the mean enhancer similarity of enhancers grouped by random was calculated. Given the relationship between number of TFs at each enhancer and enhancer similarity, the number of TFs across the randomly grouped (permuted) enhancers were matched to that of enhancers clustered by target genes. Enhancers linked to the same gene consistently show a small but significant increase in enhancer similarity compared to enhancers grouped by random (permutation test p-value <0.001 for genes with 2-10 enhancers; Figure 3.7a).

However for many genes multiple enhancer elements were assigned via the same promoter interacting regions (PIR), and therefore are close in linear distance. Although TF binding events, unlike histone marks, are not thought to spread along the DNA, enhancers close in linear distance might show a correlation in TF binding. I reasoned that the significant increase observed in similarity for enhancers grouped by target gene might therefore be driven by genes which have enhancers in close linear proximity that show correlated TF binding. In order to rule out this possibility, I repeated the previous analysis, but this time limiting genes with multiple enhancers lying on the same PIR to one enhancer element per PIR. The mean similarities of the “filtered” enhancers linked by target gene were significantly higher than those grouped by random (permutation test p-value <0.001 for genes with 2-10 enhancers; Figure 3.7b). This suggests the previous observation, that enhancers belonging to the same gene tend to be bound by more similar TFs than expected, was not driven by enhancers in very close linear proximity on the same PIR. In conclusion, it appears that genes with multiple regulatory elements tend to favour a more “shadow” type enhancer architecture rather than an “integrating” architecture.

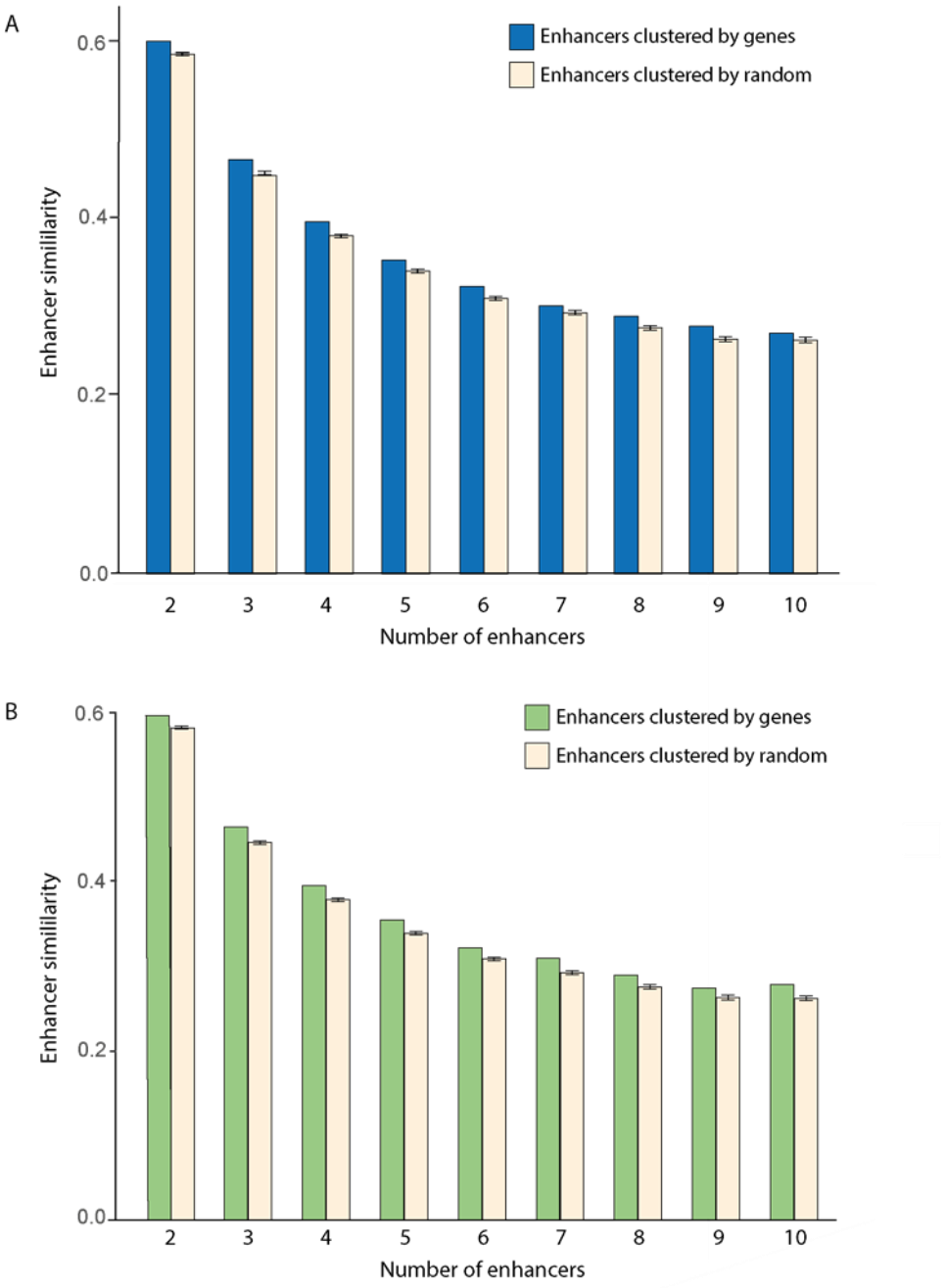


Figure 3.7 Genes tend to adopt a shadow enhancer architecture

Enhancers linked to the same gene are significantly more similar in terms of TF occupancy than enhancers randomly linked to genes, matched for numbers of TFs bound at each enhancer. (A) Blue bars show the mean similarities of enhancers linked to the same target gene (split by number of enhancers linked to each gene), and yellow bars show the mean similarities of enhancers randomly linked to genes over 1000 permutations (also split by number of enhancers). Error bars show 95% confidence intervals across permutations. Enhancers linked to the same gene are significantly more similar in terms of TFs bound than those linked to (permutation test p-value <0.001 for all numbers of enhancers). (B) Similar to (A), however here enhancers linked to the same gene were limited to one enhancer (CRM) per PIR for each gene. The green bar shows the mean similarity of enhancers over 1000 single-CRM-per-PIR draws across all PIRs and genes. Yellow bars show the mean similarities of enhancers randomly linked to genes over 1000 permutations, also limited to a single enhancer per PIR for each gene and matched for numbers of TFs at each enhancer. Error bars show 95%

confidence intervals. Enhancers linked to the same gene, limited to one enhancer per PIR, are significantly more similar than would be expected by chance (permutation test p-value <0.001 for all numbers of enhancers).

3.2.4 Genes with highly dissimilar enhancers have diverse biological functions

While in general genes tend to display a "shadow" enhancer architecture, clearly some genes adopt more of an "integrating" architecture (Figures 3.5 and 3.6). I hypothesized that genes displaying an "integrating" architecture might be involved in specific biological processes for which the binding of different sets of TFs across enhancers might be advantageous. In order to test whether genes targeted by highly dissimilar enhancers are involved in specific biological processes, I first needed to identify a set of genes with low enhancer similarities. While the enhancer similarity metric can be used to identify genes with highly dissimilar sets of enhancers, it is sensitive to the variable quality of ChIP experiments. In instances where a TF is detected at a low proportion of a gene's enhancers, it cannot be ruled out that the TF is bound more abundantly, yet not detected by ChIP. To ensure that the identification of genes with "integrating" enhancer architectures is not confounded by TF ChIP- Seq quality, I modified the enhancer similarity metric to take into account the number of ChIP- Seq peaks. To do this I made the assumption that the numbers of ChIP- Seq peaks would be generally similar for TFs binding at "shadow" versus "integrating" enhancer architectures. In brief, for a given gene the proportion of enhancers bound by each TF was converted to a z-score, based on the expected distribution of number of enhancers bound for a given number of enhancer ChIP- Seq peaks, and the mean of z-scores was taken across all TFs binding to all enhancers of this gene (see Section 2.4.1 for the equation). Figure 3.8a shows the proportion of enhancers bound by each TF and their corresponding z-scores for an example gene, *AURKAIP1*. As an example, both *ATF3* and *POU2F2* are bound at just one enhancer, and as such with the similarity metric would each contribute a score of 1/5 to the mean. However, the TFs differ in their expected distributions of proportions of enhancers bound due to having different numbers of ChIP- Seq peaks (Figure 3.8b). For *ATF3* the expected distribution is skewed to the left (mean=0.23) reflecting the fact that it has a relatively lower number of ChIP- Seq peaks, whereas for *POU2F2* the expected distribution is less skewed (mean= 0.47) due to having a greater number of ChIP- Seq peaks than *ATF3* (Figure 3.8b). As a result the z-score for *ATF3* is less negative than that of *POU2F2* (-0.4 and -1.2 respectively). This approach therefore normalises for the total number of TF ChIP- Seq peaks at enhancers.

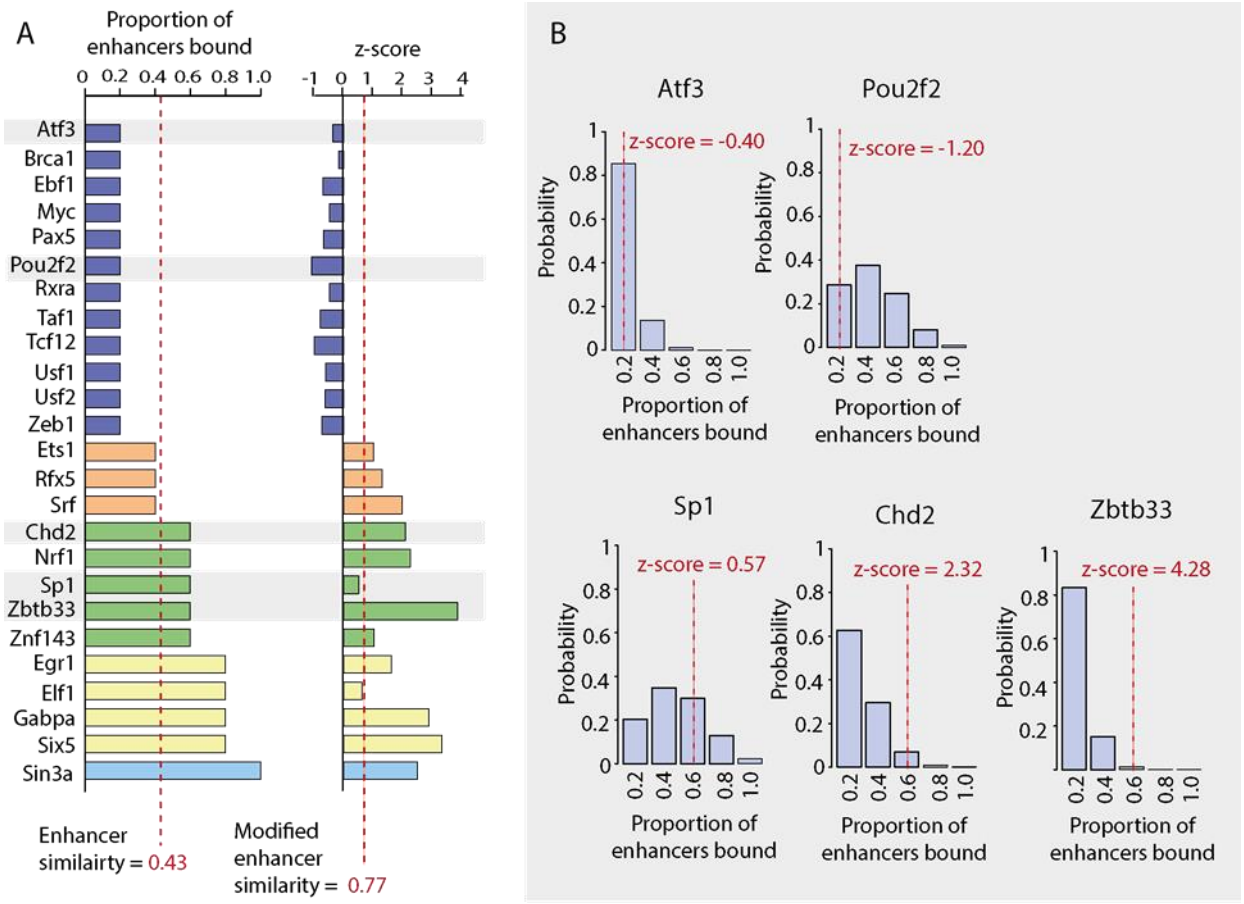


Figure 3.8 Normalising enhancer similarity score for number of TF ChIP- Seq peaks

(A) The left hand (LH) barplot shows the proportion of *AURKAIP1* enhancers bound by each TF (that is bound at at least one enhancer of *AURKAIP1*). The RH barplot shows the proportion of *AURKAIP1* enhancers bound by each TF, normalised for total number of ChIP- Seq peaks for the corresponding TF. This results in a mean normalised enhancer similarity of 0.77. (B) The expected distributions of proportions of enhancers bound (based on number of TF ChIP- Seq peaks) used in the normalisation are shown for five TFs. The expected distributions of proportion of enhancers bound (based on the number of TF ChIP-Seq peaks) are shown for ATF3 and POU2F2 (each bound at one enhancer of AURKAIP1), as well as SP1, CHD2 and ZBTB33 (all bound at three of AURKAIP1's enhancers). The more left-skewed the distributions are, the fewer the total number of ChIP-Seq peaks for the respective TF. The expected distributions are used to compute a z-score for proportion of enhancers bound at AURKAIP1 by each TF, normalising the proportion of enhancers bound for number of TF ChIP-Seq peaks.

I defined a set of genes with "integrating" enhancer architectures, by taking genes in the bottom 10th percentile according to the modified enhancer similarity metric. I first asked whether the number of unique TFs bound across enhancers differed between genes with "integrating" enhancer architectures and the remaining genes that fell between the 10th and 100th percentile for modified enhancer similarity (termed background). I observed that the background genes were bound by a greater number of unique TFs across their enhancers (median = 22) than the set of "integrating" genes (median =16) (Figure 3.9a). This difference in number of unique TFs bound across enhancers

between “integrating” and background genes may confound the analysis; genes bound by a reduced number of TFs may also be involved in distinct biological processes. To avoid number of unique TFs confounding the TF arrangement across enhancers, I thus decided to select a background set of genes that matched the unique TF distribution of the “integrating” set of genes (for details on how this was done see Section 2.4.3). This resulted in a set of background genes with a median modified enhancer similarity score of -0.02, compared to that of the integrating genes which have a median score of -0.51 (Figure 3.9b). An example of a highly “integrating” gene, *ZBTB80S*, and a background gene, *PGLYRP4*, are shown (Figure 3.10a and 3.10b). Both genes have three enhancers, bound by comparable numbers of unique TFs (15 for *ZBTB80S*, and 14 for *PGLYRP4*). However, whilst for *PGLYRP4* three of the 14 unique TFs are bound across multiple enhancers, for *ZBTB80S*, all 15 unique TFs are bound at only one enhancer i.e. each enhancer contains a different set of TFs (Figure 3.10b).

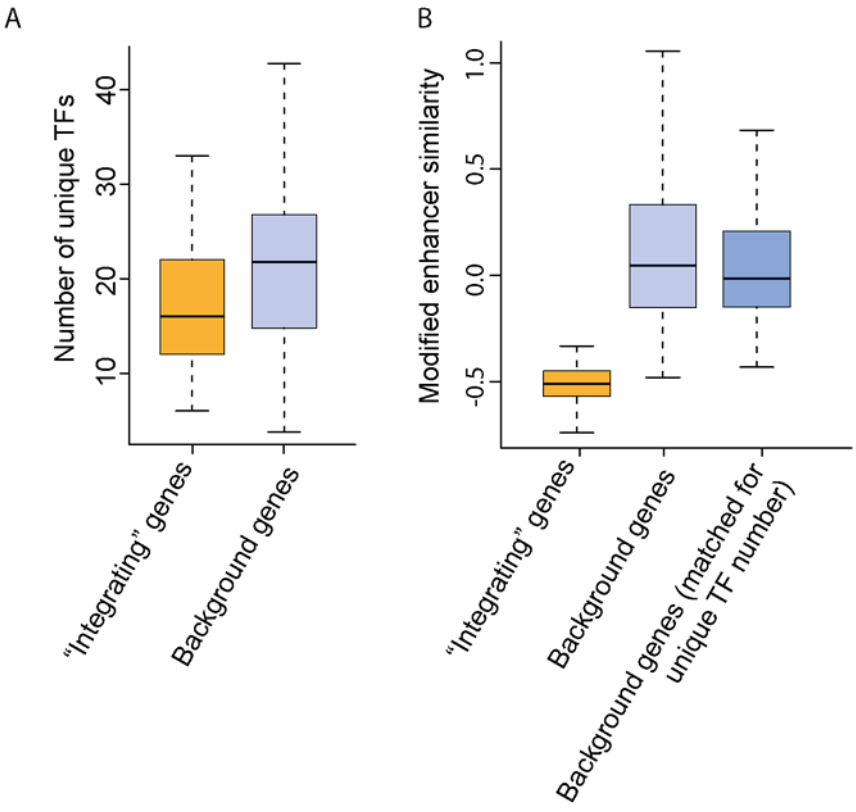
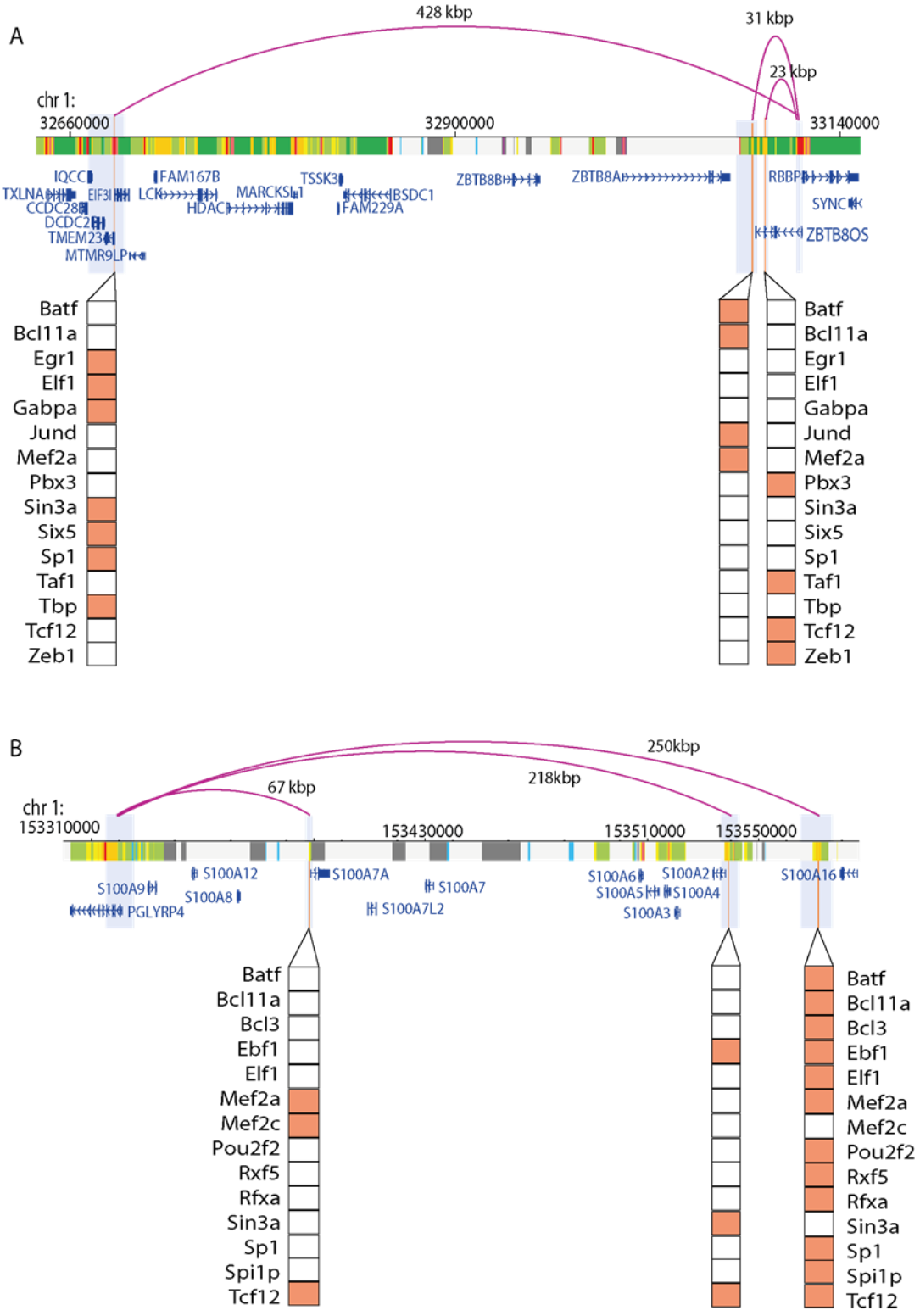


Figure 3.9 Number of unique TFs and modified enhancer similarity scores of “integrating” versus background genes

Integrating genes were taken as those in the bottom 10th percentile for modified enhancer similarity (this was done separately for genes with different numbers of enhancers, and the bottom 10th percentile from each category then combined). The remaining genes (between 10th and 100th percentile for modified enhancer similarity) were classified as background. (A) Boxplot shows the distribution of the number of unique TFs bound across all of a genes enhancers for “integrating” and background genes. (B) Boxplot shows

the distribution of modified enhancer similarity scores for integrating genes, background genes and a set of background genes matching the unique TF distribution of the “integrating” set of genes.



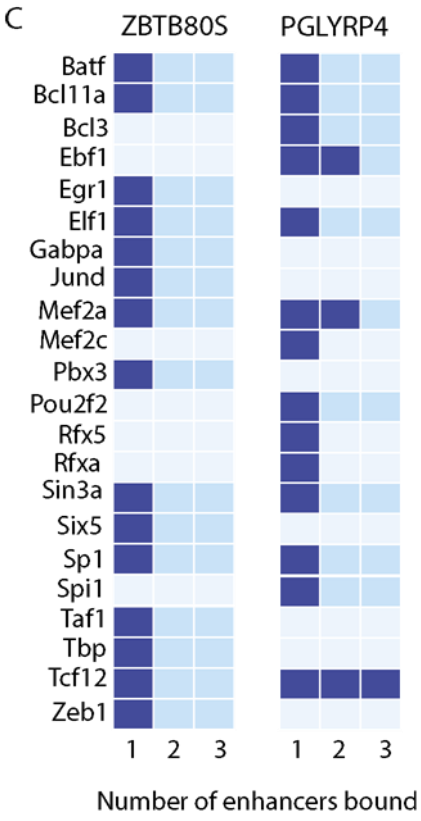


Figure 3.10 TF binding at enhancers of an integrating and background gene

Integrating genes were defined as those in the bottom 10th percentile according to the modified similarity metric, for genes split by number of enhancers. WashU browser representation of an example integrating gene (A), ZBTB80S, and background (B) gene, PGLYRP4. Significant interactions (as detected by Promoter Capture Hi-C) are shown as pink arches, with one end of the interactions containing the promoter and the other ends a promoter interacting region containing a CRM. CRMs were defined as regions bound by at least 3 different TFs according to ChIP-Seq data from ENCODE for 49 TFs (ENCODE Project Consortium, 2012a); the TFs bound at the respective regions are depicted below the browser representation (orange filled box indicates presence of TF). Genome segmentation tracks for GM12878 are shown (red, active promoter; orange, strong enhancer; yellow, weak enhancer; dark green/green, transcribed/weakly transcribed; grey, heterochromatin (Ernst et al., 2011; Kheradpour et al., 2013). (A) An example of a gene defined as having an "integrating" enhancer architecture. (B) An example of a gene in the background set. (C) Horizontal barplot showing the proportion of enhancers bound by each TF (as indicated by the dark blue squares). Only TFs bound at at least one enhancer of either of the two genes are included in the barplot. Where a TF is not present at any of the enhancers of the gene, the row corresponding to the TF is shaded in pale grey (as opposed to light blue which is used where a TF is bound at at least one, but not all of the gene's, enhancers).

Using these newly defined sets of "integrating" and background genes, I was able to ask if genes with more "integrating" enhancer architectures are enriched for certain biological functions. To do this I performed a gene ontology (GO) term enrichment analysis on the set of "integrating" genes. No GO terms were significantly enriched at the "integrating" genes, however this may be due to

the small number of “integrating” genes (total number= 101) which reduces the power of GO term enrichment analysis.

3.2.5 Properties of TFs driving the “shadow” and “integrating” architectures

I next reasoned that genes favouring different regulatory architectures might be bound by distinct TFs, which could offer insights into either the functions of genes or logic of these regulatory architectures. Following from this, I asked whether I could identify TFs that are bound across a higher (“homotypic” binding) or lower proportion (“lone” binding) of enhancers than would be expected. The problem of identifying TFs bound across a high versus low proportion of a gene’s enhancers can be re-framed to instead ask how TF binding events are distributed amongst genes. If a TF favours homotypic enhancer binding, binding events will be clustered at the enhancers of a subset of genes, and as a result the probability of TF binding at an enhancer is dependent upon the gene. In contrast for a TF favouring “lone binding”, the probability of a binding event at an enhancer is equal across genes. This difference in enhancer binding probabilities (gene dependent versus gene independent) between the two binding preference extremes enables a logistic regression to be used to distinguish between TFs of different binding preferences (see Section 2.4.4 for details).

For each TF I compared the fit of logistic regression modelling the probability that the TF is bound at an enhancer as a function of the gene the enhancer belongs to, to that of a null model where the predicted value was constant across all genes (i.e. gene independent). A greater difference in fit indicates a preference for “lone” binding, whilst a smaller difference in fit suggests homotypic enhancer binding is favoured. As previously mentioned TF-antibody affinity and thus ChIP quality can vary greatly between TFs; TFs which appear to favour lone binding might instead be TFs that immune-precipitated less well during the ChIP. To test if binding preference correlates with number of ChIP-Seq peaks, I plotted differences in model fits for each TF against the total number of TF ChIP-Seq peaks at enhancers. There is a clear correlation between binding preference and number of ChIP-Seq peaks at enhancers; TFs with fewer enhancer ChIP-Seq peaks tend to exhibit “lone” enhancer binding (Figure 3.11). To ensure that the identification of “lone binding” TFs was not driven by ChIP-quality, difference in model fit was regressed against the number of TF ChIP-Seq peaks at enhancers, and the resulting residuals used instead. This resulted in the identification of a group of 13 TFs that were bound across a significantly lower proportion of a genes enhancers than would be expected, which I termed “lone” binding TFs (Figure 3.11). I was also able to identify a

group of 12 TFs that were bound across a significantly higher proportions of a genes enhancers than expected, which I termed “homotypic” binding TFs (Figure 3.11).

