# GENOME-WIDE DISCOVERY AND ANNOTATION OF HUMAN ENHANCERS RELEVANT TO DEVELOPMENT AND DISEASE

ZHANG JINGYAO

*B.Sc. (Hons.), Nanyang Technological University 2008*

A THESIS SUBMITTED FOR THE DEGREE OF

DOCTOR OF PHILOSOPHY

NUS GRADUATE SCHOOL FOR INTEGRATIVE SCIENCES AND ENGINEERING

NATIONAL UNIVERSITY OF SINGAPORE

2015

## DECLARATION

I hereby declare that this thesis is my original work and it has been written by me in its entirety. I have duly acknowledged all the sources of information which have been used in the thesis.

This thesis has also not been submitted for any degree in any university previously.

Zhang Jingyao

9th April 2015

## ACKNOWLEDGEMENTS

I would like to express sincere thanks to everyone who has helped me in one way or another throughout the past four years. Especially, I would like to thank my supervisors Dr Lim Bing and Dr Ruan Yijun, who have been a constant source of guidance, encouragement and inspiration. I would also like to express gratitude to my TAC members, Dr Axel Hillmer and Dr Melissa Fullwood, for their insightful feedback and suggestions.

This work would not have been possible without contributions from a group of outstanding colleagues. Heartfelt thanks go to Mr Kyle Loh, Dr Ang Lay Teng, Dr Vibhor Kumar and Dr Shyam Prabhakar for their kind help for the endoderm project; Dr Li Guoliang, Dr Kuljeet Singh, Dr Zheng Meizhen, Dr Wang Ping, Ms Poh Huay Mei, Ms Joanne Lim, Ms Elaine Chew, Ms Sia Yee Yen and Ms Choi Yeen Hui for working with me on the ChIA-PET project; Dr Cathleen Teh and Dr Vladimir Korzh for their guidance in the zebrafish experiments. I also wish to acknowledge the GTB sequencing team, especially Ms Lee Yen Ling, for technical assistance and advice on next-generation sequencing. I also thank the GIS for hosting my PhD study and A*STAR for providing the scholarship and funding for this work.

Last but not least I would like to thank my family members for being a constant source of love, support and encouragement for me throughout the years.

## TABLE OF CONTENTS

## SUMMARY

The complex, spatiotemporal gene expression patterns characteristic of multicellular organisms are controlled by sequence-specific *cis*-regulatory elements distributed throughout the genome. Among the different classes of elements, enhancers are recognized as key drivers of cell-type specific transcription programs, serving as an integrative binding platform for both lineage-specific TFs and external signaling effector TFs. Due to this central role in gene regulation, disruption of enhancer function can lead to disease, and much interests have been focused on enhancer discovery and annotation in model cell lines (Bernstein et al., 2012). Despite these efforts, our understanding of enhancer activation dynamics, especially during embryonic development, remains incomplete. The annotation of all enhancers during cell differentiation and lineage commitment will help us dissect the complex patterns of developmental gene expression and understand the basis of gene dysregulation resulting in disease.

The endoderm is the inner germ layer of the embryo which gives rise to the epithelial lining of the digestive and respiratory system, as well as components of the liver, pancreas, thyroid and thymus. Despite the physiological importance of these organs, relatively little is known about how endoderm progenitor cells give rise to all differentiated derivatives. To begin addressing this knowledge gap, we employed an *in vitro* endoderm differentiation model for transcriptome and epigenome profiling. Our transcriptome profiling of endoderm and its derivatives at defined developmental stages provided a valuable resource for the investigation of novel markers during lineage specification. Epigenomic and functional annotation of these enhancers further revealed the coordination between lineage-specifying TFs and signaling effectors for endoderm differentiation, as well as a diversity of enhancer priming states. Importantly, we also

demonstrated the importance of these enhancer catalogs in facilitating the identification of causal variants of complex diseases relevant to endoderm organs.

One defining characteristic of enhancers is expression regulation through a distance-independent manner, which confounds target gene identification and enhancer functionalization. Active enhancers are recognized to regulate gene expression through physical interactions with their target loci through the looping-out of intervening DNA sequences. To map global enhancer-promoter interactions, I was involved in a large-scale ChIA-PET study, focusing on the general transcription factor RNAPII in multiple human cell types. This work not only revealed widespread cell-specific enhancer-promoter and promoter-promoter interactions, but also identified interactions involving disease-associated regulatory elements with their target genes.

In sum, the enhancer repertoire uncovered in this work represents a valuable resource for the study of human endoderm formation and patterning. In addition to elements associated with known endoderm genes, we have predicted thousands of developmental enhancers whose regulatory functions were previously unknown. Our work on chromatin organization in human cells have also revealed previously unappreciated organizational complexity between regulatory elements and target promoters for transcription control.

## LIST OF TABLES

## LIST OF FIGURES

## LIST OF ABBREVIATIONS

| | |
|---|---|
| 3C | Chromatin conformation capture |
| 5C | Chromatin conformation capture carbon copy |
| AFG | Anterior foregut |
| BLAST | Basic local alignment search tool |
| ChIA-PET | Chromatin Interaction Analysis by Paired-End Tag |
| ChIP | Chromatin immunoprecipitation |
| ChIP-seq | ChIP-sequencing |
| DE | Definitive endoderm |
| DHS | DNase I hypersensitive site |
| ENCODE | Encyclopedia of DNA elements |
| eRNA | Enhancer RNA |
| ES | Embryonic stem cell |
| FAIRE | Formaldehyde-Assisted Isolation of Regulatory Elements |
| GWAS | Genome-wide association study |
| GSEA | Gene set enrichment analysis |
| hESC | Human embryonic stem cell |
| HS | Hypersensitive |
| Ig | Immunoglobulin |
| LncRNAs | Long noncoding RNAs |
| MHG | Mid/hindgut |
| NHGRI | National Human Genome Research Institute |
| PCR | Polymerase chain reaction |

| | |
|---|---|
| PFG | Posterior foregut |
| PIC | Pre-initiation complex |
| PS | Primitive streak |
| qPCR | quantitative-PCR |
| RNAPII | RNA-polymerase II |
| RNA-seq | RNA-sequencing |
| SNP | Single nucleotide polymorphism |
| TF | Transcription factor |
| TRE | Transcription regulatory element |
| TSS | Transcription start site |
| ZED | Zebrafish enhancer detector |

# CHAPTER 1 INTRODUCTION

## 1.1 The human genome sequence

The completion of the initial human genome sequence in 2001 represented a significant milestone in biology, human genetics and biomedical research (Lander et al., 2001; McPherson et al., 2001). Subsequent refinement and analyses of the draft sequence led to a comprehensive identification and mapping of 20,000 – 25,000 protein-coding genes (Consortium, 2004), signaling the beginning of an era of "omics" revolution from the traditional gene-centric paradigm. Equipped with this extensive catalogue, genetics researchers now face immense challenges in understanding the functions of these genes and their products, as well as their regulation and coordination at a cellular and organismal level. An added layer of complexity originated from the presence of vast stretches of poorly characterized, non-coding genomic regions within which protein-coding regions (1.9% of entire human genome) reside (Lander et al., 2001). The systematic identification of all functional elements in the human genome represented a crucial requisite for comprehensive understanding of the spatial and temporal expression patterns for all identified genes, as well as the basis for altered gene expression during pathological conditions.

## 1.2 Eukaryotic transcriptional machinery and transcriptional regulatory elements (TREs)

Transcription of eukaryotic genes is dependent on three known classes of DNA-binding proteins, namely basic TFs, activators, and coactivators. Prior to each round of transcription, more than 30 basic TFs, consisting of RNAPII and a variety of accessory factors, including TFIIA, TFIIB, TFIID, TFIIE, TFIIF and TFIIH, assemble on the core promoter as a Pre-initiation Complex (PIC), constituting the "basal transcription machinery" (Conaway and Conaway, 1993; Murakami et al., 2013). On its own, the PIC initiates low levels of basal transcriptional activity, while full transcription require the presence of sequence-specific activator TFs. These activators can be classified by their DNA-binding domains, which may include basic leucine zipper (bZIP), homeobox, forkhead or helix-loop-helix (HLH) domains (Pabo and Sauer, 1992). Activator function is mediated by a transcription activation module, which is thought to facilitate PIC formation through protein-protein interactions with basic TFs (Orphanides et al., 1996). One other mechanism for activator function involves coactivator recruitment through protein-protein interactions. Coactivators exist as large complexes and modulate activator function through interactions with the basic transcriptional machinery (i.e. Mediator coactivator complex) (Malik and Roeder, 2000) or through chromatin remodeling systems (i.e. SWI/SNF complex) (Neely et al., 1999).

Components of the eukaryotic transcription machinery mediate gene expression by RNAPII through binding to specific nucleotide sequences called TREs. These elements include promoters and gene-distal elements such as enhancers, silencers, insulators and locus control regions (Fig 1.1). Generally, promoters function as the site for PIC assembly, while gene-distal elements harbor sequence motifs for activator TFs. The physical and functional properties of each of these elements are discussed further below.

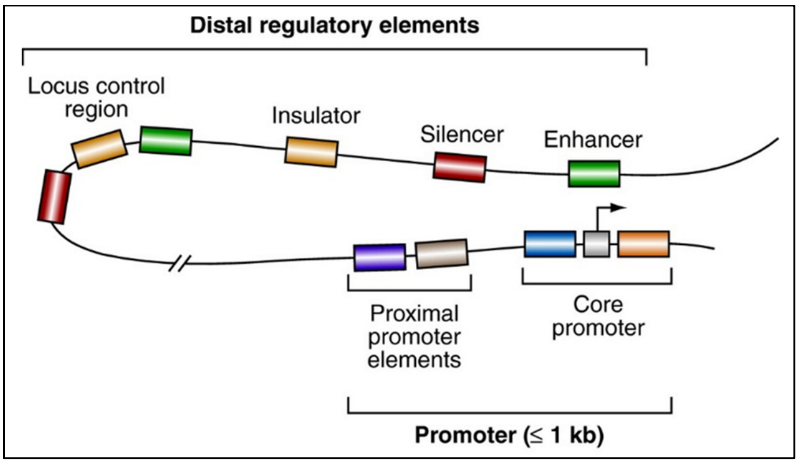


**Figure 1.1 Transcriptional regulatory elements in eukaryotic genomes (Maston et al., 2006)** TREs can be generally classified as promoter/promoter-proximal or distal element. These elements function through binding to distinct elements of the transcription machinery.

### 1.2.1 Promoters

Core promoters are located at the transcriptional start site (TSS) of genes. They are required and sufficient for transcription initiation by RNAPII through assembly of the PIC. Although considered as a single class of regulatory elements, core promoters are structurally and functionally diverse, harboring different types of Core Promoter Elements (CPEs), including the TATA box, Initiator element (Inr), Downstream Promoter Element (DPE), Downstream Core Element (DCE), Motif Ten Element (MTE), TFIIB-Recognition Element (BRE) and CpG islands (Kadonaga, 2012). The diversity of CPEs has been proposed to provide combinatorial regulation of transcription initiation, increasing the number of possible gene expression patterns in complex organisms (Gershenzon and Ioshikhes, 2005; Smale and Kadonaga, 2003). Sequence analyses revealed that CPEs are not universal across promoters – a large number of functional core promoters lack any of these known CPEs, suggesting the existence of other CPEs which are yet to be discovered (Hartmann et al., 2013).