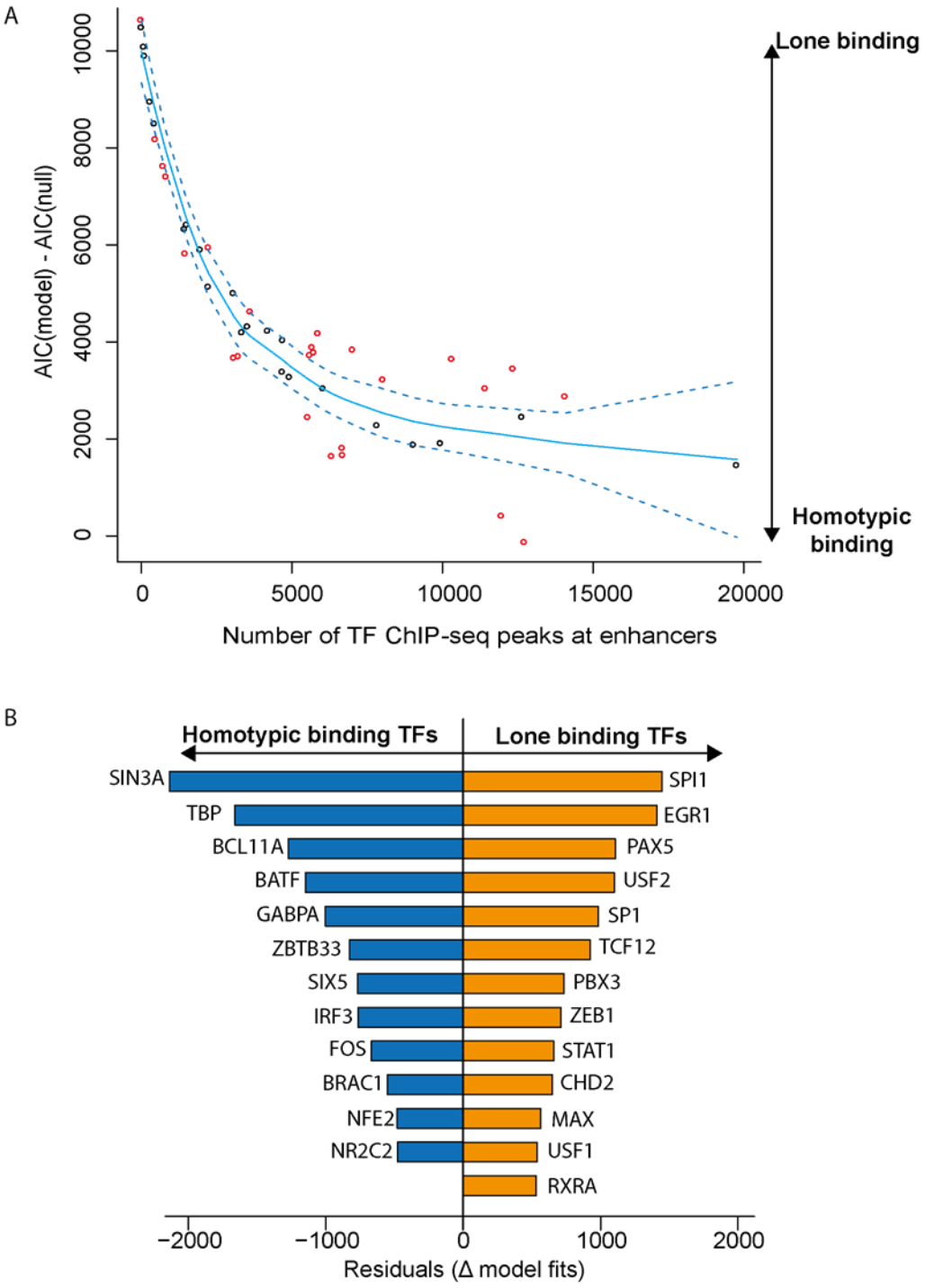


Figure 3.11 “Lone” versus “homotypic” binding preferences of TFs

For each TF a logistic regression was used to model the probability that the TF is bound at an enhancer as a function of the gene the enhancer belongs to (gene dependent). The Akaike information criterion (AIC) (Akaike, 1974) of a null model (where the predicted value is fixed/gene independent) was subtracted from

the AIC of the gene dependent model for each TF. A greater difference in model fit indicates a preference for "lone" binding (bound at a low proportion of a gene's enhancers) and a smaller difference a preference for "homotypic" binding (bound at a high proportion of a gene's enhancers). (A) Difference in model fits plotted against number of enhancer TF ChIP-Seq peaks for each TF (depicted by the coloured circles). The smooth curve fitted (using LOESS) is shown in blue; the dashed blue lines show the 95% confidence limits. TFs (coloured circles) above the 95% confidence upper limit (coloured red) were taken as "lone" binding TFs; TFs below the 95% confidence lower limit were taken as "homotypic" binding TFs (also coloured red). (B) Residuals from the fitted curve for "homotypic" and "lone" binding TFs.

I next examined the properties of the "lone" versus "homotypic" binding TFs, reasoning that this may give an insight into the different regulatory logics conferred by the contrasting TF enhancer distributions. I first observed that the "lone" binding TFs were, with the exception of SP1, all key B-cell specific TFs, whilst the homotypic binding TFs, with the exceptions of Batf and Bc1lla, tended to be more general TFs. This could suggest that house-keeping genes, which you might expect to be regulated by more general TFs, tend to adopt a more "shadow" style architecture, and conversely genes activated in a cell type specific manner have more distinct regulatory information at their enhancers. However this is speculative, and further investigation would be needed to confirm this hypothesis. I wanted a more systematic way to compare properties between the two groups of TFs. To this end, I performed a comparative feature enrichment analysis on the two sets of TFs. As expected both groups of TFs were significantly enriched for positive regulators (log p-value =10), however strikingly only "homotypic" TFs were enriched for negative regulators (log p-value = 5.44), with 7 out of the 11 TFs having negative regulatory potential, compared to one of 12 "lone" binding TFs. "Lone" binding TFs were significantly enriched for responses to external stimuli (5 out of the 11 TFs, log p-value=4.38), which may be expected given the previous observation that the majority of "lone" binding TFs are cell-type specific.

Given the finding that "homotypic" binding TFs tend to have negative regulatory potential, I hypothesized that negatively regulated genes may favour a "shadow" regulatory architecture. If this is the case, I reasoned that genes with a "shadow" architecture might be more lowly expressed than those with an "integrating" architecture. To test this I used the modified similarity metric to select genes targeted by the most similar enhancers, and compared their expression level to a background set (matched for number of unique TFs). No significant difference was found between the median expression levels of genes targeted by the most similar enhancers and the background genes ($p>0.05$). This does not rule out the possibility that genes under negative regulation favour a "shadow" architecture; it is likely that many positively regulated genes also adopt this arrangement, potentially masking any expression differences, especially given the small number of genes tested. Given that all but one of the "homotypic" TFs with negative regulatory potential can

also act as positive regulators, the lack of difference in expression may alternatively suggest that they are not acting exclusively as negative regulators when they are bound "homotypically".

3.3 Discussion

The aim of this chapter was to investigate how TFs are arranged across multiple enhancers targeting the same gene, particularly with respect to whether enhancers tend to bind similar ("shadow" enhancers) or diverse sets of TFs ("integrating" enhancers). I observed that a gene's enhancers show more similar TF occupancies than would be expected at random suggesting a "shadow" architecture is prevalent. However, numerous examples of genes with "integrating" enhancers were observed, where the same gene showed large variation in TF binding across enhancers. Distinct groups of TFs were found to associate with these contrasting models of TF enhancer binding.

The binding of similar sets of TFs at enhancers has been previously observed for hundreds of Drosophila developmental loci (Cannavò et al., 2016; Hong et al., 2008). However, since these studies only focused on genes with enhancers bound by similar sets of TFs, the extent to which genes are targeted by enhancers binding diverse sets of TFs remained unknown. Due to only examining developmental genes, it was also unclear whether this phenomenon extended beyond developmental genes. Examples of individual genes targeted by enhancers with both similar and diverse sets of TFs had been observed in both Drosophila and mammals (Hay et al., 2016; Staller et al., 2015). The work in this section has thus provided the first global insight into TF binding across enhancers targeting the same gene, revealing that, at least in B cells, they generally bind similar TFs. This suggests that in B cells the shadow architecture is pervasive and not limited to key developmental genes. Although it is clear that genes still bind a considerable number of TFs at only a subset of their enhancers, perhaps differing from the image where almost identical sets of TFs are bound at enhancers presented by Hong et al (2008). Although the degree of similarity in TF binding across enhancers, and consequent classification of shadow versus integrating, depends on both the number and identity of TFs included in the analysis, as well as the metric used. Consequently, caution should be applied when comparing these results with other studies that have examined TF binding at enhancers regulating the same gene. It is also perhaps misleading to classify enhancers as shadow versus integrating, as the degree of similarity is on a continuous scale.

Further to the global nature of this approach, another unique aspect was the use of Promoter Capture Hi-C to assign enhancers to target genes. Much of the work on multi-enhancer logic is

carried out on single loci in Drosophila, with enhancers for a given gene identified based on their ability to drive endogenous expression patterns of reporter constructs. This can result in some enhancers remaining unidentified, as often when the expression pattern is fully recapitulated the search for further enhancers is stopped. It is also plausible that not all enhancers regulating a gene will drive expression of a reporter construct when tested individually, especially when they act synergistically with other enhancers. This was found to be the case for several enhancers of the mammalian Troponin I genes (Guerrero et al, 2010). The authors discovered secondary enhancers for each of the three Troponin genes that showed no activity in reporter assays, but acted synergistically with the previously characterised enhancers in-vivo to increase expression level (ten-fold) and spatial precision (Guerrero et al., 2010). This may bias the discovery of enhancer's towards those that have the same activity and act redundantly or additively; enhancers that are perhaps more likely to bind similar TFs. Using Promoter Capture Hi-C to assign enhancers is not likely to be subject to these potential biases. Often, even when epigenetic features such as TF binding are used to identify enhancers, they are assigned to the closest active gene, as was the case for Cannavò et al. (2016). Many enhancers have been shown not to regulate the closest gene, which may result in enhancers being mis-assigned (Mifsud et al., 2015; Sanyal et al., 2012). While linking enhancers to target genes with Promoter Capture Hi-C overcomes these limitations/biases, the caveat is that the enhancers have not been functionally validated in the cell line used.

This analysis presented several challenges, a major one of which was quantifying the similarity of TF occupancy across enhancers belonging to the same gene. This was key in enabling the investigation of TF binding across enhancers targeting the same gene on a global scale. While the binding of an individual TF across enhancers can be simplified as a binary event, with the presence versus absence of a ChIP-Seq peak, enhancers are bound by multiple TFs. Quantifying similarity in TF occupancy thus needs to capture similarities in the binding across enhancers of these multiple different TFs. In addition, different genes have different numbers of enhancers, which presents a further challenge due to introducing another level of multiplicity. For this analysis I used the average fraction of a gene's enhancers bound by each TF (present at the given gene) to characterise genes in terms of enhancer occupancies across enhancers. Using this metric has enabled the investigation of TF occupancies across enhancers targeting the same gene, giving the first global insight into TF binding at multi-enhancer genes. However, the metric is by no means perfect. Genes with different numbers of enhancers could not be analysed simultaneously. As a result for gene level analyses, genes with different numbers of enhancers had to be treated separately, (i.e. a matched control group for each generated) and following this were pooled. The metric is also influenced by the number of TFs bound at each enhancer. For example, one enhancer bound by a

much large number of TFs than other enhancers, will reduce the total similarity of TF occupancy across the enhancers. TFs are also treated independently; however, it is known that many TFs show preferences for binding with certain combinations of other TFs (Bernstein et al., 2012; Gerstein et al., 2012; Stefflova et al., 2013). It may be interesting in the future to incorporate co-binding patterns into the metric.

One caveat with using ChIP-data to annotate TF binding is the potentially limited sensitivity of this assay. Binding events may not be detected due to aspects of the ChIP protocol, for example the antibody not binding to the TF. Alternatively the region may have shown an enrichment in the ChIP-Seq experiment, but fall under a signal threshold used for peak-calling. The extent to which this impacts the results could be investigated by examining the raw read counts or lowering the thresholds for peak calling at other enhancers, when at least one enhancer is found to harbour a peak above the threshold. Another potential confounding factor in the detection of TF binding is that the enhancers targeting a given gene are in close spatial proximity to each other. It has been shown that TF binding at one region may be picked up by ChIP-Seq of another region in close spatial proximity, termed an indirect ChIP-Seq peak (J. Liang et al., 2014). In theory (although unlikely, as discussed below), this could result in overestimating the extent to which TFs are bound across multiple enhancers. However it was found that these indirect ChIP-Seq peaks have lower intensity, and the majority fall under standard thresholds used for peak detecting (J. Liang et al., 2014). As such it not likely that this phenomenon significantly affects the main conclusions of my analysis.

Enhancers occupied by highly similar TFs are likely to exert comparable effects on the promoter. The pervasive use of shadow enhancer architecture seen here therefore suggests that genes tend to use multiple enhancers with similar activities, possibly to confer robustness as seen in Drosophila (Frankel et al., 2010;Perry et al., 2010). However it is becoming apparent that enhancers bound by diverse TFs are also able to exert similar effects on the promoter (Liberman & Stathopoulos, 2009b; Zinzen et al., 2009). One such example are the two *krüppel* enhancers in Drosophila which drive near identical patterns of expression despite binding non-overlapping sets of activators (Wunderlich et al., 2015). Many of the integrating enhancers may therefore also act redundantly and potentially confer robustness. As such, the use of enhancers with similar activities may be even more pervasive than was observed here.

Given that extrinsic fluctuations often affect TF abundance and activity, it is tempting to speculate that the degree of similarity in TF binding across enhancers might influence the level of potential robustness conferred to such fluctuations. It could be hypothesized that a reduction in concentration of a TF which is only bound at a subset of a gene's enhancers will have less of an

impact on regulation than a TF that is bound across a larger proportion of a gene's enhancers. This is due to the activity of only one of the enhancers being impacted. Alternatively, it is possible that having a TF bound across a larger proportion of enhancers may confer robustness. For example if a gene is activated in response to a specific stimuli by a TF, having an increased number of enhancers with binding sites for the TF will increase the probability of it binding at least one of the enhancers. If at least one enhancer is sufficient to achieve expression, as has been demonstrated for numerous genes, this may ensure robust activation (Hay et al., 2016; Xiong, Kang, & Raulet, 2002). Potentially this is analogous to how having several binding sites for a given TF at an individual regulatory element (homotypic binding) increase robustness to mutations (Kilpinen et al., 2013; Spivakov et al., 2012). This may be the case for the two *snail* enhancers in Drosophila, both of which bind Dorsal, and upon reduction in the level of Dorsal confer robustness to expression (Perry et al., 2010).

While reduced expression variability is advantageous for many genes, for some genes stochastic expression may be important to poise a subset of cells for differentiation (Chang et al., 2008; Kalmar et al., 2009). For such genes it is tempting to speculate that this may be achieved in part through varying the degree of similarity in TF binding across enhancers. Single-cell RNA-Seq provides the opportunity to test this experimentally. Single cell RNA-Seq can give a measure of cell to cell variability in expression, and as such could be used to compare expression variability between genes binding similar versus diverse sets of TFs at their enhancers (Marinov et al., 2014). This may be an interesting area for future investigation.

It remains possible that some of the integrating enhancers observed do possess different activities. The binding of different TFs at each enhancer might enable the gene to employ different regulatory logics at different enhancers. Several examples of this have been demonstrated. In mouse myeloid cells, two enhancers were found to act synergistically to drive high expression of the TF PU.1 in an auto-regulatory fashion. One of the elements was found to bind myeloid cell specific C/EBP-α, which was able to increase chromatin accessibility and permit PU.1 binding at the second element (Leddin et al., 2011). Expression of hunchback is also controlled by two enhancers with differing logics; the more distal enhancer appears to attenuate the activity of the proximal enhancer due to binding of repressors, restricting expression in the anterior pole of the Drosophila embryo (Perry et al., 2011). As such, enhancers with differing regulatory logics may be important in ensuring tight control of expression for genes where misexpression has severe consequences.

The abundance and activity of TFs is often controlled by signalling cues (Zhang & Glass, 2013), thus the integrating enhancers discovered may enable genes to use different enhancers to respond to

different cellular signals. The TFs bounds at integrating enhancers, while all active in LCLs, may be differentially active at other developmental time points. This idea is supported by the finding of Cannavò et al. (2016) from *Drosophila*, that whilst enhancers associated with the same gene commonly showed overlapping activity in space and time, this overlap tended to be partial. The observation that many key B-cell specific regulators favoured "lone" binding (i.e were bound at a low proportion of a gene's enhancers) might reflect the differential use of the enhancers at other points in development. Alternatively, the TFs bound at a subset of a gene's enhancers may respond to very similar signalling and developmental inputs; high levels of redundancy have been demonstrated in regulatory networks, and thus many TFs lie downstream of individual master regulators (Macneil & Walhout, 2011). The integration of the TF binding profiles across enhancers with signalling networks may shed further light on whether TFs bound at a subset of a gene's enhancers bring unique or the same cellular information. Incorporation of signalling networks may also be interesting with respect to the previously discussed potential role in robustness of integrating enhancers. The impact on robustness of varying levels of similarity in TF binding across enhancers may be influenced by whether TFs bound at a subset of enhancers are regulated by the same or different pathways.

The identification of distinct groups of TFs that favour "lone" versus "homotypic" binding is challenging to interpret. As previously mentioned the observation that cell-type specific TFs tend to bind at a low proportion of a gene's enhancers, might reflect the modular use of enhancers throughout development. The finding that most TFs which tend to bind across a large proportion of a gene's enhancers can act as positive or negative regulators is intriguing. A recent study found that at the *eve* locus in Drosophila, the TF *hunchback* activates one enhancer for strip seven and represses the other (Auton et al., 2015). If bi-functional TFs are commonly bound across multiple enhancers for a gene, and exert opposing effects on enhancer activity this may underlie these observations.

In conclusion the work in this section has provided a global insight into TF binding across multiple enhancers targeting the same gene, revealing that genes tend to favour a shadow enhancer architecture (enhancers bind similar sets of TFs). Future work on multi-enhancer logic is likely to shed further light on the role and significance of these contrasting enhancer architectures.

4 Investigating the effect of TF binding variation on target gene expression

4.1 Introduction

Recent advances in DNA sequencing technologies have provided an unparalleled insight into human genetic variation. This has revealed that variation amongst humans is extensive, with over 80 million DNA sequence variants identified (Auton et al., 2015). Some of this variation falls at non-coding, regulatory regions and has been implicated in expression variation between individuals and disease (Albert & Kruglyak, 2015; Corradin & Scacheri, 2014). Here I have utilised this naturally occurring genetic variation to investigate the effects of TF binding changes at distal regulatory regions on target gene expression, with the ultimate aim of further elucidating principles of enhancer regulation. To do this I took advantage of a panel of lymphoblastoid cell lines (LCLs) derived from 359 healthy individuals, which have been genotyped by the 1000 Genomes Project, and for which RNA-Seq data are publicly available from the GEUAVDIS project (Auton et al., 2015; Lappalainen et al., 2013). I used TF binding models to computationally predict which variants found in these individuals impact TF binding. I then utilised a Promoter Capture Hi-C (PCHi-C) dataset for one of the LCLs (Mifsud et al., 2015) to link predicted TF binding variants to their target promoters, making it possible to test their association with target gene expression.

The approach taken here differs from the classical eQTL approach in that only variants that are predicted to impact TF binding and are in close physical proximity to the promoter are tested for association with gene expression. This vastly reduces the number of association tests performed per gene, reducing the multiple testing burden and consequently increasing the power to detect expression-modulating variants. The reduced number of tests performed per gene also enables the effects of multiple variants linked to the same gene to be jointly tested. Both the increased power and joint testing may be especially relevant for the discovery of enhancer variants, which often have a lower effect on expression individually.

4.2 Results

4.2.1 TF binding variation predicted across 359 LCLs

I first set out to predict TF binding variation at regulatory regions, including distal enhancers, across a panel of 359 LCLs derived from healthy individuals. The approach I took is outlined in the schematic (Figure 4.1). First, I used the PCHi-C dataset to link distally interacting regions to target genes in an LCL derived from one individual, NA12878, which I then integrated with TF binding data for 52 TFs from ENCODE for the same LCL, to annotate TF binding at these distally interacting regions. Promoter Capture Hi-C is unable to detect interactions with either of the fragments immediately adjacent to the baited fragment due to high levels of noise. To identify proximal and short-range regulatory regions that may have been missed due to this, I profiled TF binding in a proximal window around the TSS of all genes. The proximal window size was set to the average length of three HindIII restriction fragments (TF-bound, and either baited or significantly interacting restriction fragments) which is 18kb. To illustrate this the resulting regulatory binding annotation for an example gene, *KIAA0141*, is shown in Figure 4.2a. A region lying 231kb upstream of *KIAA0141*, and a second region 35kb further upstream, were detected as interacting with *KIAA0141* by PCHi-C. Each fragment was found to harbour a single regulatory region, with three and eight TFs respectively. In addition, three proximal TF-bound regions were identified. Carrying this out for all genes resulted in a map of TF binding at regulatory regions for 13,080 genes in one LCL.

LCLs derived from 359 different individuals have been genotyped by the 1000 Genomes Project, enabling variants falling at the TF bound regions in the LCL derived from NA12878 to be identified. I reasoned that some of these variants might disrupt the TF binding motifs, potentially impacting binding in individuals with the respective allele. Without TF binding data for the remaining 359 LCLs, the impact of variants on binding cannot be directly assessed. However, position weight matrices (PWMs), which model the sequence binding preferences of TFs and enable sequences to be scored based on their binding energy for respective TF, can be used to predict when sequence variants might impact TF binding. Following from this rationale, I used PWMs to assess the impact of variants on the binding of TFs found at the respective regulatory regions in GM12878 (see Section 2.4.5 and 2.5.5 for further details). Notably I tested the joint effect of the combination of all variants at each regulatory region, resulting in a binding impact prediction for entire the TF bound region. The rationale behind this is that multiple variants, in particular SNPs, may fall at the same TF binding site, and thus their joint effect needs to be assessed. For the previous example gene, *KIAA0141*, binding variation was predicted for one TF, SPI1, at the distal regulatory region lying 266kb

upstream of the gene. At this 509bp distal regulatory region, a single G to A substitution reduced the predicted total affinity of the sequence for SPI1 by 1.7 fold (Figure 4.2b).

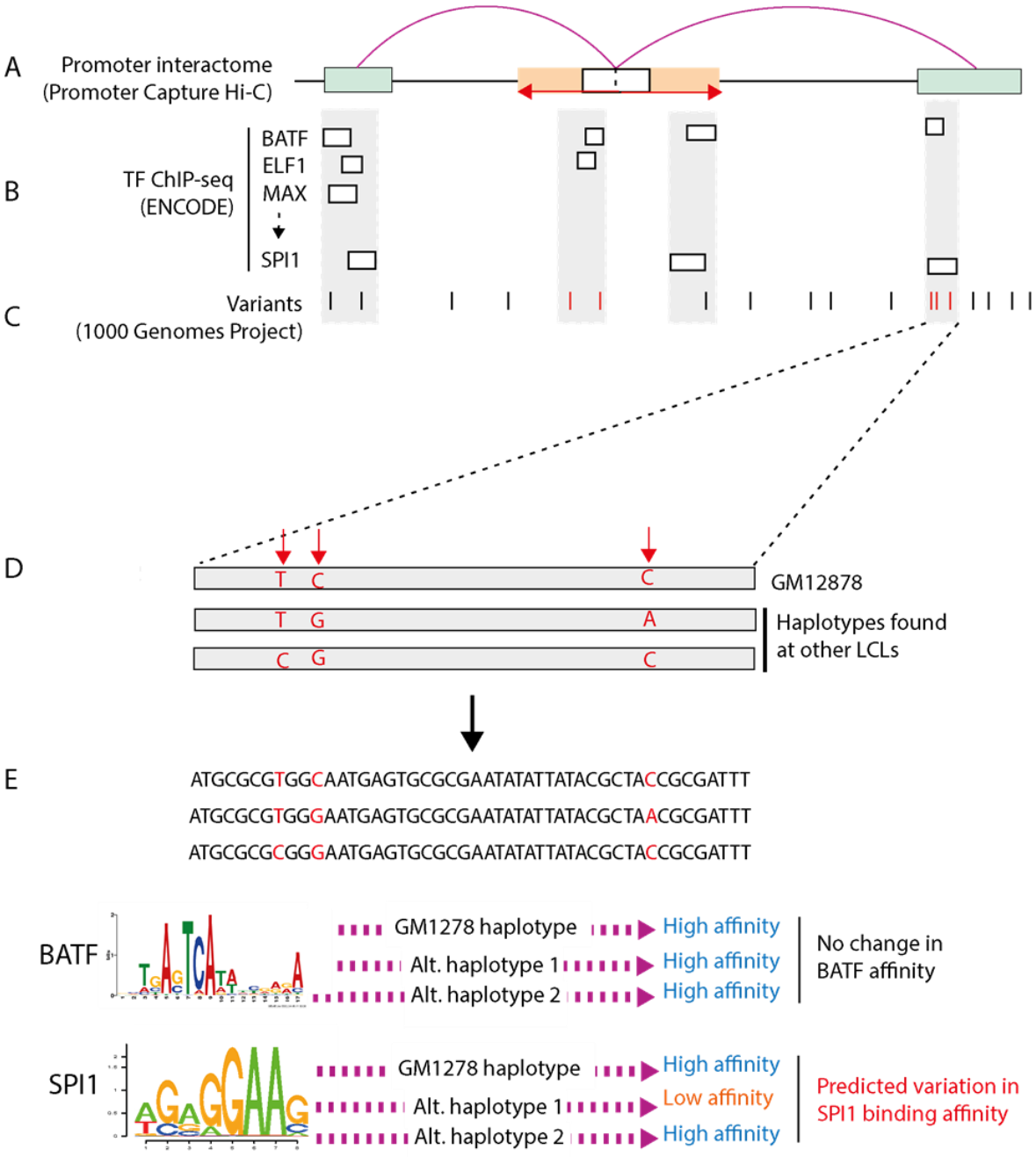


Figure 4.1 Schematic outlining binding variant predictions

(A) Promoter capture Hi-C (Mifsud et al., 2015a) was used to identity significant promoter-interacting regions at HindIII restriction fragment resolution in the LCL GM12878. The baited restriction fragment is depicted by an orange rectangle, and non-baited restriction fragments as green rectangles. Two significant promoter interacting regions (linked to baited fragment by pink arches) were identified for this hypothetical gene. An 18kb proximal window was defined, centred on the mid-point of the promoter-containing fragment. (B) TF ChIP-Seq data for 52 TFs from ENCODE (ENCODE Project Consortium, 2012b) was used to annotate TF binding at promoter interacting regions and the promoter proximal window. TF-bound regions for four TFs are shown here, depicted by solid white boxes. TF-bound regions for all TFs were overlapped, to produce a set of

composite TF-bound regions which are depicted by the grey shaded regions. (C) Variants from the 1000 Genomes Project (Auton et al., 2015b) overlapping with TF-bound regions at the promoter-interacting fragments and promoter-proximal region (red bars) were identified. (D) For TF bound region containing at least one variant, the genotypes of all 359 LCLs were examined to identify a set of unique haplotypes for the region. Three SNPs fall at the hypothetical TF bound region highlighted, and LCLs were found to have one of three unique haplotypes (TCC (GM12878 haplotype), TGA or CGC) at their alleles. (E) The genome sequence of the TF bound region was extracted, and the genotypes for each haplotype "injected" to create a haplotype-specific sequence. For each of the TFs bound at the region for which a position weight matrix (PWM) was available (41 out of 52 TFs), PWMs from ENCODE (Kheradpour & Kellis, 2014) were used to calculate a normalised TF binding affinity for each of the haplotypes. The fold change in affinity relative to that of GM12878 was calculated for each TF at each alternative haplotype; haplotypes with changes over a threshold were taken as binding variants for that given TF, and labelled as a low affinity allele. For the highlighted TF bound region normalised binding affinities for BATF and SPI1 (found bound in GM12878) were computed for each haplotype. SPI1 binding affinity at the alternative haplotype 1 (but not alternative haplotype 2) showed a decrease relative to GM12878, resulting in this haplotype being classified as low affinity and the TF bound region as harbouring an SPI1 binding variant. No change over the threshold was found between the alternative haplotypes and GM12878 haplotype for BATF binding affinity.