Located up to several hundred base pairs upstream of the core promoters are the proximal promoters. These elements have been demonstrated to regulate transcription through binding of specific TFs (Hock et al., 2004; Landry et al., 2005). A functional link between proximal promoters and distal enhancers was first suggested by the finding that these two classes of elements activate transcription through binding to the same ‘general’ activation protein domains (Seipel et al., 1992). In one example, promoter deletion studies showed that a proximal promoter of the glucagon receptor gene functioned as a *cis*-acting enhancer and regulates constitutive gene expression (Geiger et al., 2001). Besides harboring intrinsic enhancer activities, proximal promoters also interact with distal enhancers and mediate transcription in a synergistic manner

(Grzeskowiak et al., 2000; Vorachek et al., 2000; Wood et al., 1998). Due to these properties, proximal promoters are considered functionally indistinguishable from enhancers. Indeed, recent computational analyses revealed sequence features of proximal promoters which reliably predicted distal enhancers (Taher et al., 2013).

### 1.2.2 Enhancers

Transcriptional enhancers are short, non-coding regulatory elements responsible for activating target gene expression during differentiation and development (Visel et al., 2009a). Enhancer activity was first documented over 30 years ago within a region of the SV40 virus genome which, when cloned and introduced into a human cell line, increased transcription of a reporter gene by several hundred folds (Banerji et al., 1981). Subsequent to this seminal discovery, the first mammalian enhancers, driving cell-type specific expression of the Ig heavy chain gene, were identified (Banerji et al., 1983; Gillies et al., 1983). Due to a unique property of enhancers to function independent of the distance to their target gene as well as the underlying chromatin context, enhancers can be 'hijacked' through an enhancer trap, allowing unbiased identification and screenings for spatial enhancer activities (Korzh, 2007). Leveraging on this, enhancer trap assays in *Drosophila* demonstrated that distal *cis*-regulatory elements drive spatial and temporal gene expression patterns during development (O'Kane and Gehring, 1987). More recently, studies employing high-throughput genomic profiling of TF-binding, histone modifications and open chromatin suggested the presence of hundreds of thousands of enhancers across multiple cell types, often residing far from their target promoters (Bernstein et al., 2012; Ernst et al., 2011; May et al., 2011; Rada-Iglesias et al., 2011). These genome-wide observations led to the current view of enhancer activation as a highly dynamic and cell-type-specific process, and raised several key mechanistic questions regarding how enhancers function in TF binding and how binding information is relayed to target promoters.

#### 1.2.2.1 Enhancer architecture

A general feature of enhancers is the enrichment of clusters of functional TF binding sites in various combinations (Arnone and Davidson, 1997; Berman et al., 2002; Gotea et al., 2010; Levy et al., 2001), which is a critical molecular mechanism to ensure robustness of TF recruitment for tight gene expression regulation, while preventing unwanted activation through randomly occurring binding sites. Detailed studies of enhancer architecture in *Drosophila* and zebrafish have revealed the existence of regulatory modules, each containing one or several TF binding sites, within each enhancer (Arnosti et al., 1996; Kulkarni and Arnosti, 2003; Liu and Posakony, 2012; Rastegar et al., 2008). These regulatory modules function as separate functional units, occupy flexible positions within an enhancer and can independently regulate gene expression. These observations have led to the 'billboard' model, where enhancers function as 'information display' platforms with the overall functional output of an enhancer depends on net sum of all modules contained within (Fig 1.2A). This model, established through independent studies of individual enhancers, was further supported by a massively parallel reporter assay which tested almost 5,000 synthetic enhancers (Smith et al., 2013). The organizational flexibility underlying the billboard model has been suggested to confer evolutionary flexibility to enhancers, as demonstrated by sequence and binding site divergence of the *even-skipped* enhancer between the Sepsidae and Drosophilidae species (Hare et al., 2008), and may explain the weak sequence conservation of certain human enhancers (Blow et al., 2010).

Despite strong experimental support for flexible motif positioning within a large number of enhancers, strict motif positioning, including spacing, order and orientation, has been observed in other enhancers (Senger et al., 2004). One of the best-characterized of these is the interferon-β

(IFN-β) enhancer (Thanos and Maniatis, 1995), where small sequence changes to individual motifs, or the spacing between them, is sufficient to impair binding of all eight known factors binding the enhancer (Fig 1.2B) (Panne et al., 2007). The striking sequence constraint at this locus led to an alternative model, the 'enhanceosome' model of enhancer architecture (Merika and Thanos, 2001), where individual TFs assemble in a cooperative  and highly ordered manner relative to each other, forming unique, stable complexes or 'enhanceosomes'. This model can explain why presence of individual TFs do not activate *IFN-β* transcription (Thanos and Maniatis, 1995). The requirement for cooperative TF binding leads to a sharp, "all-or-none" activation effect, characteristic of rapid biological processes such as the immune response. Indeed, enhanceosomes assembly have been reported for the activation of other immune-related genes, such as interleukin-6 and interleukin-2 receptor α (Vanden Berghe et al., 1999; John et al., 1995). Unlike the enhancers functioning under the 'billboard' model, the strict sequence constraints at enhanceosome binding sites make these enhancers more susceptible to inactivating mutations (Panne et al., 2004, 2007).

Accumulating evidence suggests that both the billboard and enhanceosome models may represent two extreme ends of a spectrum of architectural diversity, and that no generalizable motif rules exist when large numbers of enhancers are examined. For example, during cardiac specification in *Drosophila*, five cardiac developmental TFs co-bind bind and activate heart enhancers in the absence of any consistent motif patterns (Junion et al., 2012). In the absence of one TF, all remaining TFs failed to mediate enhancer activation, emphasizing the importance of cooperative binding. This third model, termed 'TF collective', emphasizes a critical role of protein-protein interactions in determining overall enhancer activities, beyond that of a linear nucleotide-based code.

The increasing availability of high-throughput reporter assays (Levo and Segal, 2014) enabling quantitative output measurement of large numbers of enhancers may further challenge these existing models founded on a principled understanding of enhancer function. The current research focus on enhancer architecture lies in gaining further mechanistic understanding beyond the description of specific enhancers. This is exemplified by current research focus moving from simple models of DNA binding codes to higher levels of protein DNA interactions, including protein side-chain flexibility, DNA shape-readout, docking geometries and protein allosteric effects (Siggers and Gordân, 2014).

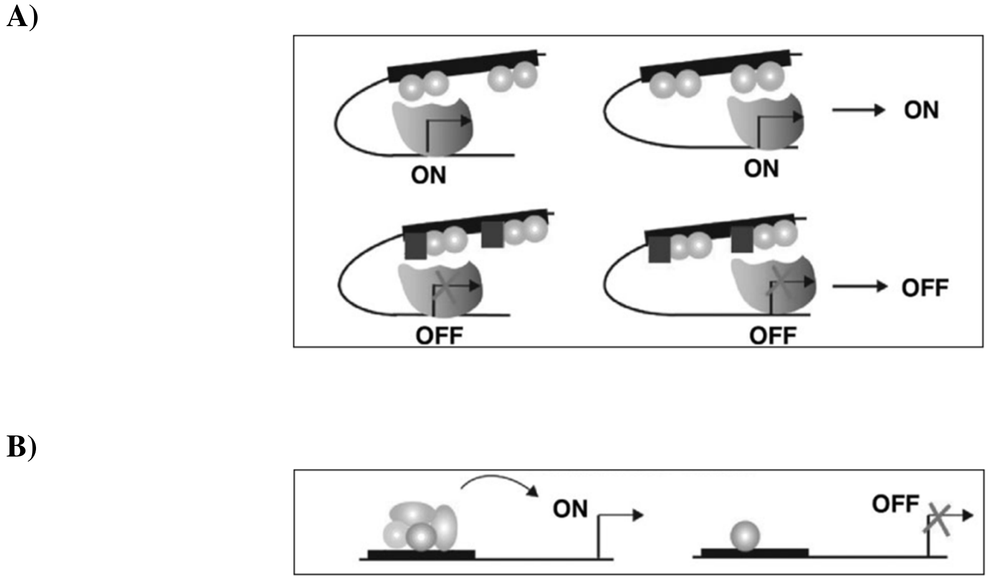


**Figure 1.2 Two models of enhancer function (Arnosti and Kulkarni, 2005)**

A) In the billboard model, regulatory modules within enhancers function independently and occupy flexible positions. Enhancer cooperativity is not necessary, and net enhancer output depends on net effect of individual interactions between TFs (grey circles) and repressors (black boxes) and the basal transcription machinery, resulting in transcription activation (top) or repression (bottom).

B) In the enhanceosome model, TFs (grey ovals and circles) bind to a strictly defined motif structure, where all TFs assemble cooperatively for gene activation. Disruption of motif sequence, or displacement of a single motif resulting in lack of a single TF, causes the enhancer to be inactive.

#### 1.2.2.2 Enhancer-promoter interactions

The highly organized TF binding patterns on enhancers raised further questions regarding how binding information is relayed to target promoters, and several models have been proposed to address this question (Fig 1.3). Binding of sequence-specific TFs have been proposed to induce the transcription complex to 'track' from the distal enhancer, through small steps along the intervening DNA, until it encounter its target promoter (Bulger and Groudine, 2011). This model has only been tested on a few enhancers and requires more detailed validation. Focused investigations at several independent loci suggest that enhancers mediate long-range chromatin loops through various mechanisms which lead to the juxtaposition of enhancer and promoter elements. Through chromatin conformation capture (3C) and gene knockdown assays, these studies have demonstrated critical roles for lineage-specific TFs in mediating chromatin loop formation and subsequent target gene expression (Yun et al., 2014; Zhang et al., 2013a; Zhou et al., 2013). A novel insight in enhancer function was illustrated by a recent finding that enhancer looping mediated by the LIM domain-binding protein 1 (LDB1) may simultaneously enhance target gene expression and repress other target genes at the level of promoter pausing in pituitary corticotrope cells (Zhang et al., 2015).

An alternative enhancer–promoter looping mechanism involving chromatin architectural proteins was supported through the genome-wide observation that CTCF binding sites significantly overlap enhancer elements (Shen et al., 2012). This mechanism was subsequently supported by findings of CTCF-mediated chromatin looping at multiple individual gene loci  (Eldholm et al., 2014; Gosalia et al., 2014; Majumder and Boss, 2010), as well as the PCDH gene cluster involving more than 50 different exons (Golan-Mashiach et al., 2012). The finding that CTCF directly recruits TATA-binding protein-associated factor 3 (TAF3) to promoter distal sites for DNA looping suggests that CTCF may directly interact with the core transcription machinery complex to tether enhancers to their target promoters (Liu et al., 2011).