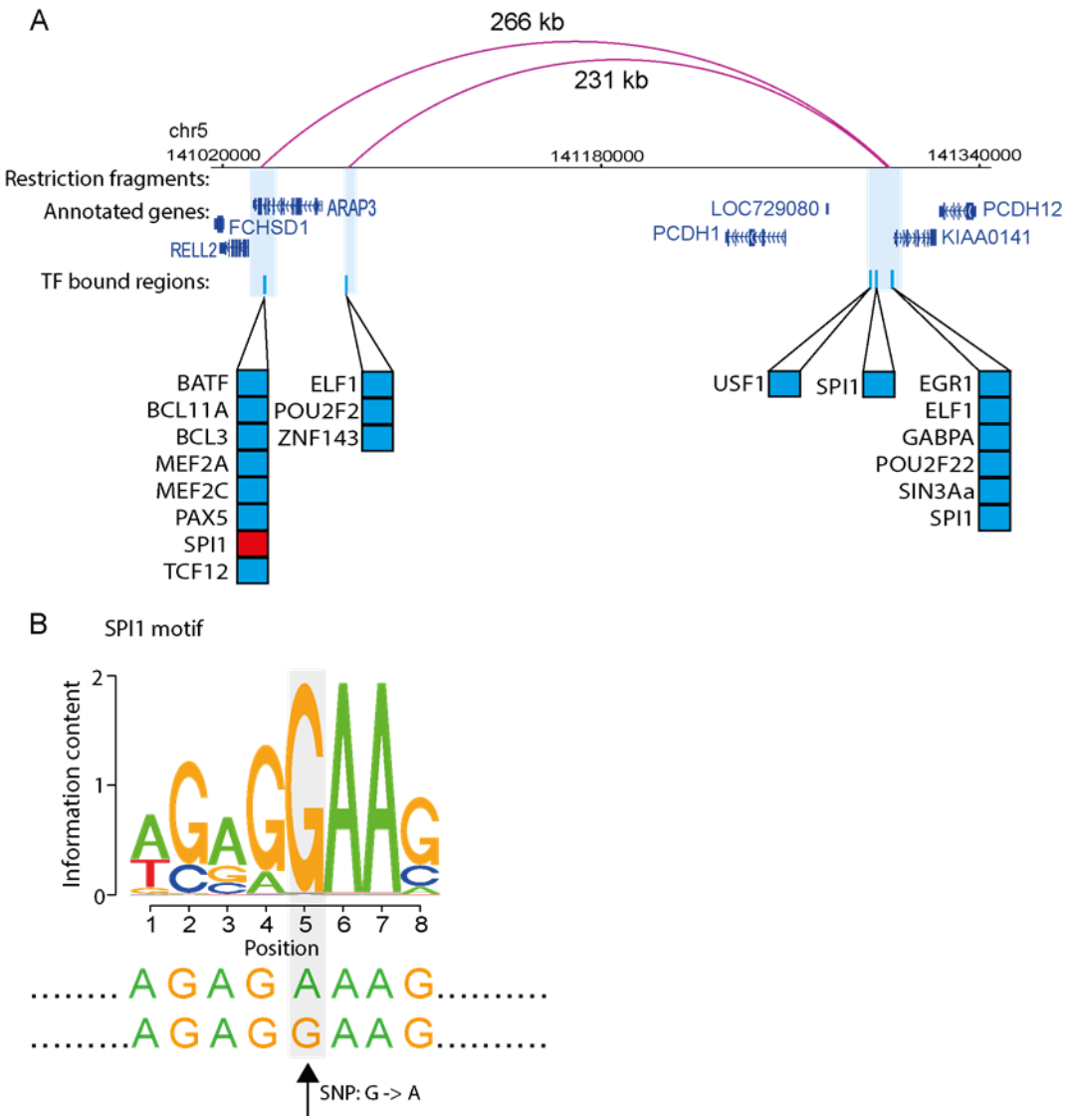


Figure 4.2 TF binding mapped at distal and proximal regulatory regions of KIAA0141, and TF binding variation predicted an SPI1 binding variant identified

(A) Promoter capture Hi-C identified two fragments (shaded in in pale blue, and linked by pink arches) significantly interacting with KIAA0141 in the LCL GM12878. ChIP-Seq data for 52 TFs were used to map TF binding at these distal fragments and within a proximal window around the TSS of KIAA0141 in GM12878. The location of the TF bound regions are shown in blue; the TFs found at each of these regions are listed below. For each of the TF bound regions, the different haplotypes amongst 359 other LCLs were identified; PWMs for each of the TF bound were used to predict changes in binding affinity for each of the haplotypes compared to GM12878. In this way binding variation was predicted for SPI1 at the distal region lying 266kb away from the promoter of KIAA0141. (B) An example of one of the SPI1 PWMs used to identify haplotypes (of the 520bp TF-bound region) with altered SPI1 binding affinity compared to GM12878. Amongst haplotypes predicted to show reduced SPI1 affinity compared to GM12878, a single G to A substitution at an 8bp SPI1 binding site was responsible for the predicted reduction in SPI1 binding affinity.

Using the approach outlined above I predicted 1,491 TF binding variants, at 1,244 out of 41,399 TF bound regions (2.4%). This resulted in a set of 1,765 genes with predicted binding variation at a regulatory region. The number of genes with a binding variant at a regulatory region exceeds the number of predicted TF binding variants because some enhancers connect to multiple genes. I examined the proportion of variants that lie at regulatory regions targeting multiple genes, revealing that 45% (665) of the binding variants connect to multiple genes (Figure 4.3a). Out of these multi-gene binding variants just over half were linked to more than two genes. Given that the majority of genes are targeted by multiple regulatory elements, I next asked whether there are some genes which have multiple predicted binding variants across their regulatory regions. Indeed, I found that 37% (660) of genes were predicted to have multiple binding variants, out of which 43% were predicted to have more than two binding variants (Figure 4.3b).

4.2.2 Predicted TF binding variants are enriched for sites of differential chromatin accessibility.

To validate this set of predicted TF binding variants, I took advantage of a DNase I sensitivity (a measure for chromatin accessibility) quantitative trait loci (dsQTLs) dataset by Degner et al. (2012). This consists of ~9000 dsQTLs in LCLs derived from 70 Yoruba individuals. Given that dsQTLs frequently associate with changes in TF binding (Degner et al., 2012), if my predicted TF binding variants reflect actual differences in TF binding I hypothesised that would be enriched for dsQTLs. To test this, I compared the proportion of SNPs predicted to alter TF binding that are detected as dsQTLs by Degner et al. (2012) to the same proportion for randomly selected TF-bound SNPs not predicted to alter TF binding (Figure 4.4). This also presents an opportunity to examine how the

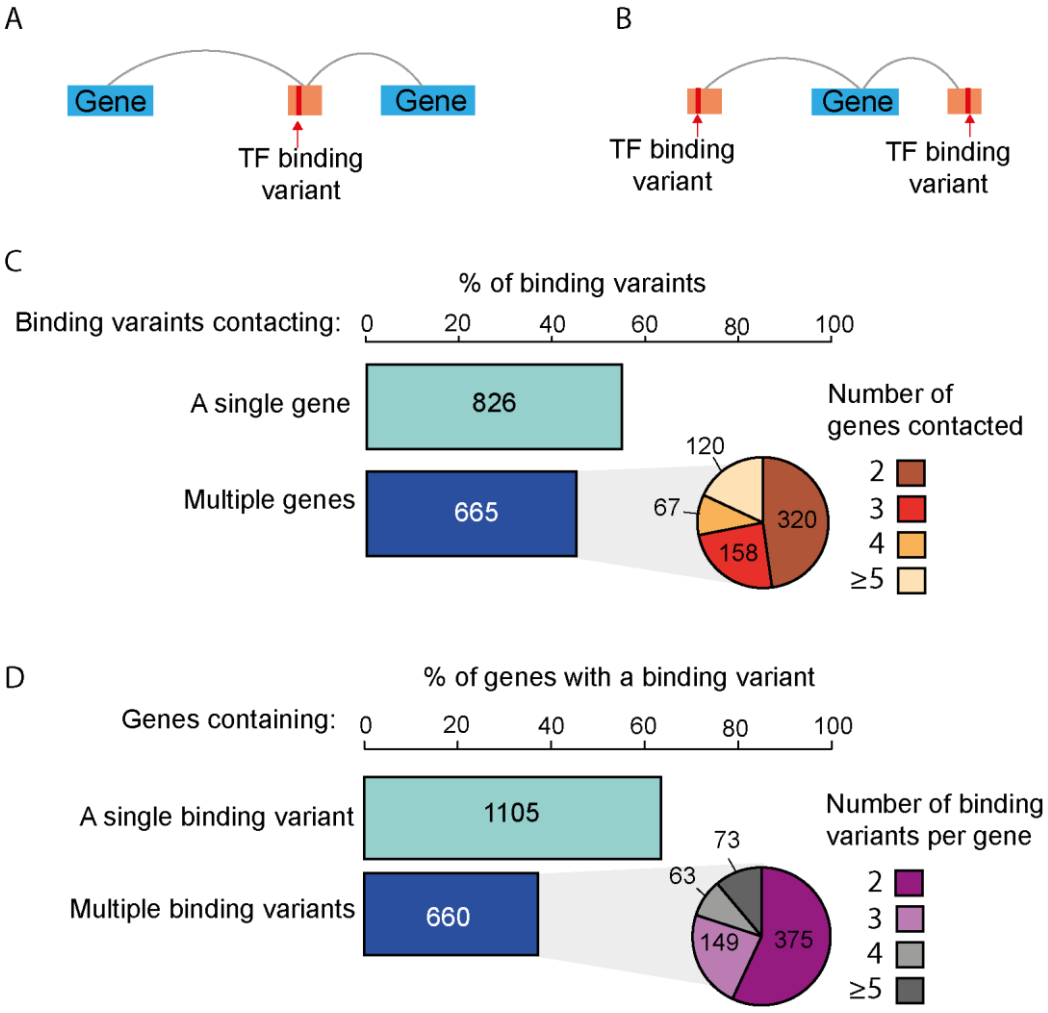


Figure 4.3 Multi-connectivity between TF binding variants and genes in the LCL GM12878

A single predicted binding variant can be linked (via Promoter Capture Hi-C or a promoter proximal window) to multiple genes (A), whilst a given gene can also be targeted by multiple binding variants (B). (C) A barplot shows the percent of binding variants (out of all those connected to at least one gene) that contact a single gene versus those that contact more than one gene (as illustrated in (A)). Out of the binding variants that contact multiple genes, the proportion of binding variants contacting two, three, four and five or more genes are shown in a piechart. (D) A barplot shows the percent of genes which are targeted by a single binding variant versus multiple binding variants. Out of the genes that harbour multiple binding variants across their regulatory regions, the proportion containing two, three, four and five or more binding variant are shown in a piechart.

choice of threshold used for defining what magnitude change in TF binding affinity constitutes a change in binding (is "binding-altering"), affects potential dsQTL enrichment. To this end, I used a range of thresholds to define sets of "binding-altering" SNPs. I observed that SNPs predicted to alter binding were significantly enriched for dsQTLs compared to predicted binding-invariant SNPs

for all thresholds used (excluding the control threshold of zero), and that the enrichment tends to increase with increasing threshold level (Figure 4.4). The observed enrichments suggest that the predicted "binding-altering" variants are more likely to be dsQTLs, providing evidence that the predicted TF binding variants likely reflect actual binding differences in vivo. The increase in the proportion of binding variants that are dsQTLs with increasing threshold plateaus at a threshold of around 0.3. I therefore chose this threshold of 0.3 for defining predicted TF binding variants. At this threshold, SNPs predicted to alter binding show a 1.8-fold enrichment for dsQTLs than those not predicted to alter TF (Figure 4.4).

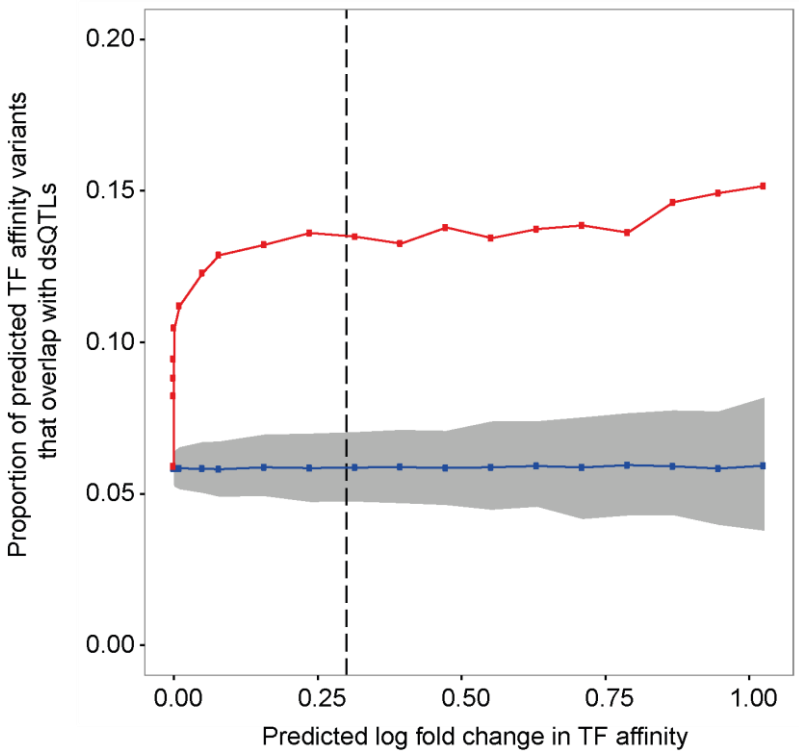


Figure 4.4 Predicted TF binding variants are significantly enriched for DNase hypersensitivity QTLs (dsQTLS)

SNPs underlying the affinity changes at each regulatory region that showed an affinity change relative to GM12878 over a range of thresholds (x-axis) for any TF, were identified ("predicted TF affinity variants"). The proportion of predicted affinity variants identified at each threshold overlapping with dsQTLs are shown (red points). For each threshold a control set of SNPs (matched in number to the predicted binding variants) falling at regions showing no affinity change/no affinity change under threshold were randomly selected, and the proportion overlapping with dsQTLs were calculated. This was repeated 1000 times and the mean proportion of binding invariant SNPs overlapping with dsQTLs calculated for each threshold (blue points). The grey shaded area indicates the limits of the 5% and 95% confidence intervals. Predicted TF binding variants at all thresholds excluding zero were significantly enriched for dsQTLs ($p<0.0001$ for all thresholds).

4.2.3 Identification of expression-modulating TF binding variants

Having predicted TF binding variation at regulatory regions across a panel of 359 LCLs, I next set out to investigate the effect of binding variation on target gene expression. I first filtered the set of predicted binding variants to those connected to genes whose expression ranked in the top 50th percentile, to avoid spurious associations caused by genes with low read counts which are typically nosier. This resulted in a set of 1,194 predicted binding variants connected to 1,530 genes. In brief, for the 1,110 genes (out of 1,530) with a single predicted TF binding variant, I used linear regression to test for association between target gene expression (obtained from the GEUVADIS project) and the binding variant genotype in the corresponding LCL (homozygote for the high-affinity binding allele, heterozygote and homozygote for the low-affinity binding allele). The presence of a significant association suggests that the binding variant affects gene expression. To illustrate this, the data for an example gene, *KLF6*, for which I detected a significant positive association between a binding variant and target gene expression, is shown in Figure 4.5. For *KLF6* I predicted variation

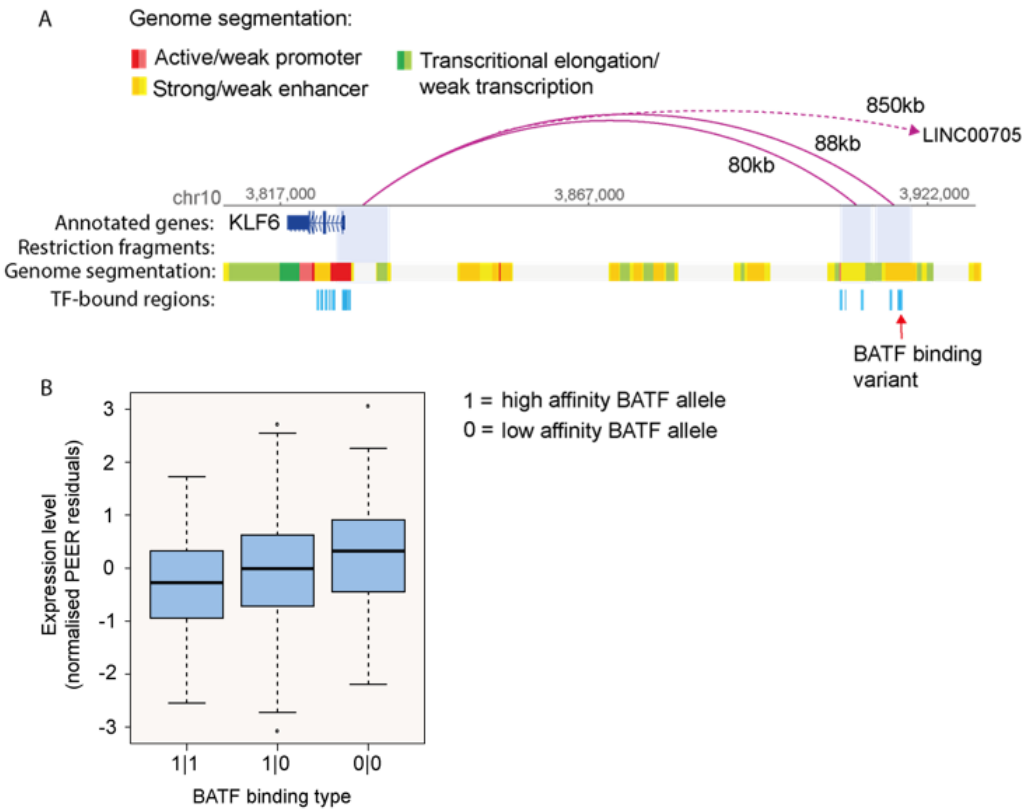


Figure 4.5 Expression of KLF6 significantly associates with a predicted BATF binding variant located at a distal element 88kb away from the KLF6 promoter

(A) Genome browser representation of the distal promoter interactions (pink arches) of KLF6 in the LCL GM12878, as detected by Promoter Capture Hi-C (Mifsud et al., 2015). Two out of the three fragments interacting with KLF6 are shown; the third fragment, which is located 850kb away from the KLF6 promoter

and contains the gene LINC00705, was omitted due to space constraints. Genome segmentation tracks for GM12878 are shown (Ernst et al., 2011; Kheradpour et al., 2013). TF bound regions at the two distally interacting fragments and TSS-proximal window are depicted in azure blue. The far-right TF bound region, which interacts with the KLF6 promoter 88kb away, harbours a variant predicted to impact BATF binding across an additional panel of 359 LCLs. (B). Boxplot showing mRNA levels (as measured with RNA-seq) of the LCLs, split according to their predicted BATF binding type (homozygote for the predicted high affinity allele, heterozygote and homozygote for the low affinity allele). KLF6 expression is significantly associated with BATF binding type (p-value=1.8×10^{-4}, effect size =1.51); LCLs with one or two copies of the allele with reduced BATF affinity show increased *KLF6* expression. RNA-Seq data from Lappalainen et al. (2013) was used, where the PEER algorithm had been applied to remove hidden confounding factors, resulting in PEER residuals which were further transformed to a normal distribution.

in the binding of the TF BATF at a distal region located 88kb away from the *KLF6* promoter. *KLF6* expression correlates with BATF binding affinity (p-value=1.8×10^{-4}, effect size=1.51), with individuals homozygous for the high-affinity binding allele for BATF showing the lowest expression, and individuals homozygous for the low-affinity binding alleles showing the highest level of expression (Figure 4.5B). This suggests that BATF acts as a negative regulator of *KLF6*. For the remaining 420 genes with at least two TF binding variants across their regulatory regions, I used a multiple regression to jointly test the effects of the variants on gene expression. This is not usually possible with a standard eQTL analysis since the number of terms (i.e. variants tested per gene) in the regression model would be overwhelmingly high.

Out of the 2,268 binding variant-target gene expression associations tested, 261 (12%) showed a significant association at a gene-level FDR of 10% (refer to Methods for details on multiple testing correction), involving 245 genes. Out of these 261 associations, 101 involved variants at distal regulatory regions and 160 variants at proximal regions. I next compared the proportion of distal versus proximal variants that associate with gene expression. I found that 6% of distal variant-expression associations were significant, compared to 26% of proximal-variant expression associations (Figure 4.6). Despite detecting a lower proportion of distal-variant expression associations, an appreciable number of distal variants still showed significant association with gene expression and were considered further.

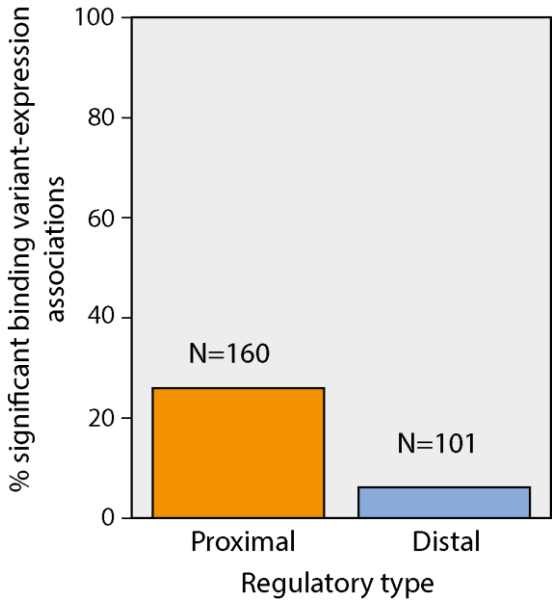


Figure 4.6 Proportion of proximal and distal predicted TF binding variants that associate with target gene expression

All predicted TF binding variants were tested for association with target gene expression, by regressing target mRNA level against binding variant genotype in 359 LCLs. Significant variant-expression associations were selected at a gene-level 10% FDR. The percent of proximal binding variants (assigned to target gene via a proximal window around the TSS) and distal binding variants (assigned to target gene using promoter capture Hi-C) that significantly associate with target gene expression are shown.

4.2.4 The majority of variant-expression associations identified are novel

I first asked how the significant distal and proximal variant-expression associations compared to those identified by the GEUVADIS consortium, which carried out a traditional eQTL analysis on the same panel of LCLs with the same genotype and expression data. As GEUVADIS used an FDR of 5% to correct for multiple testing across genes (as opposed to 10% FDR used here), I adjusted the multiple testing correction I used to 5% FDR to make the results comparable. As such, any associations identified here but not by GEAUVDIS will be due to either looking at distal regions mapping larger than their "cis"-window, increased sensitivity due to a reduced multiple testing burden or using a multiple regression to jointly test variants. Out of the 209 binding variant-expression associations I identified that were still significant at 5% FDR and were caused by a single underlying SNP, 81 (39%) were identified in the GEUVADIS analysis whilst the remaining 128 (61%) were not identified by the GEUVADIS analysis and thus novel (Figure 4.7). Strikingly 39% of the novel associations identified involve a variant outside a promoter region, resulting in 51 novel distal variant associations (Figure 4.7). Interestingly, of the novel associations identified, 57% involved genes already identified as eGenes (73 genes) by GEUAVDIS but with different variants, whilst 43%

were with genes for which GEUVADIS found no significant associations (55 genes) – and as such are novel eGenes. However, GEUVADIS were able to identify significant associations with 3,124 genes for which I found no association; the variants underlying these associations were not predicted to affect TF binding and therefore not considered in my approach.

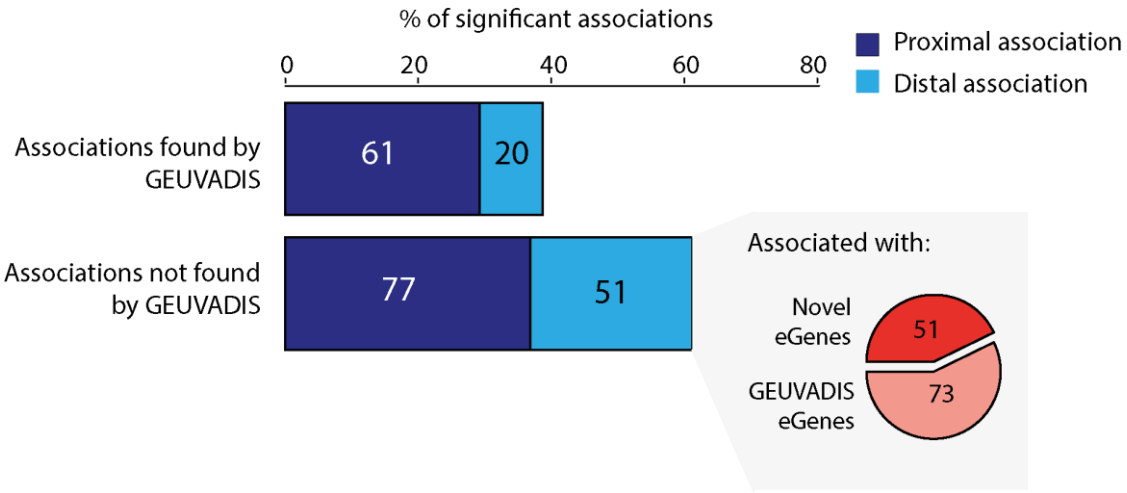


Figure 4.7 The majority of the variant-expression associations were not previously identified by a standard eQTL analysis carried out on the same data

Significant variant-expression associations were compared to those previously identified by the GEUAVDIS consortium, which carried out a standard eQTL analysis on the same LCLs, using both the same variation and RNA-Seq data (Lappalainen, Sammeth, Friedländer, 't Hoen, et al., 2013b). The barplot shows the percent of significant variant-expression associations identified in this study which were also discovered by the GEUAVDIS analysis (39%) versus those that were not previously identified by GEUAVDIS (61%). The bars are split according to whether the variant was defined as proximal (falling within 9kb of the target gene TSS; dark blue) or distal (at an interacting fragment detected by Promoter Capture Hi-C; light blue) in this study. The piechart shows the proportion of novel variant-expression association which involve genes for which the GEUVADIS analysis identified other variant-expression associations (73) versus those that involve genes for which GEUVADIS did not identify any significant variant-expression associations (51).