A third mechanism through which enhancers mediate chromatin looping involves the transcription of non-coding RNAs (ncRNA) on enhancers, termed enhancer RNAs (eRNAs) (Kim et al., 2010; Wang et al., 2011). Although eRNAs are transcribed from a large proportion of enhancers, the mechanistic details of their functions, if any, remained unclear. Antisense oligonucleotide blockage of eRNAs at both the *NRIP1* and *GREB1* loci resulted in disruption of enhancer-promoter interactions with concomitant reduction in target gene expression (Li et al., 2013). Similarly, knockdown of the enhancer-assocation ncRNA *CCAT1-L* resulted in reduced interaction frequency between *CCAT1-L* locus and decreased *MYC* transcription in colorectal cancer cells (Xiang et al., 2014), strengthening support for a role of eRNAs in physically tethering enhancer-promoter loops. Collectively these studies highlight the versatility of enhancers in harnessing TFs, architectural proteins and ncRNAs as mechanistic tools to transmit regulatory information to their target genes.

Enhancer regulation of gene expression may occur at different stages of transcription, including transcription initiation, elongation or termination. The direct involvement of components of the general transcription machinery in enhancer-promoter loop formation suggests that regulation occurs at the level of transcription initiation (Koch et al., 2011; Liu et al., 2011; Ren et al., 2011). In support, the mediator complex co-occupies ESC enhancers with the pluripotency factors NANOG, OCT4 and SOX2 and facilitate loop formation to RNAPII through direct interactions with cohesin (Kagey et al., 2010). More recent work have revealed a novel role of enhancer loops in regulating promoter-proximal pause release and led to a new class of enhancers termed ‘anti-pause’ enhancers (Liu et al., 2013). Looping of these enhancers result in activation of the P-TEFb complex and release of RNAPII for elongation. Global chromatin contact maps during *Drosophila* embryogenesis further revealed that enhancer-promoter contacts remained similar across developmental stages and frequently associated with paused RNAPII, suggesting release of paused polymerase as a key transcriptional mechanism facilitating rapid activation of developmental genes (Ghavi-helm et al., 2014). These studies provided further mechanistic insights into enhancer-target promoter regulation.

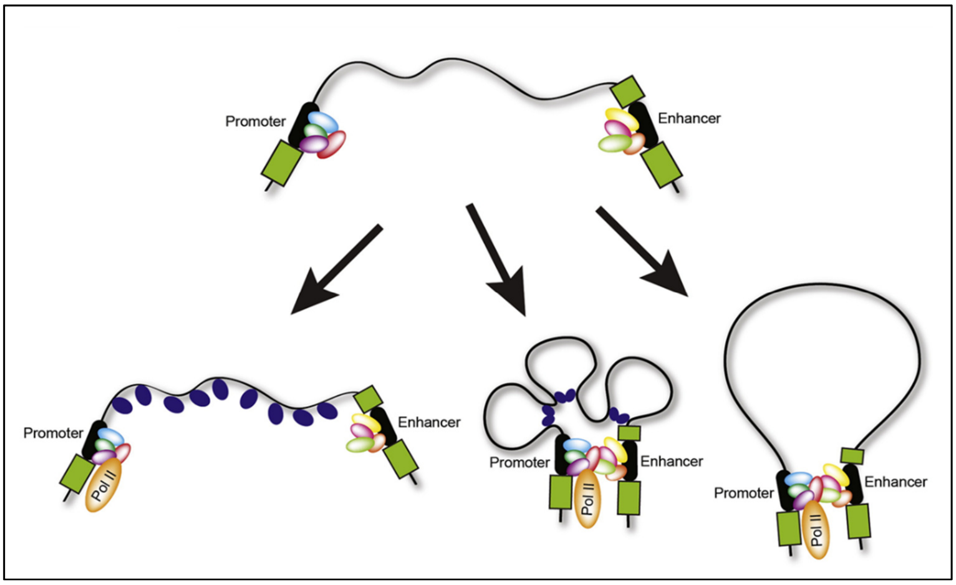


**Figure 1.3 Proposed model for enhancer function (Williamson et al., 2011)**

Activator TFs binding to enhancers may recruit additional co-activators, which reorganize chromatin along the intervening region facilitating tracking of the transcription complex towards the target genes (bottom left). Alternatively, spatial colocalization between enhancers and target promoters may be facilitated through miniloops (bottom middle) or a single loop (bottom right).

### 1.2.2.3 Super-enhancers

A set of studies focusing on enhancer discovery through chromatin modifications and TF binding proposed a unique class of TREs termed 'super-enhancers' (Hnisz et al., 2013; Lovén et al., 2013; Whyte et al., 2013). These studies suggest that a small subset of enhancers act as key regulators of cell fate, cell-type-specific gene expression and important drivers of oncogenic progression. Super-enhancers tend to span large genomic distances and generally have a median size of an order of magnitude larger than normal enhancers. Super-enhancers can be identified through the 'stitching' of adjacent enhancers and assessment of Med1 enrichment levels (Whyte et al., 2013). Since the initial proposal of the super-enhancer concept, various publications have employed differing defining criteria for super-enhancers. For example, Med1, H3K27ac and the master TF MyoD have each been used for super-enhancer identification (Hnisz et al., 2013; Lovén et al., 2013). To date, the only consistent feature of super-enhancers is the exceptionally high TF or histone modification enrichment levels revealed by ChIP-seq assays.

Despite the lack of a clear criterion for super-enhancer identification, the current definitions have revealed several important properties of these elements. Genomic profiles of lineage-specific TFs and H3K27ac have revealed widespread prevalence of super-enhancers in over 90 different cell types and tissues (Hnisz et al., 2013; Whyte et al., 2013). In a broad range of cancer cells, super-enhancers were found to be specifically enriched at multiple oncogenes (Hnisz et al., 2013; Lovén et al., 2013). Notably, binding of the BRD4 transcriptional coactivator at super-enhancers and expression of associated oncogenes were specifically impaired by Brd4 inhibition, highlighting a potential mechanism in which bromodomain inhibitors achieve cancer therapeutic effects through super-enhancers (Chapuy et al., 2013). Like enhancers, super-enhancers are enriched for disease-associated variants. This was illustrated for super-enhancers found in human brain tissues and immune cells, which were associated with SNPs associated with Alzheimer's disease and rheumatoid arthritis respectively (Hnisz et al., 2013)

Given that super-enhancers are identified through the clustering of individual enhancers into a single broad element, it is unclear whether the functions observed for super-enhancers are in fact a reflection of individual enhancers used to define super-enhancers. The assessment of cooperativity between individual enhancers may reveal whether super-enhancers function more than the sum of its parts. To date, this possibility has not been conclusively tested and remains a major goal of the super-enhancer concept. One confounding technical problem lies in the difficulty to accurately quantify small differences in TF binding through ChIP-seq datasets (Park et al., 2013; Teytelman et al., 2013). Despite contrasting views on whether super-enhancers provide any novel insights on enhancer function, blockage of super-enhancer function through inhibition of general TF or coactivator binding led to disruption of oncogene expression ion tumor cells, highlight super-enhancers as a potential therapeutic target for selective inhibition of oncogenic transcriptional programs and cancer progression.

### 1.2.3 Silencers

Silencers, as the name suggests, repress gene transcription. Like enhancers, silencers are sequence-specific elements and function in a distance- and orientation-independent manner, although location-dependent silencers have been reported (Dong and Lim, 1996; Moffat et al., 1996). Silencers can be found on promoters, intergenic regions, or within introns. They function through the recruitment of repressive TFs called repressors, which exert a negative influence on transcription through several mechanisms. Repressors targeted to the core promoter can directly repress RNAPII through post-translational modifications (Hengartner et al., 1998), or indirectly through blockage of basal TF binding (Li and Manley, 1998). Repressors can also antagonize activator binding: the BCL-6 repressor blocks its target promoters through competition with activator TFs for proximal promoter binding (Harris et al., 2005). The mechanisms of repressor function are broadened through the recruitment of co-repressor complexes (Privalsky, 2004; Reynolds et al., 2013). These complexes, such as the Polycomb Group proteins, repress transcription through chromatin compaction by modifying histones (Srinivasan and Atchison, 2004). More recent genome-wide analyses of repressor and co-repressor binding sites revealed widespread binding of repressors on actively transcribed regions of the genome (Wang et al., 2009), raising the possibility that repressors may in fact function as activators which prime or fine-tune transcription levels (Dovey et al., 2010).

### 1.2.4 Insulators

The specificity of the long-range regulatory effects of enhancers and silencers are enforced by a unique class of elements, called insulators, which partition the genome into discreet domains of expression (Valenzuela and Kamakaka, 2006). Insulators function either as barrier elements, preventing the spread of repressive chromatin domains, or enhancer blockers, disrupting enhancer-promoter communication (Fig 1.4). Sequences involved in insulator function typically resided between compact and de-condensed regions of the chromosome, and were traditionally identified through functional tests involving the integration of reporter constructs into cell lines or transgenic animals (Chung et al., 1993; Kellum and Schedl, 1991). These efforts led to the identification of several insulator binding proteins in *Drosophila*, such as BEAF, Zw5 and CTCF (Gaszner et al., 1999; Roy et al., 2007; Zhao et al., 1995). In contrast, the only *Drosophila* insulator protein with a vertebrate orthologue is CTCF (Bell et al., 1999). Contrary to its classical role as an insulator protein, recent 5C analyses revealed CTCF enrichment on looping interactions between active distal enhancers and promoters, and that 79% of all distal enhancer-promoter interactions are not blocked by intervening CTCF binding sites (Sanyal et al., 2012). At individual gene loci, CTCF has also been demonstrated to regulate complex gene clusters with distal regulatory elements (Golan-Mashiach et al., 2012; Xu et al., 2011). Collectively, these studies are leading to an emerging concept of CTCF as an architectural protein, establishing genome topology and mediating interactions between regulatory elements and their target genes (Ong and Corces, 2014).

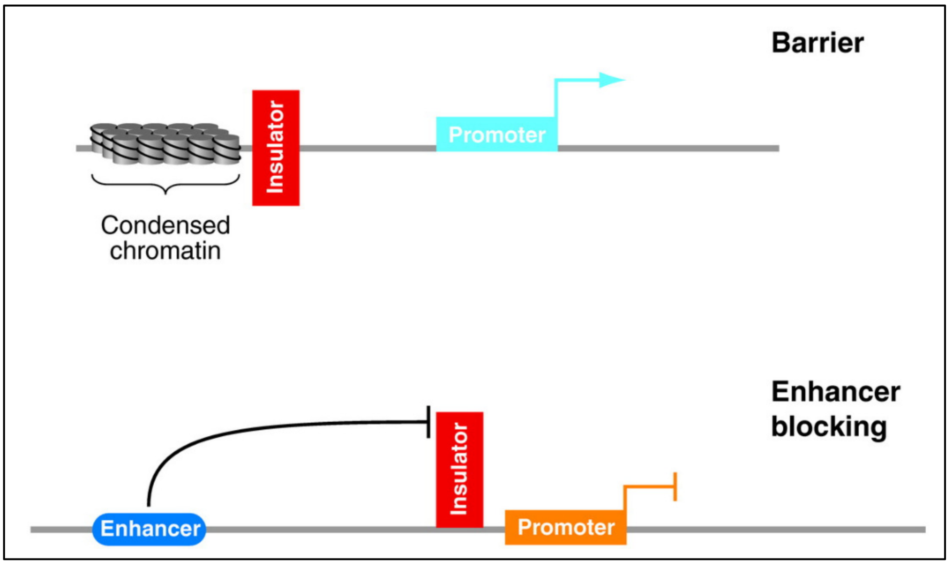


**Figure 1.4 Mechanisms of insulator function (Valenzuela and Kamakaka, 2006)**

Barrier insulators prevent the spread of heterochromatin regions, protecting target promoters from silencing (Top). Enhancer blocking focuses enhancer function to a single locus (bottom).