All of the novel associations found had distances <1MB between the binding variant and affected gene, and were therefore tested in the GEUVADIS analysis. As such, they have been detected here due to either the increased sensitivity of my approach or the use of a multiple regression to jointly test variants connected to the same gene. Increased sensitivity may enable detection of associations with smaller effect sizes (betas). To test if I was able to identify associations with smaller betas than GEUAVDIS, I compared the betas of associations identified by both GEUVADIS and my approach, to those only identified in my approach. Given that significant associations detected uniquely by my approach have a higher proportion of distal variants compared to those

also identified by GEUVADIS, I compared the coefficients separately for proximal and distal associations. The coefficients were squared, and the square root then taken to make them all positive, allowing the magnitudes to be compared. Coefficients detected uniquely in my approach show a significantly smaller coefficient than those also identified by GEUVADIS for both proximal and distal variants (Figure 4.8, two-sample Wilcoxon test, proximal: w=737, p-value=5.0×10^{-12}; distal: w=204, p-value=9.4×10^{-5}).

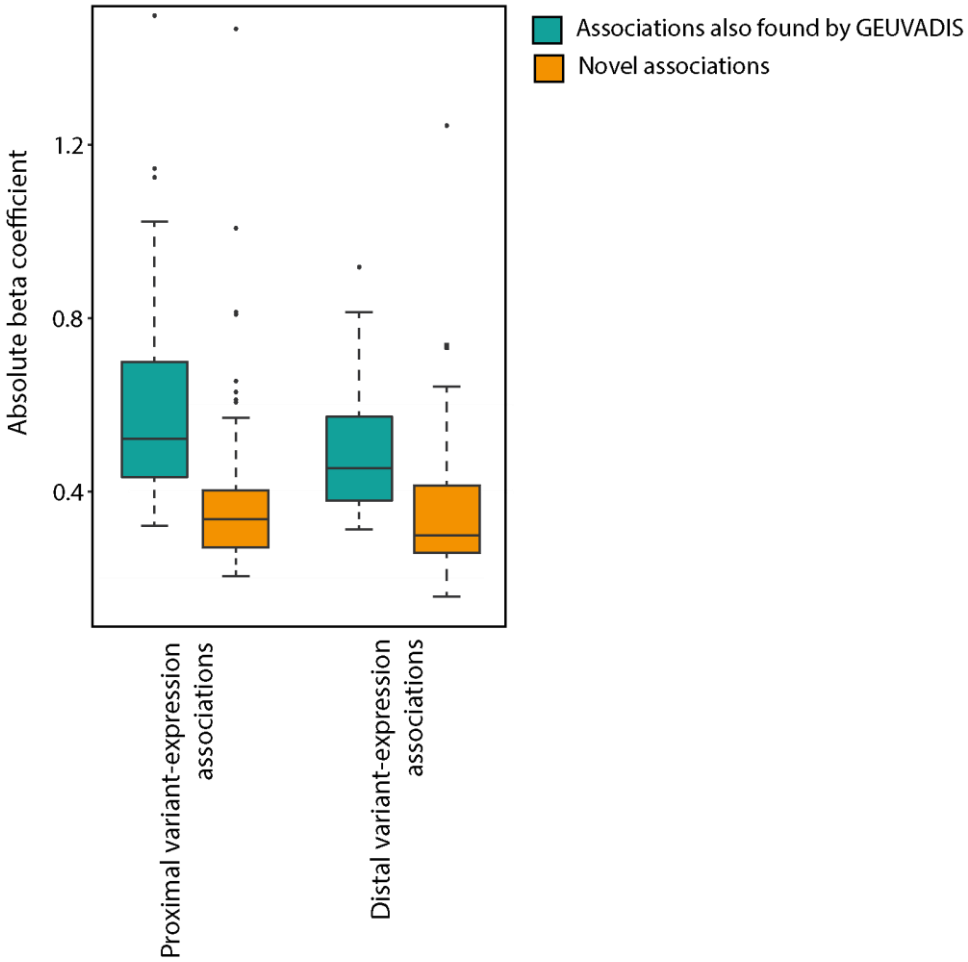


Figure 4.8 Comparison of the magnitude of regression coefficients between significant expression-variant associations also found by GEUVADIS and novel associations

The regression coefficients of all significant variant-expression associations (at 5% gene-level FDR) were squared and the square root taken, to transform to absolute values. The boxplot shows the transformed coefficients for associations involving proximal (within 9kb of the target gene TSS) versus distal (on fragments interacting with the target gene as detected by promoter capture Hi-C) variants, split by whether the associations were previously identified by the GEUAVDIS analysis (cyan) or not (orange).

To demonstrate that the increased sensitivity of my approach is due to the reduced multiple testing burden, I compared the proportion of multiple testing-corrected GEUVADIS association p-values that were above the significance threshold, performing the correction for all variants tested by GEUVADIS versus or for only those variants prioritised in my approach (those predicted to alter TF binding at regulatory regions). Applying multiple testing across only prioritised variants resulted in a higher percentage of significant associations than when multiple testing correction was applied across all GEUVADIS tested variants (11% and 7% respectively at 5% FDR). This also held at 10% FDR, where 14% of p-values corrected for multiple testing across prioritised variants were significant, compared to 8% when correcting across all GEUAVDIS tested variants (Figure 4.9a). This

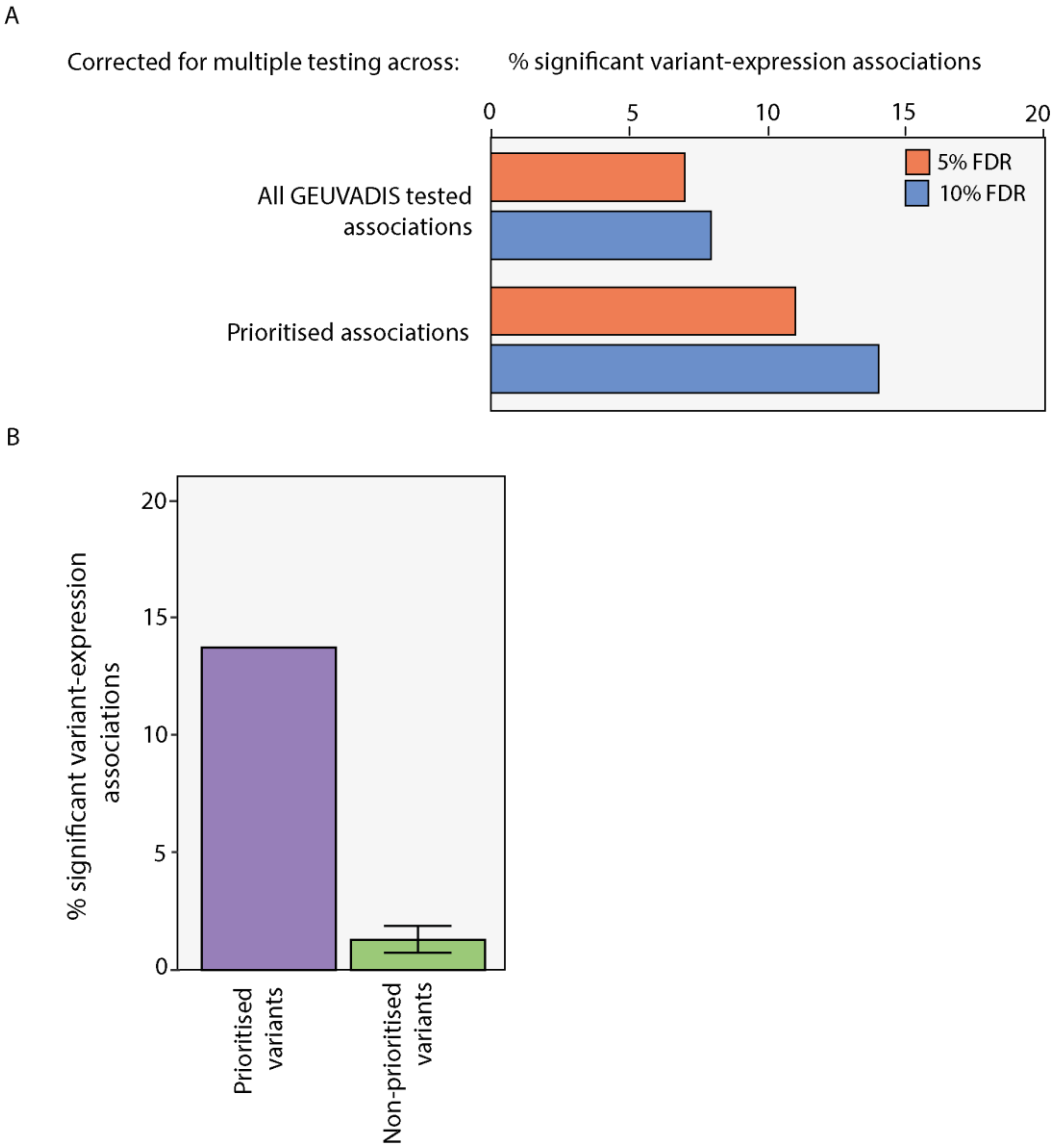


Figure 4.9 Increased sensitivity over GEUAVDIS analysis due to reduced multiple testing burden

(A) Proportion of GEUVADIS association p-values, for variants prioritised in this study, that were above the significance threshold, correcting for multiple testing across all variants tested by GEUVADIS versus correcting for multiple testing across only those variants prioritised in my approach (those predicted to alter TF binding

at regulatory regions). Correcting for multiple testing across only prioritised variants retained a higher percentage of significant associations (11% at 5% FDR, 14% at 10% FDR) than when multiple testing correction was applied across all GEUVADIS tested variants (7% at 5% FDR, 8% at 10% FDR) (B) The proportion of prioritised variants that are eQTLs versus the proportion of randomly selected variants (matched in number to the prioritised variants) that are eQTLs. The mean proportion of random variants that are eQTLs is shown from 1000 permutations. The error bar represents 95% confidence interval.

demonstrates that the variant prioritisation approach that I took does indeed increase sensitivity through reducing the burden of multiple testing. It is possible that the prioritisation approach reduces the multiple testing burden, and thus increases the power, simply through testing a reduced number of variants; perhaps the same number of randomly selected variants would allow the recovery of a comparable number of significant associations? To test this, I compared the proportion of significant variant-expression associations between the prioritised variants and randomly selected variants. Almost none of the randomly chosen variants showed significant associations with this approach (Figure 4.9b). Thus, the identity of the variants prioritised is important in reducing the multiple testing burden.

4.2.5 Genes impacted by multiple TF binding variants

I next asked whether I was able to identify any genes with multiple independent TF binding variant-expression associations. Out of the 420 genes for which I predicted more than a single TF binding variant, using a multiple regression I was able to identify 16 genes whose expression showed significant associations with multiple independent TF binding variants. A further 61 genes showed an association with a single one of the variants. As an example, a gene for which I detected two expression-associated binding variants located at discrete distal regulatory regions is shown (Figure 4.10a). Expression of the nuclear receptor gene, *NR2F6,* was significantly associated with variation in SMC3 binding at a distal regulatory region 41 kb away, as well as in the binding of SRF at a distal element 19 kb away (multiple regression p-value=4.1×10^{-7}, SMC3 term p-value=3.0×10^{-4}, effect size=0.26; SRF term p-value=1.2×10^{-7}, effect size= 0.61). To illustrate that the effects of the SMC3 and SRF binding variants on expression are independent, for each binding variant I limited the analysis to LCLs with just one genotype for the other variant, and plotted expression against the genotype of this binding variant (Figure 4.10b and c). The expression of *NR2F6* increases with both the loss binding of SRF and SMC3, even when the genotype of the other binding variant is constant,

demonstrating that the effects are indeed independent (Figure 4.10b and c).

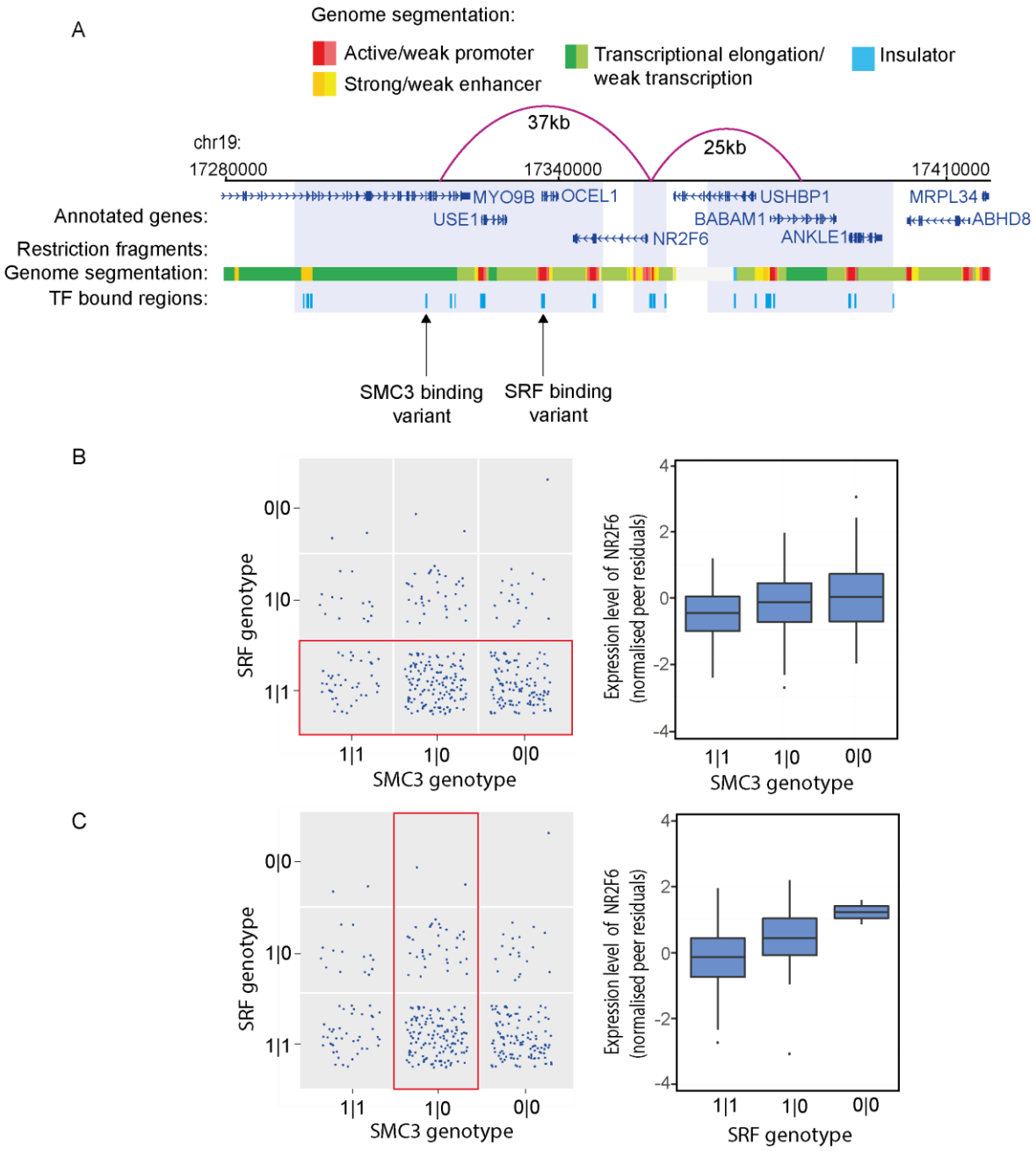


Figure 4.10 Expression of NR2F6 significantly associates with two independent binding variants

(A) Genome browser representation of NR2F6 promoter distal interactions (represented by pink arches) as detected by promoter capture Hi-C (Mifsud et al., 2015a) in the LCL GM12878. The genome segmentation track for GM12878 is also shown (Ernst et al., 2011; Kheradpour et al., 2013). TF bound regions at the distally interacting fragments (pale blue) and NR2F6 TSS-proximal window are depicted in azure blue. The distal fragment downstream of NR2F6 contains two discrete TF bound regions which harbour predicted TF binding variants: one 44kb away from the NR2F6 promoter and the other 19kb away, containing variants predicted to impact SMC3 and SRF binding respectively across the 359 LCLs. (B) The left hand (LH) plot shows the SRF and SMC3 variant genotypes of the LCLs. Each dot represents an LCL derived from a different individual, and is plotted in one of the nine squares according to the SRF and SMC3 binding type of the LCL. For example, an LCL in the bottom far left square has two high affinity alleles for SRF and also two high affinity alleles for SMC3. Positions of the LCLs within each of the squares do not mean anything, they are randomly scattered

to allow all LCLs to be visualised. LCLs homozygote for the high affinity SRF allele were selected (LCLs within red box on LH plot), and their expression level plotted against SMC3 genotype in the RH boxplot. (C) The LH plot again shows the SRF and SMC3 variant genotypes of the LCLs. This time LCLs heterozygote for the high affinity SMC3 allele were selected (LCLs within red box on LH plot), and their expression level plotted against SRF genotype in the RH boxplot.

4.2.6 TF binding variation at promoters affects expression of distally interacting genes

I next focused the analysis on the set of 101 distal expression modulating binding variants (described in Section 4.2.3), predicted to affect expression of the associated gene via impacting enhancer activity. Unexpectedly, I observed cases where the binding variant mapped to the promoter of another gene. One such example is the TCF12 binding variant, which is located at the promoter of *BEND6*, yet affects the expression of *RAB23* that is located 266kb away (p-value=4.1×10^{-12}, effect size=-0.19; Figure 4.11). Strikingly 62 out of the 107 distal variant-expression associations detected involved variants that lay on a restriction fragment containing another gene's promoter (illustrated in Figure 4.12a). The median length of significantly interacting TF-bound restriction fragments is 6000bp, and they can be much larger; it is thus possible that these variants may be located outside the promoter region, for example, mapping to a close-range enhancer. To establish if these variants are more likely to be at proximal enhancers or promoters, I examined the distance between the binding variant and the TSS of the nearest gene. I also took advantage of the publicly available genome segmentation data for the LCL GM12878 to investigate the chromatin state of the regions surrounding the TF binding variants. The median distance between the binding variant and TSS of the closest gene was ~700bp, consistent with their likely location within the promoter region (Figure 4.12b). Furthermore, I found that 70% of the binding variants have promoter-associated chromatin states, with just under 60% defined as active promoters and ~10% defined as weak or poised promoters (Figure 4.12c). The remaining 30% of binding variants were largely defined as having an active enhancer state. This suggests that the majority of binding variants fall within promoter regions, while a subset of them are likely at close-range enhancers.

Two possible scenarios may lead to the observed associations between a promoter variant and expression of a distal gene. Binding variation at the promoter might impact expression of the proximal gene, the protein product of which is involved in regulation of the distal gene, either directly or indirectly via downstream signalling processes. Alternatively, the promoter might act in an enhancer-like manner for the distal gene. If the former scenario is true, the expression of the gene whose promoter contains the binding variant must also be associated with the variant, which is not a requirement for the latter model. I therefore investigated whether the expression of the

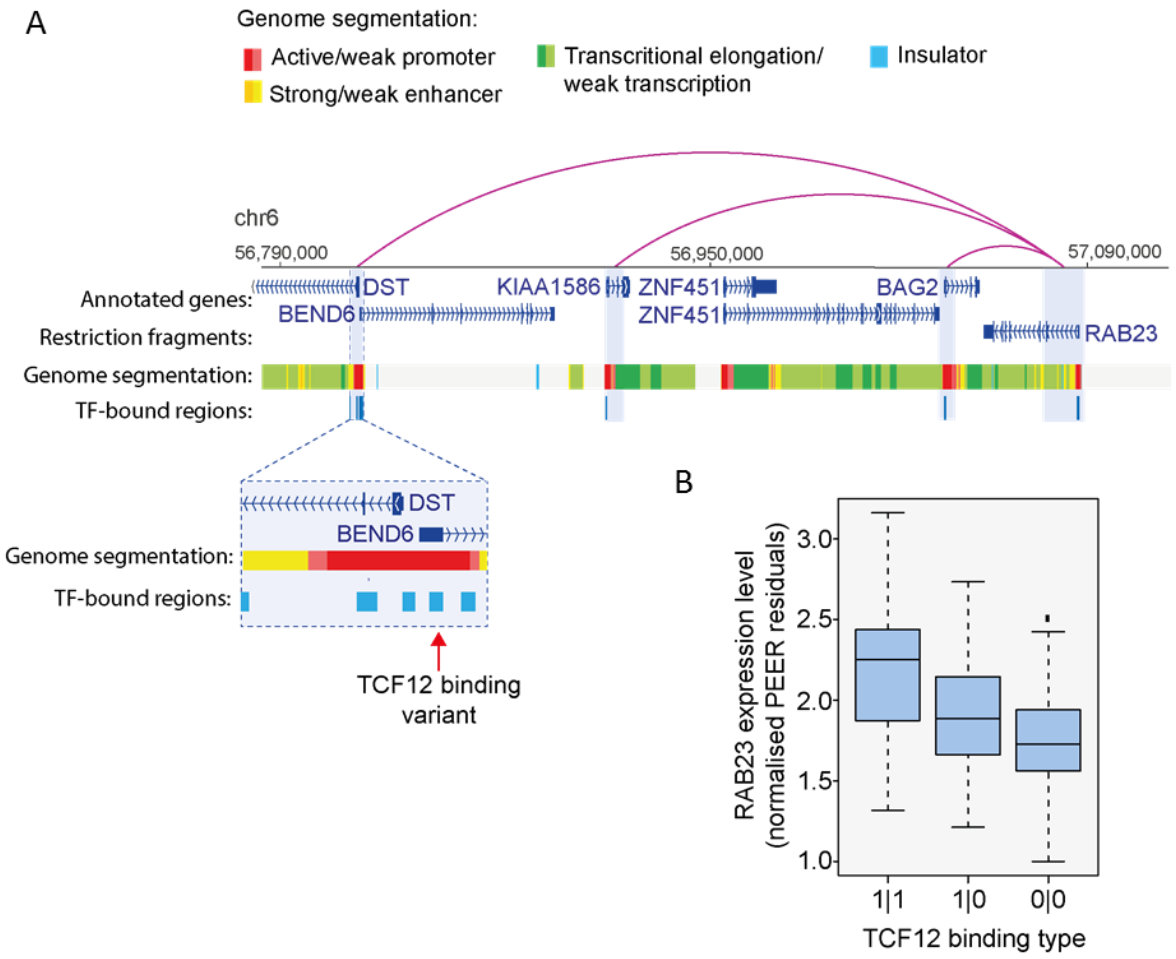


Figure 4.11 A TCF12 binding variant sat at the promoter of BEND6 associates with expression of the distally interacting RAB23 across 359 LCLs

(A) Genome browser representation of the distal interactions detected by promoter capture Hi-C (Mifsud et al., 2015a) for *RAB23*. Three significant distal interactions were detected, all with fragments containing promoters of other genes. The genome segmentation track for GM12878 is shown (Ernst et al., 2011; Pouya Kheradpour et al., 2013). TF bound regions (depicted in dark blue) identified at each fragment using TF ChIP-Seq data from ENCODE (ENCODE Project Consortium, 2012) and proximal *RAB23* TSS window are also shown. One of the TF bound regions at the most distally interacting fragment (266kb from the RAB23 promoter) harbours a predicted TCF12 binding variant. A zoomed-in view of this fragment is shown below the main genome browser view. The TCF12 variant falls at the annotated TSS of *BEND6*; a region defined as an active promoter in genome segmentation of GM12878. (B) The association between RAB23 expression and TCF12 binding variant. The 359 LCLs were split according to their genotype at the predicted TCF12 binding variant (homozygote for the high affinity TCF12 allele, heterozygote for the high affinity TCF12 allele and homozygote for the low affinity TCF12 allele) and their RNA-levels plotted in a boxplot. Publically available RNA-Seq data was used (Lappalainen, Sammeth, Friedländer, 't Hoen, et al., 2013b), where the PEER algorithm had been applied to remove hidden confounding factors, resulting in PEER residuals which were then further quantile normalised. RAB23 expression significantly associates with TCF12 binding type (p-value=4.1×10^{-12}, effect size=-0.19); LCLs with two copies of the TCF12 low affinity allele show the lowest expression.

genes containing these variants at their promoters also associates with the respective variant. Interestingly, for 15% of the genes containing a distal expression-modulating variant, no measurable gene expression was detected by RNA-seq. One example of such a gene is the previously mentioned *BEND6*. *BEND6* contains a predicted TCF12 binding variant at its promoter region which associates with expression of the distal gene *RAB23*, yet *BEND6* expression is undetectable by RNA-Seq (Figure 4.11). Out of the genes which contained a distal expression associated variant at their promoter and for which expression was detected, 60% showed no association between the binding variant and expression. This suggests that, at least for the majority of observed cases where a binding variant sits at the promoter of a gene yet affects expression of another distal gene, the promoter might indeed act in an enhancer-like manner as opposed to affecting the gene expression in trans.