### 1.2.5 Locus Control Regions (LCRs)

LCRs are composite regulatory elements which regulate a gene cluster or locus, directing physiological and tissue-specific gene expression. (Li, Kenneth R., X. Fang, 2002). LCRs consist of multiple *cis*-acting elements, including enhancers, insulators and silencers, which can be identified by a cluster of DNase-hypersensitive (HS) sites. Each HS site may exhibit variable TF binding, histone modifications and chromatin interactions (Kim et al., 2012), resulting in variable effects on gene expression between each element. The overall function of an LCR depends on the collective activities of individual HS sites, although most LCRs exhibit strong enhancer activity. LCR-associated elements can also operate in an additive or synergistic manner (Engel and Tanimoto, 2000). The different combinations of elements thus confer LCRs great diversity and properties in driving tissue-specific expression.

The β-globin LCR was the first LCR to be identified, consisting of a 16kb genomic span with 5 HS sites driving erythrocyte-specific expression (Grosveld et al., 1987). Since then, LCRs have been discovered as key drivers for a large number of mammalian gene clusters, including the T cell receptor, apolipoprotein and immunoglobulin loci (Li, Kenneth R., X. Fang, 2002). Because of their central role in transcriptional control, mutations in LCRs have been implicated in human diseases, such as thalassemia (Driscoll et al., 1989).

## 1.3 Enhancers in development

Enhancers are the most abundant and variable type of TREs between different cell types, and are recognized as the primary drivers of cell-type specific gene expression patterns (Heintzman et al., 2009; Parker et al., 2013; Xi et al., 2007). The dynamics of such cell-type specific gene regulation is particularly evident during early embryonic development, where gene expression patterns and cell fate decision outcomes depend on the integration of multiple signaling pathways. Because the signaling effectors of these pathways occupy enhancers together with cell-specific master regulatory TFs (Trompouki et al., 2011), enhancers have been considered "information integration hubs" (Buecker and Wysocka, 2012), where external signals, regulatory TFs and genomic sequence information converge to establish the complex, dynamic expression patterns during development.

During early embryonic development, pluripotent and multipotent stem cells are exposed to multiple extrinsic signaling pathways, which guide these progenitors to undergo self-renewal or specify them towards specific cell fates (Pera and Tam, 2010). These signaling pathways evoke different cellular responses not only through induction of tissue-specific TF expression, but also by differential enhancer activation. For example, during hepatocyte differentiation, Hippo signaling mediates HNF4A and FOXA2 to bind and activate distinct sets of enhancers, allowing these master liver TFs to express different genes and fulfill distinct roles during cell differentiation and organ development (Alder et al., 2014). Such developmental gene regulation is also evident at the tissue and organ level, where distinct sets of master TFs in different *Drosophila* appendages differentially activate a shared set of enhancers (McKay and Lieb, 2013).

In order for enhancer activation to occur in a timely and dynamic manner during development, enhancer specification occurs early during development before their target genes are expressed. Early observations through *in vivo* footprinting analyses revealed that FoxA and GATA factors, required for *Alb1* gene expression in liver, bound the *Alb1* enhancer in endoderm cells and were required for subsequent *Alb1* enhancer function (Bossard and Zaret, 1998; Gualdi et al., 1996; Liu et al., 1991). Because chromatin occupancy by these factors preceded other TFs in liver, they were termed 'pioneer factors'. Not only were pioneer factors characterized by the timing of DNA binding, they were able to bind condensed chromatin. Subsequently, pioneer factors in other cell types were discovered, such as PU.1 and RUNX1 in B cells, and FoxD3, Sox2 and Sp1 in ES cells (Raghu Ram and Meshorer, 2009). Pioneer factor binding on enhancers may facilitate rapid transcriptional activation by reducing the number of binding events required later in development (Fig 1.5). Alternatively, pioneer factors may also actively organize the local chromatin to facilitate binding of other regulatory TFs. For example, the pioneer factors FoxA1, GATA-4 and TFE3 can generate local DNase hypersensitivity in chromatin (Cirillo et al., 2002; Ishii et al., 2004). In addition to pioneer factor binding, an additional property of many developmental enhancers is that they are held in a primed state, marked by H3K4me1 and H3K37me3, but devoid of H3K27ac, in ES cells (Rada-Iglesias et al., 2011). Upon differentiation, these 'poised' enhancers acquire H3K27ac and were associated with gene activation. Subsequently, H3K9me3 was also identified as a mark of poised enhancers independent of H3K27me3. A poised enhancer state thus allows for timely expression of key developmental genes upon signaling and developmental cues.