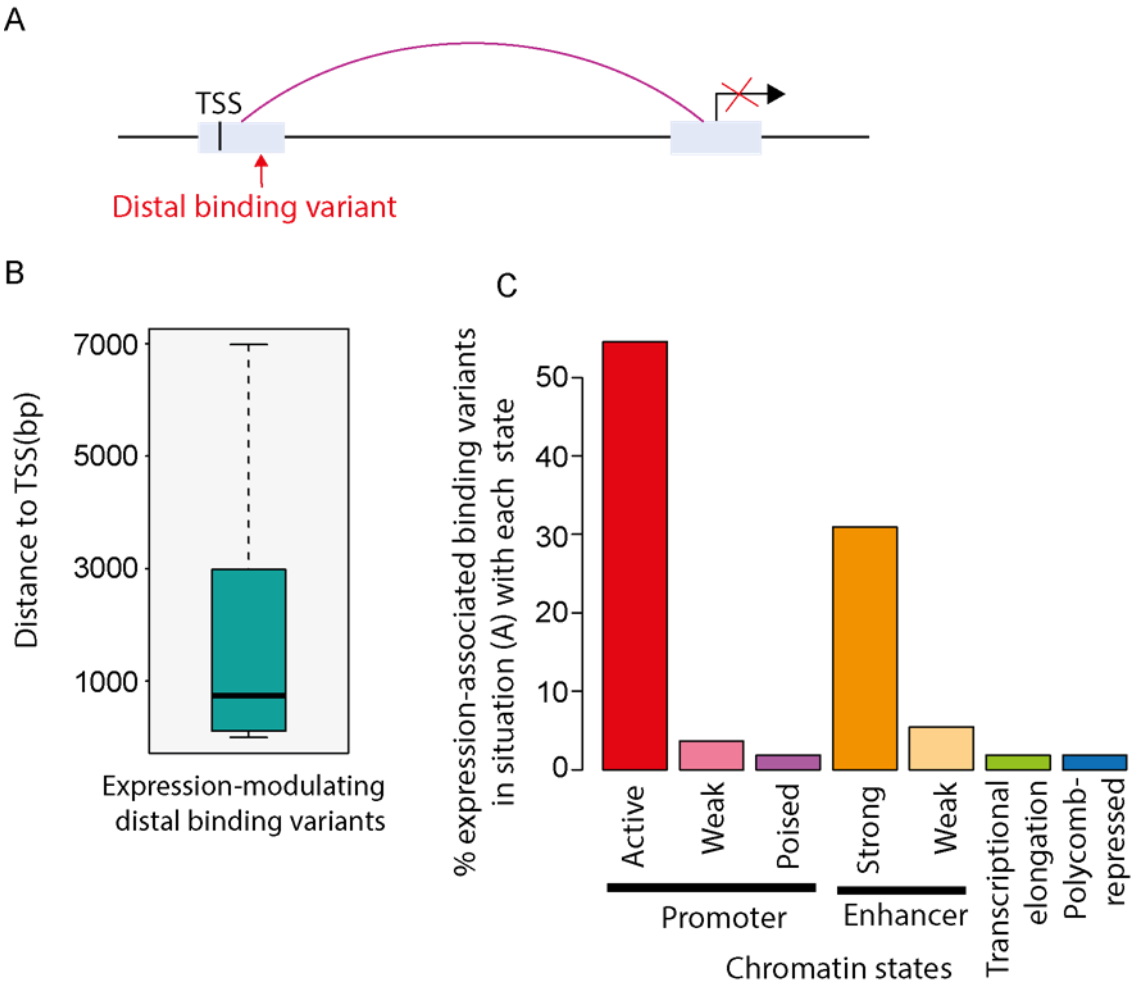


Figure 4.12 Characterisation of variants at promoter-containing fragments that associate with expression of a distal gene

(A) Schematic illustrating the situation where a binding variant associates with expression of a distally interacting gene *and* sits on the same HindIII restriction fragment (used by promoter capture Hi-C to link to other genomic regions) as another promoter. The analyses described in (B) and (C) was carried out on this

specific subset of variants. (B) Boxplot displaying the distances between variants (see (A)) and the closest TSS on the HindIII restriction fragment. (C) Barplot showing the proportion of variants (see (A)) overlapping each genome segmentation category (Ernst et al., 2011; Kheradpour et al., 2013) for GM12878.

I next asked how many predicted binding variants at promoters that contact other genes, associate with expression of the distal gene. Carrying this out revealed that 6% of such variants (mapping to promoter-containing fragments and distally interacting with another gene promoter) significantly associated with expression of the distal gene. Strikingly this is comparable to the proportion of variants sat at non-promoter containing fragments that impact expression of distally interacting genes (7%; typical enhancer variant). This suggests that promoter-promoter interactions involved in such cis-regulation may be widespread.

4.3 Discussion

In this section, I have used a population genetics approach to investigate the effects of changes in TF binding at regulatory regions, in particular at distal elements, on target gene expression using LCLs as a model system. As expected, only a small proportion of predicted TF binding variation in LCLs associated with expression of the target promoter. This is likely due to the known buffering of regulatory variation, especially at distal regions (Cannavò et al., 2016). Nonetheless, hundreds of expression-associated binding variants were identified, the majority of which were not discovered by a previous eQTL analysis using the same panel of LCLs and as such are novel. Strikingly, the majority of predicted TF binding variants showing association with distal gene expression were located within the promoter regions of other genes that physically contacted the target gene. This suggests that some promoters may act as enhancers of other genes (“epromoters”; (Dao et al., 2017)).

4.3.1 Use of epigenomic and interactome data in population-based genetic approaches

Numerous other studies have also utilised epigenomic data in population genetic based approaches to study the effects of sequence variation. Many of these have used epigenomic data post-eQTL/GWAS analysis to collectively examine the properties of expression/trait-associated, with the aim to infer global causal mechanisms. For example, examining the enrichment of eQTLs/GWAS-

loci at TF binding peaks, histone modifications and promoter-interacting regions as detected by high resolution chromosome capture conformation techniques (e.g. Blauwendraat et al., 2016; Javierre et al., 2016; Lappalainen et al., 2013). Epigenomic data has also been used to fine-map the causal variant. Due to LD, a large number of variants will associate with a given trait/expression of a certain gene and the strongest association is not always the causal one. This has led to efforts to identify the casual variants within LD blocks, which amongst other methods includes integrating epigenomic data.

The use of epigenomic data in fine-mapping eQTLs and GWAS loci follows a similar rationale to that of this approach: variants that lie in regions harbouring regulatory marks such as DNase HS, promoter/enhancer-associated histone marks and TF binding peaks are assumed more likely to be the causal variant. The epigenomic data can be incorporated post GWAS/eQTL analysis to identify causal eQTL/GWAS loci amongst those in LD. Several tools have now been developed that enable the large-scale systematic annotation of millions of variants with both epigenomic data, and also binding variant predictions (McLaren et al., 2016; Wang, Li, & Hakonarson, 2010). Recently several studies have used Promoter Capture Hi-C to identify variants at distally interacting fragments (Jäger et al., 2015; Javierre et al., 2016; Martin et al., 2015). Alternatively, functional annotations can be incorporated before association testing, as priors in Bayesian fine-mapping approaches; this effectively up-weights variants with regulatory annotations (e.g. Kichaev et al., 2014; Wen et al., 2015). The approach taken in this section thus represents a novel way to incorporate epigenomic data; like the Bayesian fine-mapping strategies it is incorporated before association testing, but differs in that annotations are used to prioritise a reduced set of predicted functional variants (as opposed to as priors). As a result, this approach offers increased power due a reduced multiple testing burden. Perhaps most significantly though, this approach is the first global eQTL-type analysis to integrate both Promoter Capture Hi-C and TF affinity predictions.

4.3.2 Many known eQTLs are not prioritised in this approach

While I was able to identify novel distal and proximal associations due to increased power, compared to the standard eQTL approach far fewer total associations were found. In fact, the vast majority of GEUVADIS eQTLs were not detected in this study. The eQTLs identified uniquely by GEUVADIS were not in LD with any prioritised variants, suggesting a large number of non-prioritised variants affect expression. This begs the question: through what mechanism might these variants impact gene expression?

A significant number of such variants may still impact expression through altering TF binding, but could not be detected as TF binding variants and thus not tested for association. To predict TF binding variants, I first required the TF to be detected as bound in the LCL GM12878 by ChIP-seq. As such TF variant predictions were limited to TFs for which ChIP-Seq data and PWMs were available– a total of 41 TFs. While the total number of TFs expressed in LCLs are unknown, the number of TFs surveyed here likely represent only a small fraction. For example current estimates for the total number of human TFs are ~1900 (Messina et al., 2004; Vaquerizas et al., 2009). It is thus plausible that eQTLs detected only by GEUVADIS impact binding of TFs not included in this study.

It is also possible that there are many cases where the impact of a variant falling at the binding site of a TF, which was included in this study, was not assessed. Some of these instances may be due to binding of the given TF in GM12878 not being detected by ChIP-seq. Also any variant which perturbs binding at both alleles in the LCL GM12878 was not be analysed – due to requiring evidence of binding in this individual to restrict the false positive rate of PWM binding predictions. Another possibility is that a variant alters binding at a low affinity site, which due to requiring a PWM score over a certain threshold in GM12878, will not be detected. There is emerging evidence to suggest that low affinity TF binding at enhancers may be functionally important (Burgess, 2016; Crocker, Preger-Ben Noon, & Stern, 2016). For example several studies have demonstrated that mutating a low affinity site to a high affinity one at an enhancer results in ectopic expression of the target gene. How widespread the use of low affinity TF binding at enhancers is remains to be seen. For example Wang et al. found that one TF binds and regulates genes containing both high and low affinity binding sites at regulatory regions (Wang et al., 2015). However a study in Drosophila embryos suggested that a large proportion of such low affinity TF binding might be non-functional (Li et al., 2008). If some low affinity binding is functional it also presents the intriguing possibility that, when TF affinity variants associate with expression, perhaps the high affinity allele is the “deleterious” one?

Out of all TF bound regions, only a tiny fraction were predicted to show TF binding variation across this panel of LCLs. This fits with findings from recent studies that suggest only a minority of variants at motifs disrupt binding (Cavalli et al., 2016; Kilpinen et al., 2013b; Maurano et al., 2012c; Spivakov et al., 2012; Tehranchi et al., 2016b). For example one such study found between 0.5 % and 3% of TF bound variants resulted in changes in binding as detected by ChIP-Seq (Tehranchi et al., 2016b). However there is growing evidence to suggest that a significant proportion of variation in TF binding does not result from disruption of known motifs (Gallone et al., 2017; Tehranchi et al., 2016; Wong

et al., 2016). There is currently much work being done to learn additional sequences features, both within and outside of the core site, that influence TF binding (Dror et al. , 2015; Levo et al., 2015).

Another possibility is that eQTLs identified by GEUAVDIS impact expression through mechanisms other than TF binding. A large number of histone modification QTLs and DNA methylation QTLs have been identified, which are often associated with eQTLs (Alasoo et al., 2017; Banovich et al., 2014b; Chen et al., 2016; Degner et al., 2012; Kasowski et al., 2013; Kilpinen et al., 2013; McVicker et al., 2013). However, recent studies suggest these chromatin-QTLs may be primarily driven by changes in TF binding, and as such most expression-associated changes should be captured through examining changes in TF binding (Kilpinen et al., 2013; Lee et al., 2015). Alternatively, variants may impact expression via posttranscriptional mechanisms, for example, through affecting rate of RNA-decay and miRNA binding. RNA decay QTLs, as well as miRNA binding QTLs are often associated with eQTLs, and as such might underlie variation in expression (Lu & Clark, 2012; Pai et al., 2012; Wang et al., 2009). In LCLs it was estimated that 19% of eQTLs might be driven by differences in rate of mRNA decay (Pai et al., 2012). The impact of variants on post-transcriptional mechanisms has tended to receive less attention, in part due to challenges with assaying post-transcriptional mechanisms on a large scale.

4.3.3 Expression is robust to regulatory variation

The finding that only a small fraction of predicted binding variants associate with changes in target gene expression, is consistent with the notion that regulatory variation is often buffered (does not impact gene expression) (Spivakov, 2014). Buffering can occur at the level of TF binding, whereby a TF is still recruited despite disruption of its binding motif. For example, cooperative binding may enable a TF to be recruited via TF-TF interactions (Spivakov et al., 2012). Alternatively, other TFs at the regulatory element may act to “buffer” the loss of a certain TF by maintaining the activity of the regulatory element. The fact that many experimental TF binding perturbations do not result in changes in gene expression support this notion (Cusanovich et al., 2014; Drewell, 2011; Spivakov et al., 2012). It is likely that the identity of the TF whose binding is perturbed affects the impact on the regulatory activity. For example, Tehranchi et al (2016) found that SNPs falling at CTCF motifs were associated with changes in binding of five other TFs, as assayed by pooled ChIP-seq. This suggests that CTCF may be a “pioneer” TF, potentially altering the chromatin environment to permit the binding of other TFs (Tehranchi et al., 2016). Binding variation at such pioneer TFs is more likely to disrupt the regulatory activity of the element, and impact expression. In addition, many

enhancers have been shown to act redundantly, with gene expression unaffected by removal of an individual enhancer under normal conditions (Frankel et al., 2010; Lam et al., 2015; Perry et al., 2010; N. Xiong et al., 2002). As such, the activity of an entire regulatory element can be buffered by the presence of another redundant enhancer. This enhancer level buffering may explain why a smaller proportion of distal variation associated with target gene expression than proximal variants.

The approach taken in this study provides a unique opportunity to test the effect of other features of regulatory architecture on buffering TF binding variation, potentially identifying novel features that may contribute to robustness to genetic perturbations. This is due to having distinct sets of binding variants (for which the identity of the perturbed TF is inferred) which do and do not associate with target gene expression, and importantly for which the regulatory architecture of the target gene is known. In particular the other TFs binding at the regulatory element harbouring the variation are known, as well as those binding other regulatory elements of the loci. A regression approach can be used to learn features of regulatory architecture or the affected TFs that increase the likelihood that expression will be impacted. Alternatively the effects of multiple variants on gene expression across multiple genes can be analysed jointly based on their shared properties (pooled by the properties of the target genes or affected TFs), which has the advantage of further increasing the sensitivity of such analyses.

4.3.4 Epromoters

My finding of promoters acting as enhancers for other genes is consistent with two very recent studies. The first of them used a high throughout reporter-based assay to initially assess enhancer activity of all promoters in human coding genes; they found that 2%-3% of all promoters have enhancer activity, which they termed “epromoters” (Dao et al., 2017). They then used CRISPR/Cas9 to delete several of these epromoters and demonstrate that they are involved in *cis*-regulation of distal gene expression in-vivo, and as such function as true enhancers (Dao et al., 2017). The second study carried out a high-throughput CRISPR/Cas-mediated-mediated mutagenesis screen around the *POU5F1* locus to identify sequences with enhancer function (Diao et al., 2017). They found 40% of the identified cis-regulatory sequences contained annotated promoters of other genes, and these formed spatial contact with the *POU5F1* promoter, analogous to enhancers (Diao et al., 2017). Whilst these two studies were the first to demonstrate the function of epromoters in vivo, as well as giving as indication of their prevalence, they followed on from an accumulating body of

evidence suggesting that promoters may be able to function as enhancers. Emerging similarities between promoter and enhancers, for example chromatin marks and bidirectional transcription, led to the idea the promoter-enhancer distinction may not be as clear as initially thought (Andersson, 2015; Core et al., 2014; Kim & Shiekhattar, 2015). In addition the finding that promoters are frequently engaged in interactions with other promoters, suggested that they may have a regulatory role (Li et al., 2012a; Sanyal et al., 2012; Schoenfelder et al., 2015). Previous reporter assays have also found that promoters were able to function as enhancers (Arnold et al., 2013; G. Li et al., 2012; Nguyen et al., 2016; Zabidi et al., 2014). The work in this section has thus added to a growing body of evidence supporting an enhancer-like role for promoters, and suggested that this type of regulation may be common. Significantly, it also provides a catalogue of potential epromoters for which, unlike those identified in reporter assays, the distal target gene has been identified. This may be of use in the further characterisation of epromoter properties.

What might be the role of such epromoters? The dual promoter-enhancer function may ensure co-ordinated regulation of the gene associated with the epromoter and the distally interacting gene. This has similarities to enhancer sharing, which is widespread and has been demonstrated to play a role in co-regulation (Jin et al., 2013; Schoenfelder et al., 2015). Dao et al. found that expression of the proximal and distal epromoter associated genes was highly correlated, however the same was seen for all physically interacting genes. Indeed several studies have shown that genes in close spatial proximity tend to show co-regulation (Li et al., 2012). Epromoters might represent one mechanism underlying the co-regulation of physically associated genes, along with enhancer-sharing and common TF environment. If this is indeed the case, what unique properties might epromoters confer to coordinated regulation over for example enhancer sharing?

Dao et al. suggested that epromoters may play a role in coordinating rapid responses to external stimuli. This followed from the observation that a significant number of epromoter associated genes were key interferon response genes. Indeed they were able to demonstrate for two loci that epromoters were necessary to activate distal expression in response to interferon. A similar finding was observed by Li et al., who found that loss of ERα binding at one promoter impacted expression of physically interacting genes, which do not bind ERα at their promoters (G. Li et al., 2012). The ability to activate expression in response to external stimuli is a well-characterised feature of enhancers, thus enhancer-sharing might also facilitate co-ordinated responses to stimuli. It is tempting to speculate epromoters versus enhancer-sharing might offer slightly different properties for such co-ordinated regulation. For example, a recent study found that the closer a given enhancer is to a gene, the more frequent the transcriptional bursts are it induces (Fukaya, Lim, & Levine, 2016). It is plausible that transcriptional bursts will be more frequent at the gene proximal

to the epromoter than the distal gene, whereas a shared enhancer may induce similar burst frequencies at both genes. Other key questions include whether the looping dynamics are similar to traditional enhancer-promoter contacts. In addition, what is the role of epromoters in transcription factories? Future studies will likely start to address these questions.

Whilst there is initial evidence to support a role of epromoters in co-ordinated regulation, the finding that 15% of epromoter associated genes show no detectable expression suggests that this is not the case for all epromoters. This is consistent with Dao et al, who found between 14% and 41% of epromoter associated genes did not have an active TSS in the given cell line (Dao et al., 2017). They hypothesized that there are two distinct types of epromoters: one that co-ordinately regulates the proximal and distal gene, and another that can function as either a promoter or enhancer, thus regulating the proximal and distal genes in different cell types. Regulatory elements that appear to be able to function as either promoters or enhancers have been previously identified in a human cross-tissue/cell type study. Leung et al. observed that ~15% of strong promoters were predicted enhancers in other tissues/cell types (Leung et al., 2015). Further these regulatory elements were able to function according to their predicted role (enhancer or promoter) in a reporter assay. Similarly, a study in mice found that intragenic enhancers were able to act as alternative tissue-specific promoters (Kowalczyk et al., 2012). If the epromoters can indeed act as an enhancer in one cell type and a promoter in another, this presents an intriguing question. How does the gene proximal to the active enhancer, which is capable of acting as a promoter in another cell type, avoid activation? Perhaps insulator type elements prevent activation of the proximal gene. Examining properties of expressed versus non-expressed genes that harbour a distal expression-modulating variant at their promoter region, might lead to insights on how activation is avoided.

One perplexing finding is that in the majority of cases where a variant at an epromoter impacts expression of the distal gene, expression of the proximal gene remains unperturbed. It is theoretically possible that expression of the proximal gene is also impacted, but not detected. However given that epromoters appear to be highly expressed, which increases power to detect changes, it seems unlikely that undetected associations explain all such instances. Furthermore, the allele frequency, which also influences power to detect associations, is identical in both the proximal and distal association tests as they are testing for association with the same variant. Another possibility is that the proximal gene may be under the control of an alternative promoter, as was discussed previously. However if the finding from Dao et al., that only ~30% of epromoter proximal genes have alternative TSSs, holds for this set of LCL epromoters, this would not explain all instances where expression of the epromoter proximal gene remains unaffected. If the

epromoter is indeed functioning simultaneously as a promoter and enhancer, an intriguing question is why only expression of the distal gene appears is impacted.

One possible explanation for these findings is that change in TF binding at the epromoter results in a loss of physical interaction between the epromoter and distal gene. In this case activation of the distal gene by the epromoter would be completely lost, whereas expression of the proximal gene might be only slightly impacted due to loss of binding of the given TF or robust to this change. There are several examples of mutations at classical enhancers that disrupt looping, and consequently target gene expression (e.g. Majumder et al. 2008; Visser, Kayser, & Palstra, 2012). For example, Visser et al. (2012) demonstrated that a mutation at one allele that reduced binding of the TF HLTF, resulted in allele-specific reduction in looping to the pigment gene OCA2, accompanied by allele-specific loss of expression. Loss of CTCF binding has also been shown to disrupt looping; a CTCF knockdown reduced physical interactions between an enhancer normally bound by CTCF, and the promoters of *HLA-DRB1* and *HLA-DQA1* (Majumder et al., 2008). The reduction in enhancer interactions led to a reduction in expression of the two contacted genes (Majumder et al., 2008). This hypothesis could be tested using putative epromoter variants identified in this study that are heterozygote in GM12878, by re-analysing the promoter capture Hi-C data to test for allele-specific looping interactions with the distal impacted gene. Further, the impacted TFs can be compared between epromoters that do and do not associate with proximal gene expression, with the hypothesis that those impact only distal expression may be factors involved in looping.

5 General Discussion

In this thesis I have used a combination of functional genomic and population genetics approaches to interrogate principles of enhancer regulation. In Chapter 3 I found that genes regulated by multiple enhancers favour a "shadow" enhancer architecture, whereby their enhancers recruit similar sets of TFs. This provided the first global insight into the principles of TF binding at the enhancers concurrently regulating the same gene. In Chapter 4 I predicted hundreds of TF binding variants at distal regulatory elements, a small proportion of which were found to associate with target gene expression. Robustness observed by the "shadow" enhancers observed in Chapter 3 is likely one reason why most TF binding variants do not associate with gene expression changes under normal conditions. It will be interesting to see whether shadow enhancers also confer robustness under stress conditions, as has been observed in Drosophila (Frankel et al., 2010; Perry et al., 2010). Strikingly a large proportion of variants that impacted expression fell at the promoters of other genes, suggesting that promoters may be able to act in an enhancer-like manner (epromoters), as observed by two very recent studies (Dao et al., 2017; Diao et al., 2017). Significantly, with the identification of promoter variants that impact distal gene expression, this work has added to a small catalogue of predicted epromoters. These will likely be valuable in further elucidating the properties and biological significance of epromoters. In addition, the identification of sets of predicting binding variants at enhancers that do and do not associate with target gene expression provides a unique opportunity to learn properties of general regulatory architecture that confer robustness. This work has thus set the scene for many exciting follow-up studies.

Population genomics approaches such as that used in Chapter 4 are likely to become an increasingly powerful way to study principles of enhancer regulation. As next-generation sequencing costs reduce, we are able to sequence the genomes and transcriptomes of an increasingly large numbers of individuals. For example, the recently launched 100,000 Genomes project aims to sequence the genomes of 70,000 cancer and rare-disease patients (Caulfield et al., 2017). Samples from each patient are being kept to enable RNA-Seq and epigenomic assays to be performed in the future (Caulfield et al., 2017). With increasing sample sizes and accompanying epigenomic information, population genomic approaches will be able to detect variants with weaker effects, complex interactions between variants as well as better detection of causal variants. These are likely to be accompanied by novel statistical techniques and approaches for incorporating epigenomic data. In addition, the increasing bulk of sequencing data may enable the application of deep learning algorithms to population genomic approaches. Deep learning algorithms have already been used

to predict molecular traits, including chromatin profiles and expression, from sequence (Kelley, Snoek, & Rinn, 2016; Leung et al., 2014; Xiong et al., 2015; Zhou & Troyanskaya, 2015). In these instances, the models were trained upon sequence features within a single genome. It could be envisioned that variation between individuals could also be used, for example, to train a model to predict when and when not variation at regulatory regions associates with changes in gene expression. Such algorithms may be particularly suited to learning complex features of regulatory architecture that confer robustness, as they cope well with non-linear dependencies and interaction effects that are likely to exist. With a combination of statistical and deep learning approaches, it is likely that the full potential of population genomics approaches for elucidating principles of enhancer function will be realised.

Insights and findings from population genomic approaches, including those in this study, however still require experimental validation due to correlative nature of associations. Until very recently it was both challenging and time-consuming to directly perturb enhancers in their native context. Reporter assays, while easy to carry out, do not capture the genomic context within which the enhancer resides. For example, they may not acquire chromatin marks, and also enhancer-promoter looping as well as the 3D genome environment is lacking. This is particularly relevant for validating and investigating the function of epromoters identified in this work, where 3D interactions are likely to play a key role. Recent developments in genome editing techniques, in particular the advent of the CRISPR-cas9 system (Cong et al., 2013; Mali et al., 2013), now enable perturbations to be introduced at endogenous loci relatively quickly and simply. Cas9 can be targeted to almost all genomic loci by sequence-specific guide RNAs to induce double stranded RNA breaks. These are repaired by non-homologous end joining which results in small insertions or deletions. In addition, cas9 can be used to produce targeted deletions through the use of two guide RNAs flanking the target region. Significantly inactivated cas9 can still be targeted specifically to DNA (Dominguez, Lim, & Qi, 2015). As such it can fused to either an activator, repressor or epigenetic regulator to directly modulate the activity or epigenetic state of a targeted enhancer (Dominguez et al., 2015; Lopes, Korkmaz, & Agami, 2016). Although the off-target effects still need to be investigated, as it is possible that the activator/repressor impacts promoter activity directly. Both the genome-editing and activity-modulating versions of CRISPR-cas9 offer a powerful way to validate and further explore the functional characteristics of predicted epromoters identified in this work.

In this work, I have used Promoter Capture Hi-C to link enhancers to target genes. This has a resolution of 4kb on average, likely larger than most regulatory regions. In addition, it has a “blind” window, (approximately three restriction fragment in length) around the centre of the baited

fragment, which limits the detection of short-range enhancers. CRISPR-cas9, due to it's amenability to high-throughput approaches, can be used to systematically interrogate sequences for regulatory potential (Lopes et al., 2016). Several studies have taken advantage of this, tiling hundreds of guide RNAs across regulatory regions for a handful of genes. This may be used, alongside Promoter Capture Hi-C, to pinpoint the regulatory regions at distal interacting regions, as well as to identify short-range enhancers. One caveat however is that redundant enhancers (at least under the conditions tested) will not be detected in a CRISPR-cas9 screen. Given the apparent pervasiveness of shadow enhancers, a large number of enhancers may act redundantly in humans. In addition, while many enhancers can be screened in parallel using high throughput CRISPR-cas9 approaches, the normally used bulk read-outs mean that it is limited to assaying the effects on a single gene.

Excitingly several recent studies have overcome this limitation by combining CRISPR-cas9 perturbations with high-throughput single cell sequencing, enabling global changes in gene expression to be measured simultaneously (Adamson et al., 2016; Dixit et al., 2016; Jaitin et al., 2016; Xie et al., 2017). While the number of enhancers that can currently be screened with this combined approach is limited due to sequencing costs (Xie et al., 2017), with future reductions in sequencing costs this approach may enable genome-wide functional analysis of enhancers. Even the ability to assay the effects on a handful of genes simultaneously might be of use for investigating the role of epromoters in TF factories, where their effect on expression of multiple genes could be investigated. Xie et al. (2017) demonstrated that a single cell repressive CRISPR-cas9 approach can also be used to investigate the combinatorial action of enhancers. They were able to show that several enhancers with a negligible effect on gene expression individually, elicited a big effect on expression in combination, suggestive of redundancy (Xie et al., 2017). If redundant enhancers are indeed pervasive in the human genome, this may represent an important strategy for identifying regulatory elements, as well as investigating multi-enhancer logic. In addition, the single-cell nature of the read out also offers a unique opportunity to investigate cell-to-cell variability in enhancer function.

In conclusion the work in this thesis has shed further light on principles of enhancer regulation. Excitingly recent advances in experimental techniques now enable these findings and biological insights to be tested experimentally.

6 Appendix

6.1 TFs used in analysis

ATF3
BATF
BCL11A**
BCL3**
BCLAF1**
BRCA1
CHD2
CTCF*
EBF1
EGR1
ELF1
ETS1
FAM48A**
FOS**
GABPA
IRF3
JUND**
KAT2A**
MAX
MEF2A
MEF2C**
MYC
NFE2
NFKB1
NR2C2
NRF1
PAX5
PBX3
POU2F2
RAD21*
REST
RFX5
RXRA
SIN3A
SIX5
SMC3*
SP1
SPI1
SRF
STAT1
STAT3
TAF1**

TBP
TCF12
USF1
USF2
WRNIP1**
YY1
ZBTB33
ZEB1
ZNF143
ZZZ3**

* TFs that were excluded from the analysis in Section 3.

**TFs for which PWMs were not available for. These were therefore not included in the TF affinity analysis part of Section 4 (but were included in the initial TF binding annotation part of Section 4).

6.2 List of abbreviations

AIC	Akaike information criterion
ANOVA	Analysis of variance
ATAC-seq	Assay for Transposase-Accessible Chromatin using sequencing
bp	Base pairs
ChIA-PET	Chromatin Interaction Analysis by Paired-End Tag Sequencing
CHiC	Capture Hi-C
ChIP	Chromatin immunoprecipitation
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
CRM	Cis-regulatory module
DNase	Deoxyribonuclease
DNase HS	Dnase hypersensitivity
dsQTL	DNase I sensitivity quantative trait loci
ENCODE	Encyclopedia of DNA Elements
eQTL	expression quantative trait loci
eRNA	Enhancer RNA
FDR	False discovery rate
FISH	Fluorescent in situ hybridisation
GO	Gene ontology
GTF	General transcription factor
GWAS	Genome wide association study
H3K27ac	Histone H3 acetylation at lysine 27
H3K4me1	Histone H3 mono-methylation at lysine 4
HAT	Histone acetly transferase
IFN	Interferon
Ig	Immunoglobulin
Indel	Insertion/deletion
kb	Kilobases
LCL	Lymphoblastoid cell line

LD	Linkage disequilibrium
miRNA	MicroRNA
MPRA	Massively paralell reporter assay
PCHiC	Promoter capture Hi-C
PCR	Polymerse chain reaction
PEER	Probabilistic Estimation of Expression Residuals
PIC	Preinitiation complex
PIR	Promoter interacting region
PWM	Postion weight matrix
RNA pol	RNA polymerase
Shh	Sonic hedgehog
SNP	Single nucleotide polymorphism
TF	Transcription factor
TSS	Transcription start site

7 References

Adamson, B., Norman, T. M., Jost, M., Parnas, O., Regev, A., Weissman Correspondence, J. S., … Weissman, J. S. (2016). A Multiplexed Single-Cell CRISPR Screening Platform Enables Systematic Dissection of the Unfolded Protein Response. *Cell*, *167*. https://doi.org/10.1016/j.cell.2016.11.048

Adelman, K., & Lis, J. T. (2012). Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nature Reviews. Genetics*, *13*(10), 720–31. https://doi.org/10.1038/nrg3293

Adkins, M. W., & Tyler, J. K. (2006). Transcriptional Activators Are Dispensable for Transcription in the Absence of Spt6-Mediated Chromatin Reassembly of Promoter Regions. *Molecular Cell*, *21*(3), 405–416. https://doi.org/10.1016/j.molcel.2005.12.010

Agalioti, T., Lomvardas, S., Parekh, B., Yie, J., Maniatis, T., & Thanos, D. (2000). Ordered recruitment of chromatin modifying and general transcription factors to the IFN-beta promoter. *Cell*, *103*(4), 667–78. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11106736

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, *19*(6), 716–723. https://doi.org/10.1109/TAC.1974.1100705

Alasoo, K., Rodrigues, J., Mukhopadhyay, S., Knights, A. J., Mann, A. L., Kundu, K., … Gaffney, D. J. (2017). Genetic effects on chromatin accessibility foreshadow gene expression changes in macrophage immune response. *bioRxiv*, 102392. https://doi.org/10.1101/102392

Albert, F. W., & Kruglyak, L. (2015a). The role of regulatory variation in complex traits and disease. *Nature Reviews Genetics*, *16*(4), 197–212. https://doi.org/10.1038/nrg3891

ALLFREY, V. G., FAULKNER, R., & MIRSKY, A. E. (1964). ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. *Proceedings of the National Academy of Sciences of the United States of America*, *51*(5), 786–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/14172992

Amano, T., Sagai, T., Tanabe, H., Mizushina, Y., Nakazawa, H., & Shiroishi, T. (2009). Chromosomal Dynamics at the Shh Locus: Limb Bud-Specific Differential Regulation of Competence and Active Transcription. *Developmental Cell*, *16*(1), 47–57. https://doi.org/10.1016/j.devcel.2008.11.011

Andersen, M. C., Engström, P. G., Lithwick, S., Arenillas, D., Eriksson, P., Lenhard, B., … Odeberg, J. (2008). In Silico Detection of Sequence Variations Modifying Transcriptional Regulation. *PLoS Computational Biology*, *4*(1), e5. https://doi.org/10.1371/journal.pcbi.0040005

Anderson, E., Devenney, P. S., Hill, R. E., & Lettice, L. A. (2014). Mapping the Shh long-range regulatory domain. *Development (Cambridge, England)*, *141*(20), 3934–43. https://doi.org/10.1242/dev.108480

Andersson, R. (2015). Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model. *BioEssays*, *37*(3), 314–323. https://doi.org/10.1002/bies.201400162

Arents, G., Burlingame, R. W., Wang, B. C., Love, W. E., & Moudrianakis, E. N. (1991). The nucleosomal core histone octamer at 3.1 A resolution: a tripartite protein assembly and a left-handed superhelix. *Proceedings of the National Academy of Sciences of the United States of America*, *88*(22), 10148–52. Retrieved from

http://www.ncbi.nlm.nih.gov/pubmed/1946434

Arnold, C. D., Gerlach, D., Stelzer, C., Boryn, L. M., Rath, M., & Stark, A. (2013). Genome-Wide Quantitative Enhancer Activity Maps Identified by STARR-seq. *Science*, *339*(6123), 1074–1077. https://doi.org/10.1126/science.1232542

Arnoult, L., Su, K. F. Y., Manoel, D., Minervino, C., Magriña, J., Gompel, N., … B., P. (2013). Emergence and diversification of fly pigmentation through evolution of a gene regulatory module. *Science (New York, N.Y.)*, *339*(6126), 1423–6. https://doi.org/10.1126/science.1233749

Auton, A., Abecasis, G. R., Altshuler, D. M., Durbin, R. M., Abecasis, G. R., Bentley, D. R., … Abecasis, G. R. (2015a). A global reference for human genetic variation. *Nature*, *526*(7571), 68–74. https://doi.org/10.1038/nature15393

Banerji, J., Olson, L., & Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. *Cell*, *33*(3), 729–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6409418

Banerji, J., Rusconi, S., & Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. *Cell*, *27*(2 Pt 1), 299–308. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/6277502

Bannister, A. J., & Kouzarides, T. (1996). The CBP co-activator is a histone acetyltransferase. *Nature*, *384*(6610), 641–643. https://doi.org/10.1038/384641a0

Bannister, A. J., & Kouzarides, T. (2011). Regulation of chromatin by histone modifications. *Cell Research*, *21*(3), 381–95. https://doi.org/10.1038/cr.2011.22

Banovich, N. E., Lan, X., McVicker, G., van de Geijn, B., Degner, J. F., Blischak, J. D., … Gilad, Y. (2014a). Methylation QTLs Are Associated with Coordinated Changes in Transcription Factor Binding, Histone Modifications, and Gene Expression Levels. *PLoS Genetics*. https://doi.org/10.1371/journal.pgen.1004663

Barolo, S. (2012). Shadow enhancers: Frequently asked questions about distributed cis-regulatory information and enhancer redundancy. *BioEssays*. https://doi.org/10.1002/bies.201100121

Battle, A., Mostafavi, S., Zhu, X., Potash, J. B., Weissman, M. M., McCormick, C., … Koller, D. (2014). Characterizing the genetic basis of transcriptome diversity through RNA-sequencing of 922 individuals. *Genome Research*, *24*(1), 14–24. https://doi.org/10.1101/gr.155192.113

Bauer, D. E., Kamran, S. C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., … Orkin, S. H. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. *Science (New York, N.Y.)*, *342*(6155), 253–7. https://doi.org/10.1126/science.1242088

Beagrie, R. A., & Pombo, A. (2016). Gene activation by metazoan enhancers: Diverse mechanisms stimulate distinct steps of transcription. *BioEssays*, *38*(9), 881–893. https://doi.org/10.1002/bies.201600032

Becker, P. B., & Workman, J. L. (2013). Nucleosome remodeling and epigenetics. *Cold Spring Harbor Perspectives in Biology*, *5*(9). https://doi.org/10.1101/cshperspect.a017905

Bell, M. A. (1987). Interacting evolutionary constraints i n pelvic reduction of threespine sticklebacks, Gasterosteus aculeatus (Pisces, Gasterosteidae). *Biological Journal of the Linnean Society*.

Benabdallah, N. S., Williamson, I., Illingworth, R. S., Boyle, S., Grimes, G. R., Therizols, P., & Bickmore, W. A. (2017). PARP mediated chromatin unfolding is coupled to long- range

enhancer activation. https://doi.org/10.1101/155325

Berg, O. G., & von Hippel, P. H. (1987). Selection of DNA binding sites by regulatory proteins. Statistical-mechanical theory and application to operators and promoters. *Journal of Molecular Biology*, *193*(4), 723–50. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3612791

Berg, O. G., & von Hippel, P. H. (1988). Selection of DNA binding sites by regulatory proteins. *Trends in Biochemical Sciences*, *13*(6), 207–11. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3079537

Berger, S. L. (2007). The complex language of chromatin regulation during transcription. *Nature*, *447*(7143), 407–412. https://doi.org/10.1038/nature05915

Bernstein, B. E., Birney, E., Dunham, I., Green, E. D., Gunter, C., & Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Blackwood, E. M., & Kadonaga, J. T. (1998). Going the Distance: A Current View of Enhancer Action. *Science*, *281*(5373). Retrieved from http://science.sciencemag.org/content/281/5373/60

Blauwendraat, C., Francescatto, M., Gibbs, J. R., Jansen, I. E., Simón-Sánchez, J., Hernandez, D. G., … Heutink, P. (2016). Comprehensive promoter level expression quantitative trait loci analysis of the human frontal lobe. *Genome Medicine*, *8*(1), 65. https://doi.org/10.1186/s13073-016-0320-1

Boeger, H., Griesenbeck, J., Strattan, J. S., & Kornberg, R. D. (2004). Removal of Promoter Nucleosomes by Disassembly Rather Than Sliding In Vivo. *Molecular Cell*, *14*(5), 667–673. https://doi.org/10.1016/j.molcel.2004.05.013

Bothma, J. P., Garcia, H. G., Ng, S., Perry, M. W., Gregor, T., & Levine, M. (2015). Enhancer additivity and non-additivity are determined by enhancer strength in the Drosophila embryo. *eLife*, *4*. https://doi.org/10.7554/eLife.07956

Bowman, G. D., & Poirier, M. G. (2015). Post-translational modifications of histones that influence nucleosome dynamics. *Chemical Reviews*, *115*(6), 2274–95. https://doi.org/10.1021/cr500350x

Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y., & Greenleaf, W. J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nature Methods*, *10*(12), 1213–1218. https://doi.org/10.1038/nmeth.2688

Bulger, M., & Groudine, M. (1999). Looping versus linking: toward a model for long-distance gene activation. *Genes & Development*, *13*(19), 2465–77. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10521391

Bulger, M., & Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. *Cell*, *144*(3), 327–39. https://doi.org/10.1016/j.cell.2011.01.024

Burgess, D. J. (2016). Regulatory elements: Putting enhancers into context. *Nature Reviews Genetics*, *17*(7), 377–377. https://doi.org/10.1038/nrg.2016.74

BURGESS, R. R., TRAVERS, A. A., DUNN, J. J., & BAUTZ, E. K. F. (1969). Factor Stimulating Transcription by RNA Polymerase. *Nature*, *221*(5175), 43–46. https://doi.org/10.1038/221043a0

Butler, J. S., Koutelou, E., Schibler, A. C., & Dent, S. Y. R. (2012). Histone-modifying enzymes: regulators of developmental decisions and drivers of human disease. *Epigenomics*, *4*(2), 163–77. https://doi.org/10.2217/epi.12.3

Cairns, J., Freire-Pritchett, P., Wingett, S. W., Várnai, C., Dimond, A., Plagnol, V., … Spivakov, M. (2016). CHiCAGO: robust detection of DNA looping interactions in Capture Hi-C data. *Genome Biology*, *17*(1), 127. https://doi.org/10.1186/s13059-016-0992-2

Calo, E., & Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? *Molecular Cell*, *49*(5), 825–37. https://doi.org/10.1016/j.molcel.2013.01.038

Cannavò, E., Khoueiry, P., Garfield, D. A., Geeleher, P., Zichner, T., Gustafson, E. H., … Furlong, E. E. M. (2016). Shadow Enhancers Are Pervasive Features of Developmental Regulatory Networks. *Current Biology*. https://doi.org/10.1016/j.cub.2015.11.034

Canver, M. C., Smith, E. C., Sher, F., Pinello, L., Sanjana, N. E., Shalem, O., … Bauer, D. E. (2015). BCL11A enhancer dissection by Cas9-mediated in situ saturating mutagenesis. *Nature*, *527*(7577), 192–197. https://doi.org/10.1038/nature15521

Caroline Bartman, A. R., Hsu, S. C., C-S Hsiung, C., Raj, A., Blobel, G. A., & Bartman, C. R. (2016). Enhancer Regulation of Transcriptional Bursting Parameters Revealed by Forced Chromatin Looping. *Molecular Cell*, *62*, 237–247. https://doi.org/10.1016/j.molcel.2016.03.007

Carroll, S. B. (2008). Evo-Devo and an Expanding Evolutionary Synthesis: A Genetic Theory of Morphological Evolution. *Cell*, *134*(1), 25–36. https://doi.org/10.1016/J.CELL.2008.06.030

Carter, D., Chakalova, L., Osborne, C. S., Dai, Y., & Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. *Nature Genetics*, *32*(4), 623–626. https://doi.org/10.1038/ng1051

Caulfield, M., Davies, J., Dennys, M., Elbahy, L., Fowler, T., Hill, S., … Woods, K. (2017). The 100,000 Genomes Project Protocol. https://doi.org/10.6084/M9.FIGSHARE.4530893.V2

Cavalli, M., Pan, G., Nord, H., Wallerman, O., Wallén Arzt, E., Berggren, O., … Wadelius, C. (2016). Allele-specific transcription factor binding to common and rare variants associated with disease and gene expression. *Human Genetics*, *135*(5), 485–97. https://doi.org/10.1007/s00439-016-1654-x

Chan, Y. F., Marks, M. E., Jones, F. C., Villarreal, G., Shapiro, M. D., Brady, S. D., … Kingsley, D. M. (2010). Adaptive evolution of pelvic reduction in sticklebacks by recurrent deletion of a Pitx1 enhancer. *Science (New York, N.Y.)*, *327*(5963), 302–5. https://doi.org/10.1126/science.1182213

Chang, H. H., Hemberg, M., Barahona, M., Ingber, D. E., & Huang, S. (2008). Transcriptome-wide noise controls lineage choice in mammalian progenitor cells. *Nature*, *453*(7194), 544–7. https://doi.org/10.1038/nature06965

Chen, L., Ge, B., Casale, F. P., Vasquez, L., Kwan, T., Garrido-Martín, D., … Soranzo, N. (2016). Genetic Drivers of Epigenetic and Transcriptional Variation in Human Immune Cells. *Cell*, *167*(5), 1398–1414.e24. https://doi.org/10.1016/j.cell.2016.10.026

Cho, K. W. Y. (2012). Enhancers. *Wiley Interdisciplinary Reviews. Developmental Biology*, *1*(4), 469–78. https://doi.org/10.1002/wdev.53

Clapier, C. R., Iwasa, J., Cairns, B. R., & Peterson, C. L. (2017). Mechanisms of action and regulation of ATP-dependent chromatin-remodelling complexes. *Nature Reviews Molecular Cell Biology*, *18*(7), 407–422. https://doi.org/10.1038/nrm.2017.26

Cong, L., Ran, F. A., Cox, D., Lin, S., Barretto, R., Habib, N., … Zhang, F. (2013). Multiplex Genome Engineering Using CRISPR/Cas Systems. *Science*, *339*(6121), 819–823. https://doi.org/10.1126/science.1231143

Core, L. J., Martins, A. L., Danko, C. G., Waters, C. T., Siepel, A., & Lis, J. T. (2014). Analysis of nascent RNA identifies a unified architecture of initiation regions at mammalian promoters and enhancers. *Nature Genetics*, *46*(12), 1311–1320. https://doi.org/10.1038/ng.3142

Corradin, O., Saiakhova, A., Akhtar-Zaidi, B., Myeroff, L., Willis, J., Cowper-Sal lari, R., … Scacheri, P. C. (2014). Combinatorial effects of multiple enhancer variants in linkage disequilibrium dictate levels of gene expression to confer susceptibility to common traits. *Genome Research*, *24*(1), 1–13. https://doi.org/10.1101/gr.164079.113

Corradin, O., & Scacheri, P. C. (2014). Enhancer variants: evaluating functions in common disease. *Genome Medicine*, *6*(10), 85. https://doi.org/10.1186/s13073-014-0085-3

Cowper-Sal·lari, R., Zhang, X., Wright, J. B., Bailey, S. D., Cole, M. D., Eeckhoute, J., … Lupien, M. (2012). Breast cancer risk–associated SNPs modulate the affinity of chromatin for FOXA1 and alter gene expression. *Nature Genetics*, *44*(11), 1191–1198. https://doi.org/10.1038/ng.2416

Crocker, J., Preger-Ben Noon, E., & Stern, D. L. (2016). The Soft Touch. In *Current topics in developmental biology* (Vol. 117, pp. 455–469). https://doi.org/10.1016/bs.ctdb.2015.11.018

Cusanovich, D. A., Pavlovic, B., Pritchard, J. K., Gilad, Y., & Thurman, R. (2014). The Functional Consequences of Variation in Transcription Factor Binding. *PLoS Genetics*, *10*(3), e1004226. https://doi.org/10.1371/journal.pgen.1004226

Danecek, P., Auton, A., Abecasis, G., Albers, C. A., Banks, E., DePristo, M. A., … Durbin, R. (2011). The variant call format and VCFtools. *Bioinformatics*, *27*(15), 2156–2158. https://doi.org/10.1093/bioinformatics/btr330

Daniel, B., Nagy, G., & Nagy, L. (2014). The intriguing complexities of mammalian gene regulation: How to link enhancers to regulated genes. Are we there yet? *FEBS Letters*, *588*(15), 2379–2391. https://doi.org/10.1016/j.febslet.2014.05.041

Dao, L. T. M., Galindo-Albarrán, A. O., Castro-Mondragon, J. A., Andrieu-Soler, C., Medina-Rivera, A., Souaid, C., … Spicuglia, S. (2017). Genome-wide characterization of mammalian promoters with distal enhancer functions. *Nature Genetics*, *49*(7), 1073–1081. https://doi.org/10.1038/ng.3884

De Santa, F., Barozzi, I., Mietton, F., Ghisletti, S., Polletti, S., Tusi, B. K., … Natoli, G. (2010). A Large Fraction of Extragenic RNA Pol II Transcription Sites Overlap Enhancers. *PLoS Biology*, *8*(5), e1000384. https://doi.org/10.1371/journal.pbio.1000384

Degenhardt, K. R., Milewski, R. C., Padmanabhan, A., Miller, M., Singh, M. K., Lang, D., … Epstein, J. A. (2010). Distinct enhancers at the Pax3 locus can function redundantly to regulate neural tube and neural crest expressions. *Developmental Biology*, *339*(2), 519–527. https://doi.org/10.1016/j.ydbio.2009.12.030

Degner, J. F., Pai, A. A., Pique-Regi, R., Veyrieras, J.-B., Gaffney, D. J., Pickrell, J. K., … Pritchard, J. K. (2012). DNase I sensitivity QTLs are a major determinant of human expression variation. *Nature*, *482*(7385), 390–394. https://doi.org/10.1038/nature10808

Dekker, J., Marti-Renom, M. A., & Mirny, L. A. (2013). Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nature Reviews Genetics*,

14(6), 390–403. https://doi.org/10.1038/nrg3454

Dekker, J., Rippe, K., Dekker, M., & Kleckner, N. (2002). Capturing Chromosome Conformation. *Science*, *295*(5558), 1306–1311. https://doi.org/10.1126/science.1067799

DeMare, L. E., Leng, J., Cotney, J., Reilly, S. K., Yin, J., Sarro, R., & Noonan, J. P. (2013). The genomic landscape of cohesin-associated chromatin interactions. *Genome Research*, *23*(8), 1224–34. https://doi.org/10.1101/gr.156570.113

Deng, W., Lee, J., Wang, H., Miller, J., Reik, A., Gregory, P. D., … Blobel, G. A. (2012). Controlling long-range genomic interactions at a native locus by targeted tethering of a looping factor. *Cell*, *149*(6), 1233–44. https://doi.org/10.1016/j.cell.2012.03.051

Diao, Y., Fang, R., Li, B., Meng, Z., Yu, J., Qiu, Y., … Ren, B. (2017). A tiling-deletion-based genetic screen for cis-regulatory element identification in mammalian cells. *Nature Methods*, *14*(6), 629–635. https://doi.org/10.1038/nmeth.4264

Dixit, A., Parnas, O., Li, B., Chen, J., Fulco, C. P., Jerby-Arnon, L., … Regev, A. (2016). Perturb-Seq: Dissecting Molecular Circuits with Scalable Single-Cell RNA Profiling of Pooled Genetic Screens. *Cell*, *167*(7), 1853–1866.e17. https://doi.org/10.1016/j.cell.2016.11.038

Dixon, J. R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., … Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, *485*(7398), 376–80. https://doi.org/10.1038/nature11082

Dominguez, A. A., Lim, W. A., & Qi, L. S. (2015). Beyond editing: repurposing CRISPR–Cas9 for precision genome regulation and interrogation. *Nature Reviews Molecular Cell Biology*, *17*(1), 5–15. https://doi.org/10.1038/nrm.2015.2

Donnelly, P., Price, A. L., & Spencer, C. C. A. (n.d.). Progress and promise in understanding the genetic basis of common diseases. https://doi.org/10.1098/rspb.2015.1684

Drewell, R. A. (2011). Transcription Factor Binding Site Redundancy in Embryonic Enhancers of the *Drosophila* Bithorax Complex. *G3: Genes|Genomes|Genetics*, *1*(7), 603–606. https://doi.org/10.1534/g3.111.001404

Dror, I., Golan, T., Levy, C., Rohs, R., & Mandel-Gutfreund, Y. (2015). A widespread role of the motif environment in transcription factor binding across diverse protein families. *Genome Research*, *25*(9), 1268–1280. https://doi.org/10.1101/gr.184671.114

Dunipace, L., Ozdemir, A., & Stathopoulos, A. (2011a). Complex interactions between cis-regulatory modules in native conformation are critical for Drosophila snail expression. *Development (Cambridge, England)*, *138*(18), 4075–84. https://doi.org/10.1242/dev.069146

Ebright, R. H. (2000). RNA Polymerase: Structural Similarities Between Bacterial RNA Polymerase and Eukaryotic RNA Polymerase II. *Journal of Molecular Biology*, *304*(5), 687–698. https://doi.org/10.1006/JMBI.2000.4309

Edgar, R., Domrachev, M., & Lash, A. E. (2002). Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Research*, *30*(1), 207–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11752295

ENCODE Project Consortium, E. P. (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. *Science (New York, N.Y.)*, *306*(5696), 636–40. https://doi.org/10.1126/science.1105136

ENCODE Project Consortium, T. E. P. (2012a). An integrated encyclopedia of DNA elements in the human genome. *Nature*, *489*(7414), 57–74. https://doi.org/10.1038/nature11247

Ernst, J., & Kellis, M. (2010). Discovery and characterization of chromatin states for systematic annotation of the human genome. *Nature Biotechnology*, *28*(8), 817–825. https://doi.org/10.1038/nbt.1662

Ernst, J., Kheradpour, P., Mikkelsen, T. S., Shoresh, N., Ward, L. D., Epstein, C. B., … Bernstein, B. E. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, *473*(7345), 43–49. https://doi.org/10.1038/nature09906

Felsenfeld, G., & Groudine, M. (2003). Controlling the double helix. *Nature*, *421*(6921), 448–453. https://doi.org/10.1038/nature01411

Fields, D. S., He, Y., Al-Uzri, A. Y., & Stormo, G. D. (1997). Quantitative specificity of the Mnt repressor. *Journal of Molecular Biology*, *271*(2), 178–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9268651

Figueiredo, L. M., Cross, G. A. M., & Janzen, C. J. (2009). Epigenetic regulation in African trypanosomes: a new kid on the block. *Nature Reviews Microbiology*, *7*(7), 504–513. https://doi.org/10.1038/nrmicro2149

Finch, J. T., & Klug, A. (1976). Solenoidal model for superstructure in chromatin. *Proceedings of the National Academy of Sciences of the United States of America*, *73*(6), 1897–901. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/1064861

Flanagan, P. M., Kelleher, R. J., Sayre, M. H., Tschochner, H., & Kornberg, R. D. (1991). A mediator required for activation of RNA polymerase II transcription in vitro. *Nature*, *350*(6317), 436–438. https://doi.org/10.1038/350436a0

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., & Stern, D. L. (2010a). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, *466*. https://doi.org/10.1038/nature09158

Frankel, N., Davis, G. K., Vargas, D., Wang, S., Payre, F., & Stern, D. L. (2010b). Phenotypic robustness conferred by apparently redundant transcriptional enhancers. *Nature*, *466*(7305), 490–3. https://doi.org/10.1038/nature09158

Freire-Pritchett, P., Schoenfelder, S., Várnai, C., Wingett, S. W., Cairns, J., Collier, A. J., … Spivakov, M. (2017). Global reorganisation of *cis* -regulatory units upon lineage commitment of human embryonic stem cells. *eLife*, *6*. https://doi.org/10.7554/eLife.21926

Fukaya, T., Lim, B., & Levine, M. (2016). Enhancer Control of Transcriptional Bursting. *Cell*. https://doi.org/10.1016/j.cell.2016.05.025

Fullwood, M. J., Liu, M. H., Pan, Y. F., Liu, J., Xu, H., Mohamed, Y. Bin, … Ruan, Y. (2009). An oestrogen-receptor-α-bound human chromatin interactome. *Nature*, *462*(7269), 58–64. https://doi.org/10.1038/nature08497

Fussner, E., Ching, R. W., & Bazett-Jones, D. P. (2011). Living without 30nm chromatin fibers. *Trends in Biochemical Sciences*, *36*(1), 1–6. https://doi.org/10.1016/j.tibs.2010.09.002

Gaffney, D. J., Veyrieras, J.-B., Degner, J. F., Pique-Regi, R., Pai, A. A., Crawford, G. E., … Pritchard, J. K. (2012). Dissecting the regulatory architecture of gene expression QTLs. *Genome Biology*, *13*(1), R7. https://doi.org/10.1186/gb-2012-13-1-r7

Gallone, G., Haerty, W., Disanto, G., Ramagopalan, S. V., Ponting, C. P., & Berlanga-Taylor, A. J. (2017). Identification of genetic variants affecting vitamin D receptor binding and associations with autoimmune disease. *Human Molecular Genetics*, *26*(11), 2164–2176. https://doi.org/10.1093/hmg/ddx092

Gerchman, S. E., & Ramakrishnan, V. (1987). Chromatin higher-order structure studied by neutron scattering and scanning transmission electron microscopy. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(22), 7802–6. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3479765

Gerstein, M. B., Kundaje, A., Hariharan, M., Landt, S. G., Yan, K.-K., Cheng, C., … Snyder, M. (2012). Architecture of the human regulatory network derived from ENCODE data. *Nature*, *489*(7414), 91–100. https://doi.org/10.1038/nature11245

Gilad, Y., Rifkin, S. A., & Pritchard, J. K. (2008). Revealing the architecture of gene regulation: the promise of eQTL studies. *Trends in Genetics : TIG*, *24*(8), 408–15. https://doi.org/10.1016/j.tig.2008.06.001

Gillies, S. D., Morrison, S. L., Oi, V. T., Tonegawa, S., Kelley, D. E., Perry, R. P., & Hood, L. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. *Cell*, *33*(3), 717–28. https://doi.org/10.1016/0092-8674(83)90014-4

Giorgetti, L., & Heard, E. (2016). Closing the loop: 3C versus DNA FISH. *Genome Biology*, *17*(1), 215. https://doi.org/10.1186/s13059-016-1081-2

Gompel, N., Prud'homme, B., Wittkopp, P. J., Kassner, V. A., & Carroll, S. B. (2005). Chance caught on the wing: cis-regulatory evolution and the origin of pigment patterns in Drosophila. *Nature*, *433*(7025), 481–487. https://doi.org/10.1038/nature03235

Grünberg, S., & Hahn, S. (2013). Structural insights into transcription initiation by RNA polymerase II. *Trends in Biochemical Sciences*, *38*(12), 603–11. https://doi.org/10.1016/j.tibs.2013.09.002

Grundberg, E., Small, K. S., Hedman, Å. K., Nica, A. C., Buil, A., Keildson, S., … Consortium, T. M. T. H. E. R. (MuTHER). (2012). Mapping cis- and trans-regulatory effects across multiple tissues in twins. *Nature Genetics*, *44*(10), 1084–1089. https://doi.org/10.1038/ng.2394

GTEx Consortium, Gte. (2015). Human genomics. The Genotype-Tissue Expression (GTEx) pilot analysis: multitissue gene regulation in humans. *Science (New York, N.Y.)*, *348*(6235), 648–60. https://doi.org/10.1126/science.1262110

Guerrero, L., Marco-Ferreres, R., Serrano, A. L., Arredondo, J. J., & Cervera, M. (2010a). Secondary enhancers synergise with primary enhancers to guarantee fine-tuned muscle gene expression. *Developmental Biology*, *337*(1), 16–28. https://doi.org/10.1016/j.ydbio.2009.10.006

Hahn, S. (2004). Structure and mechanism of the RNA polymerase II transcription machinery. *Nature Structural & Molecular Biology*, *11*(5), 394–403. https://doi.org/10.1038/nsmb763

Hampsey, M. (1998). Molecular genetics of the RNA polymerase II general transcriptional machinery. *Microbiology and Molecular Biology Reviews : MMBR*, *62*(2), 465–503. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9618449

Hay, D., Hughes, J. R., Babbs, C., Davies, J. O. J., Graham, B. J., Hanssen, L. L. P., … Higgs, D. R. (2016). Genetic dissection of the α-globin super-enhancer in vivo. *Nature Genetics*, *48*(8), 895–903. https://doi.org/10.1038/ng.3605

Heintzman, N. D., Hon, G. C., Hawkins, R. D., Kheradpour, P., Stark, A., Harp, L. F., … Ren, B. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. *Nature*, *459*(7243), 108–12. https://doi.org/10.1038/nature07829

Hergeth, S. P., & Schneider, R. (2015). The H1 linker histones: multifunctional proteins beyond the nucleosomal core particle. *EMBO Reports*, *16*(11), 1439–53. https://doi.org/10.15252/embr.201540749

Hilton, I. B., D'Ippolito, A. M., Vockley, C. M., Thakore, P. I., Crawford, G. E., Reddy, T. E., & Gersbach, C. A. (2015). Epigenome editing by a CRISPR-Cas9-based acetyltransferase activates genes from promoters and enhancers. *Nature Biotechnology*, *33*(5), 510–517. https://doi.org/10.1038/nbt.3199

Hinrichs, A. S., Karolchik, D., Baertsch, R., Barber, G. P., Bejerano, G., Clawson, H., … Kent, W. J. (2006). The UCSC Genome Browser Database: update 2006. *Nucleic Acids Research*, *34*(90001), D590–D598. https://doi.org/10.1093/nar/gkj144

Hirschhorn, J. N., & Daly, M. J. (2005). Genome-wide association studies for common diseases and complex traits. *Nature Reviews Genetics*, *6*(2), 95–108. https://doi.org/10.1038/nrg1521

Hong, J.-W., Hendrix, D. A., & Levine, M. S. (2008). Shadow Enhancers as a Source of Evolutionary Novelty. *Science*, *321*(5894), 1314–1314. https://doi.org/10.1126/science.1160631

Hozumi, A., Yoshida, R., Horie, T., Sakuma, T., Yamamoto, T., & Sasakura, Y. (2013). Enhancer activity sensitive to the orientation of the gene it regulates in the chordategenome. *Developmental Biology*, *375*(1), 79–91. https://doi.org/10.1016/j.ydbio.2012.12.012

Hughes, J. R., Roberts, N., McGowan, S., Hay, D., Giannoulatou, E., Lynch, M., … Higgs, D. R. (2014). Analysis of hundreds of cis-regulatory landscapes at high resolution in a single, high-throughput experiment. *Nature Genetics*, *46*(2), 205–212. https://doi.org/10.1038/ng.2871

Hussain, T., & Mulherkar, R. (2012). Lymphoblastoid Cell lines: a Continuous in Vitro Source of Cells to Study Carcinogen Sensitivity and DNA Repair. *International Journal of Molecular and Cellular Medicine*, *1*(2), 75–87. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/24551762

Inoue, F., & Ahituv, N. (2015). Decoding enhancers using massively parallel reporter assays. *Genomics*, *106*(3), 159–164. https://doi.org/10.1016/j.ygeno.2015.06.005

Ip, Y. T., Park, R. E., Kosman, D., Yazdanbakhsh, K., & Levine, M. (1992). dorsal-twist interactions establish snail expression in the presumptive mesoderm of the Drosophila embryo. *Genes & Development*, *6*(8), 1518–30.

Jäger, R., Migliorini, G., Henrion, M., Kandaswamy, R., Speedy, H. E., Heindl, A., … Houlston, R. S. (2015). Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nature Communications*, *6*, 6178. https://doi.org/10.1038/ncomms7178

Jaitin, D. A., Weiner, A., Yofe, I., Lara-Astiaso, D., Keren-Shaul, H., David, E., … Amit, I. (2016). Dissecting Immune Circuits by Linking CRISPR-Pooled Screens with Single-Cell RNA-Seq. *Cell*, *167*(7), 1883–1896.e15. https://doi.org/10.1016/j.cell.2016.11.039

Javierre, B. M., Burren, O. S., Wilder, S. P., Kreuzhuber, R., Hill, S. M., Sewitz, S., … Flicek, P. (2016a). Lineage-Specific Genome Architecture Links Enhancers and Non-coding Disease Variants to Target Gene Promoters. *Cell*, *167*(5), 1369–1384.e19. https://doi.org/10.1016/j.cell.2016.09.037

Jeong, Y., El-Jaick, K., Roessler, E., Muenke, M., & Epstein, D. J. (2006). A functional screen for sonic hedgehog regulatory elements across a 1 Mb interval identifies long-range ventral forebrain enhancers. *Development*, *133*(4), 761–772. https://doi.org/10.1242/dev.02239

Jin, F., Li, Y., Dixon, J. R., Selvaraj, S., Ye, Z., Lee, A. Y., … Ren, B. (2013). A high-resolution map of

the three-dimensional chromatin interactome in human cells. *Nature*, *503*(7475), 290–4. https://doi.org/10.1038/nature12644

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E. H., Birney, E., & Furlong, E. E. M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. *Cell*, *148*(3), 473–86. https://doi.org/10.1016/j.cell.2012.01.030

Kaimal, V., Bardes, E. E., Tabar, S. C., Jegga, A. G., & Aronow, B. J. (2010). ToppCluster: a multiple gene list feature analyzer for comparative enrichment clustering and network-based dissection of biological systems. *Nucleic Acids Research*, *38*(Web Server issue), W96-102. https://doi.org/10.1093/nar/gkq418

Kalmar, T., Lim, C., Hayward, P., Muñoz-Descalzo, S., Nichols, J., Garcia-Ojalvo, J., & Martinez Arias, A. (2009). Regulated fluctuations in nanog expression mediate cell fate decisions in embryonic stem cells. *PLoS Biology*, *7*(7), e1000149. https://doi.org/10.1371/journal.pbio.1000149

Kasowski, M., Kyriazopoulou-Panagiotopoulou, S., Grubert, F., Zaugg, J. B., Kundaje, A., Liu, Y., … Snyder, M. (2013). Extensive Variation in Chromatin States Across Humans. *Science*, *342*(6159), 750–752. https://doi.org/10.1126/science.1242510

Kearns, N. A., Pham, H., Tabak, B., Genga, R. M., Silverstein, N. J., Garber, M., & Maehr, R. (2015). Functional annotation of native enhancers with a Cas9–histone demethylase fusion. *Nature Methods*, *12*(5), 401–403. https://doi.org/10.1038/nmeth.3325

Kelleher, R. J., Flanagan, P. M., & Kornberg, R. D. (1990). A novel mediator between activator proteins and the RNA polymerase II transcription apparatus. *Cell*, *61*(7), 1209–15. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2163759

Kelley, D. R., Snoek, J., & Rinn, J. L. (2016). Basset: learning the regulatory code of the accessible genome with deep convolutional neural networks. *Genome Research*, *26*(7), 990–9. https://doi.org/10.1101/gr.200535.115

Kheradpour, P., Ernst, J., Melnikov, A., Rogov, P., Wang, L., Zhang, X., … Kellis, M. (2013). Systematic dissection of regulatory motifs in 2000 predicted human enhancers using a massively parallel reporter assay. *Genome Research*, *23*(5), 800–811.

Kichaev, G., Yang, W.-Y., Lindstrom, S., Hormozdiari, F., Eskin, E., Price, A. L., … Pasaniuc, B. (2014). Integrating Functional Data to Prioritize Causal Variants in Statistical Fine-Mapping Studies. *PLoS Genetics*, *10*(10), e1004722. https://doi.org/10.1371/journal.pgen.1004722

Kilpinen, H., Waszak, S. M., Gschwind, A. R., Raghav, S. K., Witwicki, R. M., Orioli, A., … Dermitzakis, E. T. (2013a). Coordinated effects of sequence variation on DNA binding, chromatin structure, and transcription. *Science (New York, N.Y.)*, *342*(6159), 744–7. https://doi.org/10.1126/science.1242463

Kim, T.-K., Hemberg, M., Gray, J. M., Costa, A. M., Bear, D. M., Wu, J., … Greenberg, M. E. (2010). Widespread transcription at neuronal activity-regulated enhancers. *Nature*, *465*(7295), 182–187. https://doi.org/10.1038/nature09033

Kim, T.-K., & Shiekhattar, R. (2015). Leading Edge Review Architectural and Functional Commonalities between Enhancers and Promoters. *Cell*, *162*, 948–959. https://doi.org/10.1016/j.cell.2015.08.008

Knezetic, J. A., & Luse, D. S. (1986). The presence of nucleosomes on a DNA template prevents initiation by RNA polymerase II in vitro. *Cell*, *45*(1), 95–104. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3955658

Koleske, A. J., & Young, R. A. (1994). An RNA polymerase II holoenzyme responsive to activators. *Nature*, *368*(6470), 466–469. https://doi.org/10.1038/368466a0

Kornberg, R. D., & Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. *Cell*, *98*(3), 285–94. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10458604

Kouzarides, T. (2007). Chromatin Modifications and Their Function. *Cell*, *128*(4), 693–705. https://doi.org/10.1016/J.CELL.2007.02.005

Kowalczyk, M. S., Hughes, J. R., Garrick, D., Lynch, M. D., Sharpe, J. A., Sloane-Stanley, J. A., … Higgs, D. R. (2012). Intragenic Enhancers Act as Alternative Promoters. *Molecular Cell*, *45*(4), 447–458. https://doi.org/10.1016/j.molcel.2011.12.021

Kraus, W. L., Manning, E. T., & Kadonaga, J. T. (1999). Biochemical analysis of distinct activation functions in p300 that enhance transcription initiation with chromatin templates. *Molecular and Cellular Biology*, *19*(12), 8123–35. https://doi.org/10.1128/MCB.19.12.8123

Kulkarni, M. M., & Arnosti, D. N. (2003). Information display by transcriptional enhancers. *Development (Cambridge, England)*, *130*(26), 6569–75. https://doi.org/10.1242/dev.00890

Kumasaka, N., Knights, A. J., & Gaffney, D. J. (2015). Fine-mapping cellular QTLs with RASQUAL and ATAC-seq. *Nature Genetics*, *48*(2), 206–213. https://doi.org/10.1038/ng.3467

Kwasnieski, J. C., Mogno, I., Myers, C. A., Corbo, J. C., & Cohen, B. A. (2012). Complex effects of nucleotide variants in a mammalian cis-regulatory element. *Proceedings of the National Academy of Sciences*, *109*(47), 19498–19503. https://doi.org/10.1073/pnas.1210678109

Lam, D. D., de Souza, F. S. J., Nasif, S., Yamashita, M., López-Leal, R., Otero-Corchon, V., … Low, M. J. (2015). Partially Redundant Enhancers Cooperatively Maintain Mammalian Pomc Expression Above a Critical Functional Threshold. *PLOS Genetics*, *11*(2), e1004935. https://doi.org/10.1371/journal.pgen.1004935

Lappalainen, T., Sammeth, M., Friedländer, M. R., ‘t Hoen, P. A. C., Monlong, J., Rivas, M. A., … Dermitzakis, E. T. (2013). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–511. https://doi.org/10.1038/nature12531

Lappalainen, T., Sammeth, M., Friedländer, M. R., ’t Hoen, P. A. C., Monlong, J., Rivas, M. A., … Dermitzakis, E. T. (2013a). Transcriptome and genome sequencing uncovers functional variation in humans. *Nature*, *501*(7468), 506–11. https://doi.org/10.1038/nature12531

Lawrence, M., Daujat, S., & Schneider, R. (2016). Lateral Thinking: How Histone Modifications Regulate Gene Expression. *Trends in Genetics : TIG*, *32*(1), 42–56. https://doi.org/10.1016/j.tig.2015.10.007

Leddin, M., Perrod, C., Hoogenkamp, M., Ghani, S., Assi, S., Heinz, S., … Rosenbauer, F. (2011). Two distinct auto-regulatory loops operate at the PU.1 locus in B cells and myeloid cells. *Blood*, *117*(10), 2827–38. https://doi.org/10.1182/blood-2010-08-302976

Lee, D., Gorkin, D. U., Baker, M., Strober, B. J., Asoni, A. L., McCallion, A. S., & Beer, M. A. (2015). A method to predict the impact of regulatory variants from DNA sequence. *Nature Genetics*, *47*(8), 955–61. https://doi.org/10.1038/ng.3331

Lee, T. I., & Young, R. A. (2000). Transcription of Eukaryotic Protein-Coding Genes. *Annual Review of Genetics*, *34*(1), 77–137. https://doi.org/10.1146/annurev.genet.34.1.77

Lee, W., Tillo, D., Bray, N., Morse, R. H., Davis, R. W., Hughes, T. R., & Nislow, C. (2007). A high-resolution atlas of nucleosome occupancy in yeast. *Nature Genetics*, *39*(10), 1235–44.

https://doi.org/10.1038/ng2117

Lettice, L. A., Heaney, S. J. H., Purdie, L. A., Li, L., de Beer, P., Oostra, B. A., … de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Human Molecular Genetics*, *12*(14), 1725–35. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12837695

Lettice, L. A., Hill, A. E., Devenney, P. S., & Hill, R. E. (2008). Point mutations in a distant sonic hedgehog cis-regulator generate a variable regulatory output responsible for preaxial polydactyly. *Human Molecular Genetics*, *17*(7), 978–985. https://doi.org/10.1093/hmg/ddm370

Lettice, L. A., Horikoshi, T., Heaney, S. J. H., van Baren, M. J., van der Linde, H. C., Breedveld, G. J., … Noji, S. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. *Proceedings of the National Academy of Sciences of the United States of America*, *99*(11), 7548–53. https://doi.org/10.1073/pnas.112212199

Lettice, L. A., Williamson, I., Wiltshire, J. H., Peluso, S., Devenney, P. S., Hill, A. E., … Hill, R. E. (2012). Opposing Functions of the ETS Factor Family Define Shh Spatial Expression in Limb Buds and Underlie Polydactyly. *Developmental Cell*, *22*(2), 459–467. https://doi.org/10.1016/j.devcel.2011.12.010

Leung, D., Jung, I., Rajagopal, N., Schmitt, A., Selvaraj, S., Lee, A. Y., … Ren, B. (2015). Integrative analysis of haplotype-resolved epigenomes across human tissues. *Nature*, *518*(7539), 350–354. https://doi.org/10.1038/nature14217

Leung, M. K. K., Xiong, H. Y., Lee, L. J., & Frey, B. J. (2014). Deep learning of the tissue-regulated splicing code. *Bioinformatics (Oxford, England)*, *30*(12), i121-9. https://doi.org/10.1093/bioinformatics/btu277

Levine, M., Cattoglio, C., & Tjian, R. (2014). Looping back to leap forward: transcription enters a new era. *Cell*, *157*(1), 13–25. https://doi.org/10.1016/j.cell.2014.02.009

Levo, M., Zalckvar, E., Sharon, E., Dantas Machado, A. C., Kalma, Y., Lotam-Pompan, M., … Segal, E. (2015). Unraveling determinants of transcription factor binding outside the core binding site. *Genome Research*, *25*(7), 1018–29. https://doi.org/10.1101/gr.185033.114

Li, B., Carey, M., & Workman, J. L. (2007). The role of chromatin during transcription. *Cell*, *128*(4), 707–19. https://doi.org/10.1016/j.cell.2007.01.015

Li, G., Ruan, X., Auerbach, R. K., Sandhu, K. S., Zheng, M., Wang, P., … Ruan, Y. (2012). Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. *Cell*, *148*(1–2), 84–98. https://doi.org/10.1016/j.cell.2011.12.014

Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. A., Iyer, V. N., … Biggin, M. D. (2008). Transcription Factors Bind Thousands of Active and Inactive Regions in the Drosophila Blastoderm. *PLoS Biology*, *6*(2), e27. https://doi.org/10.1371/journal.pbio.0060027

Liang, G., Lin, J. C. Y., Wei, V., Yoo, C., Cheng, J. C., Nguyen, C. T., … Jones, P. A. (2004). Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome. *Proceedings of the National Academy of Sciences*, *101*(19), 7357–7362. https://doi.org/10.1073/pnas.0401866101

Liang, J., Lacroix, L., Gamot, A., Cuddapah, S., Queille, S., Lhoumaud, P., … Cuvier, O. (2014). Chromatin Immunoprecipitation Indirect Peaks Highlight Long-Range Interactions of Insulator Proteins and Pol II Pausing. *Molecular Cell*, *53*, 672–681. https://doi.org/10.1016/j.molcel.2013.12.029

Liberman, L. M., & Stathopoulos, A. (2009). Design flexibility in cis-regulatory control of gene expression: synthetic and comparative evidence. *Developmental Biology*, *327*(2), 578–89. https://doi.org/10.1016/j.ydbio.2008.12.020

Lieberman-Aiden, E., van Berkum, N. L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., … Dekker, J. (2009). Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome. *Science*, *326*(5950), 289–293. https://doi.org/10.1126/science.1181369

Lomvardas, S., & Thanos, D. (2001). Nucleosome Sliding via TBP DNA Binding In Vivo. *Cell*, *106*(6), 685–696. https://doi.org/10.1016/S0092-8674(01)00490-1

Lomvardas, S., & Thanos, D. (2002). Modifying gene expression programs by altering core promoter chromatin architecture. *Cell*, *110*(2), 261–71. https://doi.org/10.1016/S0092-8674(02)00822-X

Long, H. K., Prescott, S. L., & Wysocka, J. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, *167*, 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018

Long, H. K., Prescott, S. L., Wysocka, J., Ye, Z., Kolovos, P., Brouwer, R. W., … al., et. (2016). Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell*, *167*(5), 1170–1187. https://doi.org/10.1016/j.cell.2016.09.018

Lopes, R., Korkmaz, G., & Agami, R. (2016). Applying CRISPR–Cas9 tools to identify and characterize transcriptional enhancers. *Nature Reviews Molecular Cell Biology*, *17*(9), 597–604. https://doi.org/10.1038/nrm.2016.79

Lorch, Y., LaPointe, J. W., & Kornberg, R. D. (1987). Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones. *Cell*, *49*(2), 203–10. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/3568125

Lu, J., & Clark, A. G. (2012). Impact of microRNA regulation on variation in human gene expression. *Genome Research*, *22*(7), 1243–54. https://doi.org/10.1101/gr.132514.111

Luger, K., Dechassa, M. L., & Tremethick, D. J. (2012). New insights into nucleosome and chromatin structure: an ordered state or a disordered affair? *Nature Reviews Molecular Cell Biology*, *13*(7), 436–447. https://doi.org/10.1038/nrm3382

Luger, K., Mäder, A. W., Richmond, R. K., Sargent, D. F., & Richmond, T. J. (1997). Crystal structure of the nucleosome core particle at 2.8 Å resolution. *Nature*, *389*(6648), 251–260. https://doi.org/10.1038/38444

Macintyre, G., Bailey, J., Haviv, I., & Kowalczyk, A. (2010). is-rSNP: a novel technique for in silico regulatory SNP detection. *Bioinformatics*, *26*(18), i524–i530. https://doi.org/10.1093/bioinformatics/btq378

Macneil, L. T., & Walhout, A. J. M. (2011). Gene regulatory networks and the role of robustness and stochasticity in the control of gene expression. *Genome Research*, *21*(5), 645–57. https://doi.org/10.1101/gr.097378.109

Maeshima, K., Hihara, S., & Eltsov, M. (2010). Chromatin structure: does the 30-nm fibre exist in vivo? *Current Opinion in Cell Biology*, *22*(3), 291–297. https://doi.org/10.1016/j.ceb.2010.03.001

Majumder, P., Gomez, J. A., Chadwick, B. P., & Boss, J. M. (2008). The insulator factor CTCF controls MHC class II gene expression and is required for the formation of long-distance

chromatin interactions. *The Journal of Experimental Medicine*, *205*(4), 785–98. https://doi.org/10.1084/jem.20071843

Mali, P., Yang, L., Esvelt, K. M., Aach, J., Guell, M., DiCarlo, J. E., … Church, G. M. (2013). RNA-guided human genome engineering via Cas9. *Science (New York, N.Y.)*, *339*(6121), 823–6. https://doi.org/10.1126/science.1232033

Manke, T., Heinig, M., & Vingron, M. (2010). Quantifying the effect of sequence variation on regulatory interactions. *Human Mutation*. https://doi.org/10.1002/humu.21209

Manke, T., Roider, H. G., & Vingron, M. (2008). Statistical modeling of transcription factor binding affinities predicts regulatory interactions. *PLoS Computational Biology*. https://doi.org/10.1371/journal.pcbi.1000039

Marinov, G. K., Williams, B. A., McCue, K., Schroth, G. P., Gertz, J., Myers, R. M., & Wold, B. J. (2014). From single-cell to cell-pool transcriptomes: stochasticity in gene expression and RNA splicing. *Genome Research*, *24*(3), 496–510. https://doi.org/10.1101/gr.161034.113

Marmorstein, R. Q., & Sigler, P. B. (1989). Structure and Mechanism of the trp Repressor/Operator System (pp. 56–78). Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-83709-8_5

Martin, P., McGovern, A., Orozco, G., Duffus, K., Yarwood, A., Schoenfelder, S., … Eyre, S. (2015). Capture Hi-C reveals novel candidate genes and complex long-range interactions with related autoimmune risk loci. *Nature Communications*, *6*, 10069. https://doi.org/10.1038/ncomms10069

Maston, G. A., Landt, S. G., Snyder, M., & Green, M. R. (2012). Characterization of Enhancer Function from Genome-Wide Analyses. *Annual Review of Genomics and Human Genetics*, *13*(1), 29–57. https://doi.org/10.1146/annurev-genom-090711-163723

Matharu, N., & Ahituv, N. (2015). Minor Loops in Major Folds: Enhancer-Promoter Looping, Chromatin Restructuring, and Their Association with Transcriptional Regulation and Disease. *PLoS Genetics*, *11*(12), e1005640. https://doi.org/10.1371/journal.pgen.1005640

Maurano, M. T., Humbert, R., Rynes, E., Thurman, R. E., Haugen, E., Wang, H., … Stamatoyannopoulos, J. A. (2012). Systematic localization of common disease-associated variation in regulatory DNA. *Science (New York, N.Y.)*, *337*(6099), 1190–5. https://doi.org/10.1126/science.1222794

McCarthy, M. I., & Hirschhorn, J. N. (2008). Genome-wide association studies: past, present and future. *Human Molecular Genetics*, *17*(R2), R100–R101. https://doi.org/10.1093/hmg/ddn298

McLaren, W., Gil, L., Hunt, S. E., Riat, H. S., Ritchie, G. R. S., Thormann, A., … Cunningham, F. (2016). The Ensembl Variant Effect Predictor. *Genome Biology*, *17*(1), 122. https://doi.org/10.1186/s13059-016-0974-4

McVicker, G., van de Geijn, B., Degner, J. F., Cain, C. E., Banovich, N. E., Raj, A., … Pritchard, J. K. (2013). Identification of Genetic Variants That Affect Histone Modifications in Human Cells. *Science*, *342*(6159), 747–749. https://doi.org/10.1126/science.1242429

Mendenhall, E. M., Williamson, K. E., Reyon, D., Zou, J. Y., Ram, O., Joung, J. K., & Bernstein, B. E. (2013). Locus-specific editing of histone modifications at endogenous enhancers. *Nature Biotechnology*, *31*(12), 1133–1136. https://doi.org/10.1038/nbt.2701

Messina, D. N., Glasscock, J., Gish, W., & Lovett, M. (2004). An ORFeome-based Analysis of Human

Transcription Factor Genes and the Construction of a Microarray to Interrogate Their Expression. *Genome Research*, *14*(10b), 2041–2047. https://doi.org/10.1101/gr.2584104

Meyer, K. D., Lin, S., Bernecky, C., Gao, Y., & Taatjes, D. J. (2010). p53 activates transcription by directing structural shifts in Mediator. *Nature Structural & Molecular Biology*, *17*(6), 753–760. https://doi.org/10.1038/nsmb.1816

Mifsud, B., Tavares-Cadete, F., Young, A. N., Sugar, R., Schoenfelder, S., Ferreira, L., … Osborne, C. S. (2015). Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nature Genetics*, *47*(6), 598–606. https://doi.org/10.1038/ng.3286

Miller, J. A., & Widom, J. (2003). Collaborative competition mechanism for gene activation in vivo. *Molecular and Cellular Biology*, *23*(5), 1623–32. https://doi.org/10.1128/mcb.23.5.1623-1632.2003

Montgomery, S. B., Sammeth, M., Gutierrez-Arcelus, M., Lach, R. P., Ingle, C., Nisbett, J., … Dermitzakis, E. T. (2010). Transcriptome genetics using second generation sequencing in a Caucasian population. *Nature*, *464*(7289), 773–777. https://doi.org/10.1038/nature08903

Moyerbrailean, G. A., Kalita, C. A., Harvey, C. T., Wen, X., Luca, F., & Pique-Regi, R. (2016). Which Genetics Variants in DNase-Seq Footprints Are More Likely to Alter Binding? *PLOS Genetics*, *12*(2), e1005875. https://doi.org/10.1371/journal.pgen.1005875

Mumbach, M. R., Rubin, A. J., Flynn, R. A., Dai, C., Khavari, P. A., Greenleaf, W. J., & Chang, H. Y. (2016). HiChIP: efficient and sensitive analysis of protein-directed genome architecture. *Nature Methods*, *13*(11), 919–922. https://doi.org/10.1038/nmeth.3999

Nguyen, T. A., Jones, R. D., Snavely, A. R., Pfenning, A. R., Kirchner, R., Hemberg, M., & Gray, J. M. (2016). High-throughput functional comparison of promoter and enhancer activities. *Genome Research*, *26*(8), 1023–33. https://doi.org/10.1101/gr.204834.116

Nica, A. C., & Dermitzakis, E. T. (2013a). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Nica, A. C., & Dermitzakis, E. T. (2013b). Expression quantitative trait loci: present and future. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences*, *368*(1620), 20120362. https://doi.org/10.1098/rstb.2012.0362

Nishino, Y., Eltsov, M., Joti, Y., Ito, K., Takata, H., Takahashi, Y., … Maeshima, K. (2012). Human mitotic chromosomes consist predominantly of irregularly folded nucleosome fibres without a 30-nm chromatin structure. *The EMBO Journal*, *31*(7), 1644–1653. https://doi.org/10.1038/emboj.2012.35

Noll, M. (1974). Subunit structure of chromatin. *Nature*, *251*(5472), 249–51. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4422492

Nonet, M. L., & Young, R. A. (1989). Intragenic and extragenic suppressors of mutations in the heptapeptide repeat domain of Saccharomyces cerevisiae RNA polymerase II. *Genetics*, *123*(4), 715–24. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/2693207

Olins, A. L., & Olins, D. E. (1974). Spheroid chromatin units (v bodies). *Science (New York, N.Y.)*, *183*(4122), 330–2. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4128918

Orphanides, G., Lagrange, T., & Reinberg, D. (1996). The general transcription factors of RNA polymerase II. *Genes & Development*, *10*(21), 2657–83. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8946909

Ozsolak, F., Song, J. S., Liu, X. S., & Fisher, D. E. (2007). High-throughput mapping of the chromatin structure of human promoters. *Nature Biotechnology*, *25*(2), 244–8. https://doi.org/10.1038/nbt1279

Pai, A. A., Cain, C. E., Mizrahi-Man, O., De Leon, S., Lewellen, N., Veyrieras, J.-B., … Gilad, Y. (2012). The Contribution of RNA Decay Quantitative Trait Loci to Inter-Individual Variation in Steady-State Gene Expression Levels. *PLoS Genetics*, *8*(10), e1003000. https://doi.org/10.1371/journal.pgen.1003000

Pai, A. A., Pritchard, J. K., & Gilad, Y. (2015). The Genetic and Mechanistic Basis for Variation in Gene Regulation. *PLoS Genetics*, *11*(1), e1004857. https://doi.org/10.1371/journal.pgen.1004857

Patwardhan, R. P., Hiatt, J. B., Witten, D. M., Kim, M. J., Smith, R. P., May, D., … Shendure, J. (2012). Massively parallel functional dissection of mammalian enhancers in vivo. *Nature Biotechnology*, *30*(3), 265–70. https://doi.org/10.1038/nbt.2136

Pennisi, E. (1997). Opening the way to gene activity. *Science (New York, N.Y.)*, *275*(5297), 155–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8999545

Perner, J., & Chung, H.-R. (2013). Chromatin signaling and transcription initiation. *Frontiers in Life Science*, *7*(1–2), 22–30. https://doi.org/10.1080/21553769.2013.856038

Perry, M. W., Boettiger, A. N., Bothma, J. P., & Levine, M. (2010). Shadow enhancers foster robustness of drosophila gastrulation. *Current Biology*. https://doi.org/10.1016/j.cub.2010.07.043

Perry, M. W., Boettiger, A. N., & Levine, M. (2011). Multiple enhancers ensure precision of gap gene-expression patterns in the Drosophila embryo. *Proceedings of the National Academy of Sciences*, *108*(33), 13570–13575. https://doi.org/10.1073/pnas.1109873108

Pokholok, D. K., Harbison, C. T., Levine, S., Cole, M., Hannett, N. M., Lee, T. I., … Young, R. A. (2005). Genome-wide Map of Nucleosome Acetylation and Methylation in Yeast. *Cell*, *122*(4), 517–527. https://doi.org/10.1016/j.cell.2005.06.026

Poss, Z. C., Ebmeier, C. C., & Taatjes, D. J. (2013). The Mediator complex and transcription regulation. *Critical Reviews in Biochemistry and Molecular Biology*, *48*(6), 575–608. https://doi.org/10.3109/10409238.2013.840259

Ptashne, M., & Gann, A. (1997). Transcriptional activation by recruitment. *Nature*, *386*(6625), 569–577. https://doi.org/10.1038/386569a0

Rao, S. S. P., Huntley, M. H., Durand, N. C., Stamenova, E. K., Bochkov, I. D., Robinson, J. T., … Aiden, E. L. (2014). A 3D Map of the Human Genome at Kilobase Resolution Reveals Principles of Chromatin Looping. *Cell*, *159*(7), 1665–1680. https://doi.org/10.1016/j.cell.2014.11.021

Reimand, J., Kull, M., Peterson, H., Hansen, J., & Vilo, J. (2007). g:Profiler--a web-based toolset for functional profiling of gene lists from large-scale experiments. *Nucleic Acids Research*, *35*(Web Server issue), W193-200. https://doi.org/10.1093/nar/gkm226

Reinke, H., & Hörz, W. (2003). Histones are first hyperacetylated and then lose contact with the activated PHO5 promoter. *Molecular Cell*, *11*(6), 1599–607. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12820972

Risca, V. I., & Greenleaf, W. J. (2015). Unraveling the 3D genome: genomics tools for multiscale exploration. *Trends in Genetics : TIG*, *31*(7), 357–72.

https://doi.org/10.1016/j.tig.2015.03.010

Roeder, R. G. (1996). The role of general initiation factors in transcription by RNA polymerase II. *Trends in Biochemical Sciences*, *21*(9), 327–335. https://doi.org/10.1016/S0968-0004(96)10050-5

Roeder, R. G., & Rutter, W. J. (1969). Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms. *Nature*, *224*(5216), 234–7. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5344598

Roeder, R. G., & Rutter, W. J. (1970). Specific nucleolar and nucleoplasmic RNA polymerases. *Proceedings of the National Academy of Sciences of the United States of America*, *65*(3), 675–82. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/5267147

Roh, T.-Y., Cuddapah, S., & Zhao, K. (2005). Active chromatin domains are defined by acetylation islands revealed by genome-wide mapping. *Genes & Development*, *19*(5), 542–52. https://doi.org/10.1101/gad.1272505

Roider, H. G., Kanhere, A., Manke, T., & Vingron, M. (2007). Predicting transcription factor affinities to DNA from a biophysical model. *Bioinformatics*, *23*(2), 134–141. https://doi.org/10.1093/bioinformatics/btl565

Rojo, F. (1999). Repression of transcription initiation in bacteria. *Journal of Bacteriology*, *181*(10), 2987–91. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10321997

Sagai, T., Amano, T., Tamura, M., Mizushina, Y., Sumiyama, K., & Shiroishi, T. (2009). A cluster of three long-range enhancers directs regional Shh expression in the epithelial linings. *Development*, *136*(10), 1665–1674. https://doi.org/10.1242/dev.032714

Sanborn, A. L., Rao, S. S. P., Huang, S.-C., Durand, N. C., Huntley, M. H., Jewett, A. I., … Aiden, E. L. (2015). Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(47), E6456-65. https://doi.org/10.1073/pnas.1518552112

Sanyal, A., Lajoie, B. R., Jain, G., & Dekker, J. (2012a). The long-range interaction landscape of gene promoters. *Nature*, *489*(7414), 109–13. https://doi.org/10.1038/nature11279

Sauter, K. A., Bouhlel, M. A., O'Neal, J., Sester, D. P., Tagoh, H., Ingram, R. M., … Hume, D. A. (2013). The Function of the Conserved Regulatory Element within the Second Intron of the Mammalian Csf1r Locus. *PLoS ONE*, *8*(1), e54935. https://doi.org/10.1371/journal.pone.0054935

Schaub, M. A., Boyle, A. P., Kundaje, A., Batzoglou, S., & Snyder, M. (2012). Linking disease associations with regulatory information in the human genome. *Genome Research*, *22*(9), 1748–59. https://doi.org/10.1101/gr.136127.111

Schmitt, A. D., Hu, M., & Ren, B. (2016). Genome-wide mapping and analysis of chromosome architecture. *Nature Reviews Molecular Cell Biology*, *17*(12), 743–755. https://doi.org/10.1038/nrm.2016.104

Schoenfelder, S., Furlan-Magaril, M., Mifsud, B., Tavares-Cadete, F., Sugar, R., Javierre, B.-M., … Fraser, P. (2015). The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Research*, *25*(4), 582–97. https://doi.org/10.1101/gr.185272.114

Sebé-Pedrós, A., Ballaré, C., Parra-Acero, H., Chiva, C., Tena, J. J., Sabidó, E., … Ruiz-Trillo, I. (2016). The Dynamic Regulatory Genome of Capsaspora and the Origin of Animal

Multicellularity. *Cell*, *165*(5), 1224–1237. https://doi.org/10.1016/j.cell.2016.03.034

Seshasayee, A. S. N., Sivaraman, K., & Luscombe, N. M. (2011). An Overview of Prokaryotic Transcription Factors. In *Sub-cellular biochemistry* (Vol. 52, pp. 7–23). https://doi.org/10.1007/978-90-481-9069-0_2

Shapiro, M. D., Marks, M. E., Peichel, C. L., Blackman, B. K., Nereng, K. S., Jónsson, B., … Kingsley, D. M. (2004). Genetic and developmental basis of evolutionary pelvic reduction in threespine sticklebacks. *Nature*, *428*(6984), 717–723. https://doi.org/10.1038/nature02415

Shen, Y., Yue, F., Mccleary, D. F., Ye, Z., Edsall, L., Kuan, S., … Ren, B. (2012). A map of the cis-regulatory sequences in the mouse genome. https://doi.org/10.1038/nature11243

Shlyueva, D., Stampfel, G., & Stark, A. (2014a). Transcriptional enhancers: from properties to genome-wide predictions. *Nature Reviews Genetics*, *15*(4), 272–286. https://doi.org/10.1038/nrg3682

Sie, L., Loong, S., & Tan, E. K. (2009). Utility of lymphoblastoid cell lines. *Journal of Neuroscience Research*, *87*(9), 1953–1959. https://doi.org/10.1002/jnr.22000

Smedley, D., Haider, S., Durinck, S., Pandini, L., Provero, P., Allen, J., … Kasprzyk, A. (2015). The BioMart community portal: an innovative alternative to large, centralized data repositories. *Nucleic Acids Research*, *43*(W1), W589–W598. https://doi.org/10.1093/nar/gkv350

Smemo, S., Tena, J. J., Kim, K.-H., Gamazon, E. R., Sakabe, N. J., Gómez-Marín, C., … Nóbrega, M. A. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, *507*(7492), 371–5. https://doi.org/10.1038/nature13138

Smith, R. P., Taher, L., Patwardhan, R. P., Kim, M. J., Inoue, F., Shendure, J., … Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. *Nature Genetics*, *45*(9), 1021–1028. https://doi.org/10.1038/ng.2713

Soutourina, J. (2017). Transcription regulation by the Mediator complex. *Nature Reviews Molecular Cell Biology*, *19*(4), 262–274. https://doi.org/10.1038/nrm.2017.115

Spain, S. L., & Barrett, J. C. (2015). Strategies for fine-mapping complex traits. *Human Molecular Genetics*, *24*(R1), R111–R119. https://doi.org/10.1093/hmg/ddv260

Spitz, F., & Furlong, E. E. M. (2012a). Transcription factors: from enhancer binding to developmental control. *Nature Reviews Genetics*, *13*(9), 613–626. https://doi.org/10.1038/nrg3207

Spivakov, M. (2014). Spurious transcription factor binding: Non-functional or genetically redundant? *BioEssays*, *36*(8), 798–806. https://doi.org/10.1002/bies.201400036

Spivakov, M., Akhtar, J., Kheradpour, P., Beal, K., Girardot, C., Koscielny, G., … Birney, E. (2012). Analysis of variation at transcription factor binding sites in Drosophila and humans. *Genome Biology*, *13*(9), R49. https://doi.org/10.1186/gb-2012-13-9-r49

Staller, M. V, Vincent, B. J., Bragdon, M. D. J., Lydiard-Martin, T., Wunderlich, Z., Estrada, J., & DePace, A. H. (2015). Shadow enhancers enable Hunchback bifunctionality in the Drosophila embryo. *Proceedings of the National Academy of Sciences of the United States of America*, *112*(3), 785–90. https://doi.org/10.1073/pnas.1413877112

Stefflova, K., Thybert, D., Wilson, M. D., Streeter, I., Aleksic, J., Karagianni, P., … Odom, D. T. (2013). Cooperativity and Rapid Evolution of Cobound Transcription Factors in Closely Related Mammals. *Cell*, *154*(3), 530–540. https://doi.org/10.1016/j.cell.2013.07.007

Stormo, G. D. (2000). DNA binding sites: representation and discovery. *Bioinformatics (Oxford, England)*, *16*(1), 16–23. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10812473

Stranger, B. E., Nica, A. C., Forrest, M. S., Dimas, A., Bird, C. P., Beazley, C., … Dermitzakis, E. T. (2007). Population genomics of human gene expression. *Nature Genetics*, *39*(10), 1217–24. https://doi.org/10.1038/ng2142

Stranger, B. E., & Raj, T. (2013). Genetics of human gene expression. *Current Opinion in Genetics & Development*, *23*(6), 627–634. https://doi.org/10.1016/j.gde.2013.10.004

Swamynathan, S. K., & Piatigorsky, J. (2002). Orientation-dependent Influence of an Intergenic Enhancer on the Promoter Activity of the Divergently Transcribed Mouse *Shsp* /αB-crystallin and *Mkbp* / *HspB2* Genes. *Journal of Biological Chemistry*, *277*(51), 49700–49706. https://doi.org/10.1074/jbc.M209700200

Taatjes, D. J., Näär, A. M., Andel, F., Nogales, E., & Tjian, R. (2002). Structure, Function, and Activator-Induced Conformations of the CRSP Coactivator. *Science*, *295*(5557), 1058–1062. https://doi.org/10.1126/science.1065249

Taatjes, D. J., Schneider-Poetsch, T., & Tjian, R. (2004). Distinct conformational states of nuclear receptor–bound CRSP–Med complexes. *Nature Structural & Molecular Biology*, *11*(7), 664–671. https://doi.org/10.1038/nsmb789

Taher, L., McGaughey, D. M., Maragh, S., Aneas, I., Bessling, S. L., Miller, W., … Ovcharenko, I. (2011). Genome-wide identification of conserved regulatory function in diverged sequences. *Genome Research*, *21*(7), 1139–49. https://doi.org/10.1101/gr.119016.110

Taverna, S. D., Li, H., Ruthenburg, A. J., Allis, C. D., & Patel, D. J. (2007). How chromatin-binding modules interpret histone modifications: lessons from professional pocket pickers. *Nature Structural & Molecular Biology*, *14*(11), 1025–1040. https://doi.org/10.1038/nsmb1338

Tehranchi, A. K., Myrthil, M., Martin, T., Hie, B. L., Golan, D., & Fraser, H. B. (2016). Pooled ChIP-Seq Links Variation in Transcription Factor Binding to Complex Disease Risk. *Cell*, *165*, 730–741. https://doi.org/10.1016/j.cell.2016.03.041

Tewhey, R., Kotliar, D., Park, D. S., Liu, B., Winnicki, S., Reilly, S. K., … Sabeti, P. C. (2016). Direct Identification of Hundreds of Expression-Modulating Variants using a Multiplexed Reporter Assay. *Cell*, *165*(6), 1519–1529. https://doi.org/10.1016/j.cell.2016.04.027

Thanos, D., & Maniatis, T. (1995). Virus induction of human IFN beta gene expression requires the assembly of an enhanceosome. *Cell*, *83*(7), 1091–100. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8548797

Thompson, C. M., Koleske, A. J., Chao, D. M., & Young, R. A. (1993). A multisubunit complex associated with the RNA polymerase II CTD and TATA-binding protein in yeast. *Cell*, *73*(7), 1361–75. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/8324825

Thurman, R. E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M. T., Haugen, E., … Stamatoyannopoulos, J. A. (2012). The accessible chromatin landscape of the human genome. *Nature*, *489*(7414), 75–82. https://doi.org/10.1038/nature11232

Tillo, D., Kaplan, N., Moore, I. K., Fondufe-Mittendorf, Y., Gossett, A. J., Field, Y., … Hughes, T. R. (2010). High Nucleosome Occupancy Is Encoded at Human Regulatory Sequences. *PLoS ONE*, *5*(2), e9129. https://doi.org/10.1371/journal.pone.0009129

Todeschini, A.-L., Georges, A., & Veitia, R. A. (2014). Transcription factors: specific DNA binding and specific gene regulation. *Trends in Genetics*, *30*(6), 211–219.

https://doi.org/10.1016/j.tig.2014.04.002

Tolhuis, B., Palstra, R.-J., Splinter, E., Grosveld, F., & de Laat, W. (2002). Looping and Interaction between Hypersensitive Sites in the Active β-globin Locus. *Molecular Cell*, *10*(6), 1453–1465. https://doi.org/10.1016/S1097-2765(02)00781-5

Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A., & Luscombe, N. M. (2009a). A census of human transcription factors: function, expression and evolution. *Nature Reviews Genetics*, *10*(4), 252–263. https://doi.org/10.1038/nrg2538

Venters, B. J., & Pugh, B. F. (2009). How eukaryotic genes are transcribed. *Critical Reviews in Biochemistry and Molecular Biology*, *44*(2–3), 117–41. https://doi.org/10.1080/10409230902858785

Vernimmen, D., Lynch, M. D., De Gobbi, M., Garrick, D., Sharpe, J. A., Sloane-Stanley, J. A., … Higgs, D. R. (2011). Polycomb eviction as a new distant enhancer function. *Genes & Development*, *25*(15), 1583–8. https://doi.org/10.1101/gad.16985411

Vieira, K. F., Levings, P. P., Hill, M. A., Crusselle, V. J., Kang, S.-H. L., Engel, J. D., & Bungert, J. (2004). Recruitment of Transcription Complexes to the β-Globin Gene Locus *in Vivo* and *in Vitro*. *Journal of Biological Chemistry*, *279*(48), 50350–50357. https://doi.org/10.1074/jbc.M408883200

Villar, D., Berthelot, C., Aldridge, S., Rayner, T. F., Lukk, M., Pignatelli, M., … Odom, D. T. (2015). Enhancer evolution across 20 mammalian species. *Cell*, *160*(3), 554–66. https://doi.org/10.1016/j.cell.2015.01.006

Visel, A., Blow, M. J., Li, Z., Zhang, T., Akiyama, J. A., Holt, A., … Pennacchio, L. A. (2009). ChIP-seq accurately predicts tissue-specific activity of enhancers. *Nature*, *457*(7231), 854–858. https://doi.org/10.1038/nature07730

Visser, M., Kayser, M., & Palstra, R.-J. (2012). HERC2 rs12913832 modulates human pigmentation by attenuating chromatin-loop formation between a long-range enhancer and the OCA2 promoter. *Genome Research*, *22*(3), 446–55. https://doi.org/10.1101/gr.128652.111

Wang, J., Malecka, A., Trøen, G., & Delabie, J. (2015). Comprehensive genome-wide transcription factor analysis reveals that a combination of high affinity and low affinity DNA binding is needed for human gene regulation. *BMC Genomics*, *16 Suppl 7*(Suppl 7), S12. https://doi.org/10.1186/1471-2164-16-S7-S12

Wang, K., Li, M., & Hakonarson, H. (2010). ANNOVAR: functional annotation of genetic variants from high-throughput sequencing data. *Nucleic Acids Research*, *38*(16), e164. https://doi.org/10.1093/nar/gkq603

Wang, L., Oberg, A. L., Asmann, Y. W., Sicotte, H., McDonnell, S. K., Riska, S. M., … Thibodeau, S. N. (2009). Genome-Wide Transcriptional Profiling Reveals MicroRNA-Correlated Genes and Biological Processes in Human Lymphoblastoid Cell Lines. *PLoS ONE*, *4*(6), e5878. https://doi.org/10.1371/journal.pone.0005878

Warner, J. R. (1999). The economics of ribosome biosynthesis in yeast. *Trends in Biochemical Sciences*, *24*(11), 437–40. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/10542411

Weinmann, R., & Roeder, R. G. (1974). Role of DNA-dependent RNA polymerase 3 in the transcription of the tRNA and 5S RNA genes. *Proceedings of the National Academy of Sciences of the United States of America*, *71*(5), 1790–4. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/4525293

Wen, X., Luca, F., Pique-Regi, R., Naranbhai, V., & Wong, D. (2015). Cross-Population Joint Analysis of eQTLs: Fine Mapping and Functional Annotation. *PLOS Genetics*, *11*(4), e1005176. https://doi.org/10.1371/journal.pgen.1005176

Werner, T., Hammer, A., Wahlbuhl, M., Bösl, M. R., & Wegner, M. (2007). Multiple conserved regulatory elements with overlapping functions determine Sox10 expression in mouse embryogenesis. *Nucleic Acids Research*. https://doi.org/10.1093/nar/gkm727

Westra, H.-J., & Franke, L. (2014). From genome to function by studying eQTLs. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, *1842*(10), 1896–1902. https://doi.org/10.1016/j.bbadis.2014.04.024

Wittkopp, P. J., Vaccaro, K., & Carroll, S. B. (2002). Evolution of yellow gene regulation and pigmentation in Drosophila. *Current Biology : CB*, *12*(18), 1547–56. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/12372246

Wong, E. S., Schmitt, B. M., Kazachenka, A., Thybert, D., Redmond, A., Connor, F., … Flicek, P. (2016). Interplay of cis and trans mechanisms driving transcription factor binding and gene expression evolution. https://doi.org/10.1101/059873

Woodcock, C. L. F., Safer, J. P., & Stanchfield, J. E. (1976). Structural repeating units in chromatin: I. Evidence for their general occurrence. *Experimental Cell Research*, *97*(1), 101–110. https://doi.org/10.1016/0014-4827(76)90659-5

Wunderlich, Z., Bragdon, M. D. J., Vincent, B. J., White, J. A., Estrada, J., & DePace, A. H. (2015). Krüppel Expression Levels Are Maintained through Compensatory Evolution of Shadow Enhancers. *Cell Reports*, *12*(11), 1740–7. https://doi.org/10.1016/j.celrep.2015.08.021

Xie, S., Duan, J., Li, B., Zhou, P., & Hon, G. C. (2017). Multiplexed Engineering and Analysis of Combinatorial Enhancer Activity in Single Cells. *Molecular Cell*, *66*(2), 285–299.e5. https://doi.org/10.1016/j.molcel.2017.03.007

Xiong, H. Y., Alipanahi, B., Lee, L. J., Bretschneider, H., Merico, D., Yuen, R. K. C., … Frey, B. J. (2015). RNA splicing. The human splicing code reveals new insights into the genetic determinants of disease. *Science (New York, N.Y.)*, *347*(6218), 1254806. https://doi.org/10.1126/science.1254806

Xiong, N., Kang, C., & Raulet, D. H. (2002). Redundant and unique roles of two enhancer elements in the TCRgamma locus in gene regulation and gammadelta T cell development. *Immunity*, *16*(3), 453–63. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/11911829

Yang, J., Ferreira, T., Morris, A. P., Medland, S. E., F Madden, P. A., Heath, A. C., … Visscher, P. M. (2012). Conditional and joint multiple-SNP analysis of GWAS summary statistics identifies additional variants influencing complex traits. *Nature Publishing Group*, *44*(4). https://doi.org/10.1038/ng.2213

Yuan, G.-C., Liu, Y.-J., Dion, M. F., Slack, M. D., Wu, L. F., Altschuler, S. J., & Rando, O. J. (2005). Genome-scale identification of nucleosome positions in S. cerevisiae. *Science (New York, N.Y.)*, *309*(5734), 626–30. https://doi.org/10.1126/science.1112178

Zabidi, M. A., Arnold, C. D., Schernhuber, K., Pagani, M., Rath, M., Frank, O., & Stark, A. (2014). Enhancer–core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, *518*(7540), 556–559. https://doi.org/10.1038/nature13994

Zanton, S. J., & Pugh, B. F. (2006). Full and partial genome-wide assembly and disassembly of the yeast transcription machinery in response to heat shock. *Genes & Development*, *20*(16), 2250–2265. https://doi.org/10.1101/gad.1437506

Zhang, D. X., & Glass, C. K. (2013). Towards an understanding of cell-specific functions of signal-dependent transcription factors. *Journal of Molecular Endocrinology*, *51*(3), T37-50. https://doi.org/10.1530/JME-13-0216

Zhang, Y., Wong, C.-H., Birnbaum, R. Y., Li, G., Favaro, R., Ngan, C. Y., … Wei, C.-L. (2013). Chromatin connectivity maps reveal dynamic promoter–enhancer long-range associations. *Nature*, *504*(7479), 306–310. https://doi.org/10.1038/nature12716

Zhao, J., Herrera-Diaz, J., & Gross, D. S. (2005). Domain-Wide Displacement of Histones by Activated Heat Shock Factor Occurs Independently of Swi/Snf and Is Not Correlated with RNA Polymerase II Density. *Molecular and Cellular Biology*, *25*(20), 8985–8999. https://doi.org/10.1128/MCB.25.20.8985-8999.2005

Zhou, J., & Troyanskaya, O. G. (2015). Predicting effects of noncoding variants with deep learning–based sequence model. *Nature Methods*, *12*(10), 931–934. https://doi.org/10.1038/nmeth.3547

Zinzen, R. P., Girardot, C., Gagneur, J., Braun, M., & Furlong, E. E. M. (2009). Combinatorial binding predicts spatio-temporal cis-regulatory activity. *Nature*, *462*(7269), 65–70. https://doi.org/10.1038/nature08531

Zuo, C., Shin, S., & Keleş, S. (2015). atSNP: transcription factor binding affinity testing for regulatory SNP detection. *Bioinformatics (Oxford, England)*, *31*(20), 3353–5. https://doi.org/10.1093/bioinformatics/btv328