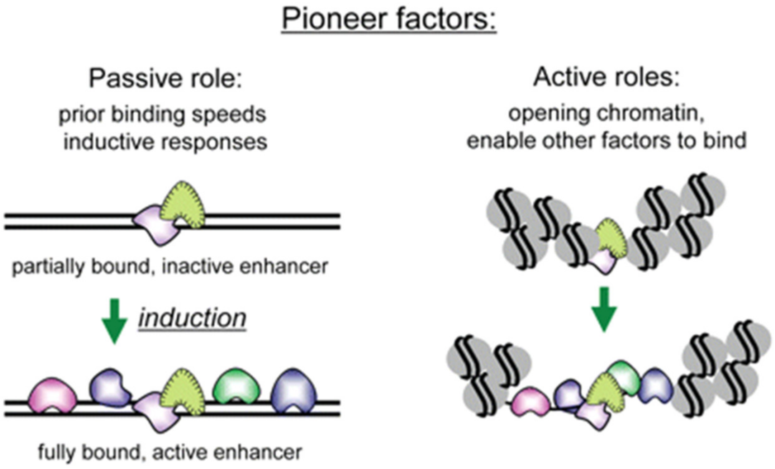

Pioneer factors:
Passive role:
prior binding speeds inductive responses
partially bound, inactive enhancer
induction
fully bound, active enhancer
Active roles:
opening chromatin, enable other factors to bind

**Figure 1.5 Active and passive roles of pioneer factors (Zaret and Carroll, 2011)**

Pioneer factors may function passively, facilitating enhancer activation through prior binding which reduces the number of binding events required subsequently. Alternatively, pioneer factors can actively remodel the underlying chromatin structure, indirectly allowing other regulators to bind.

## 1.4 TREs and enhancers in diseases

Given the pivotal role of TREs in developmental gene control, it is not surprising that a large number of mutations in TREs have been linked to human diseases. According to data compiled by the Human Gene Mutation Database (www.hgmd.cf.ac.uk), a total of 3,024 regulatory mutations have been identified, as of Oct 2014, which underlie or are associated with human inherited diseases. These documented mutations are mainly found on promoters, and their corresponding target genes and underlying defects are well-defined. Table 1 illustrates a list of classic Mendelian disorders resulting from TRE mutations. Such mutations disrupt basic TF or activator binding which are required for transcription through RNAPII, resulting in significant changes in expression levels of a single gene. Although most mutations increase disease risk through reduction of transcription, some mutations may also increase gene transcription. For example, mutation of a *COLIA1* regulatory element increases TF binding affinity and *COLIA1* transcription, resulting in an altered ratio of transcribed collagen chains leading to reduced bone mineral density (Grant et al., 1996). Besides promoters, several mutations in distal enhancers were identified to be associated with disease phenotypes. For example, the blood disorder β-thalassemia, a result of dysregulation of the β-globin gene, can be caused by translocations which remove a distal enhancer required for high-level expression of the β-globin gene in erythroblasts (Driscoll et al., 1989) (Table 1). In another example, point mutations or translocations of an upstream enhancer of the sonic hedgehog (*SHH*) gene, a regulator of limb and brain development, was found to be associated with inherited preaxial polydactyly (Lettice et al., 2003).

In contrast to monogenic Mendelian diseases, many common human disorders exhibit a more complex inheritance pattern, and disease susceptibility depends on the net effect of interactions between multiple genes and may be influenced by environmental factors. To understand the genetic changes underlying common disease susceptibility, it is useful to examine common genetic variation, such as Single Nucleotide Polymorphisms (SNPs), to detect associations between variants and the disorder. Such variants can be identified through genomewide association studies (GWASs) which employ high-density SNP arrays to scan for SNPs which are statistically associated with the trait of interest. To date, GWASs have identified hundreds of common variants statistically correlated with various traits and diseases (Welter et al., 2014). Because SNPs probed in these studies are designed to capture genome linkage disequilibrium structure, trait-associated SNPs are more likely to tag actual risk loci rather than being causal themselves. More than 90% of these variants are found on noncoding regions and are enriched in DHSs (Maurano et al., 2012) (Fig 1.6), suggesting that causal variants may disrupt regulatory element function. Large-scale epigenomic profiling further revealed a correlation of GWAS variants with the enhancer marks H3K4me1, H3K27ac and eRNA (Akhtar-Zaidi et al., 2012; Ernst et al., 2011; Farh et al., 2015), suggesting that the disruption of enhancer function and dysregulation of the relevant target gene(s) may underlie disease predisposition. In support, fine-mapping of various diseases-associated loci, including coronary artery disease (Harismendy et al., 2011), breast cancer  (French et al., 2013; Meyer et al., 2013) and obesity (Smemo et al., 2014) identified functional variants disrupting enhancer TF binding and expression of target genes. The functional annotation of all genomic TREs, especially enhancers, is expected to shed light on the etiology of common human diseases and ultimately aid in development of novel diagnostic tools and therapeutics.

**Table 1**. **TREs associated with human diseases**

| TRE | Disease | Gene affected | Reference |
|---|---|---|---|
| Core promoter | β-thalassemia | *HBB* | (Antonarakis et al., 1984) |
| Proximal promoter | Familial hypercholesterolemia | *LDLR* | (Mozas et al., 2002) |
| | Hemophilia B | *F9* | (Reitsma et al., 1988) |
| | Pyruvate kinase deficiency | *PKLR* | (Manco et al., 2000) |
| | Charcot-Marie-Tooth disease | *GJB1* | (Wang et al., 2000) |
| Enhancer | Aniridia | PAX6 | (Lauderdale et al., 2000) |
| | Preaxial polydactyl | *SHH* | (Lettice et al., 2003) |
| | Van Buchem disease | *SOST* | (Loots et al., 2005) |
| | X-linked deafness | *POU3F4* | (Naranjo et al., 2010) |
| Insulator | Hereditary spherocytosis | *ANK1* | (Gallagher et al., 2010) |
| Silencer | Asthma | *TGF-β* | (Hobbs et al., 1998) |
| | Facioscapulohumeral muscular dystrophy | 4q35 genes | (Gabellini et al., 2002) |
| LCR | β-thalassemia | *HBB* | (Driscoll et al., 1989) |
| | α-thalassemia | *HBA1* | (Hatton et al., 1990) |

A)

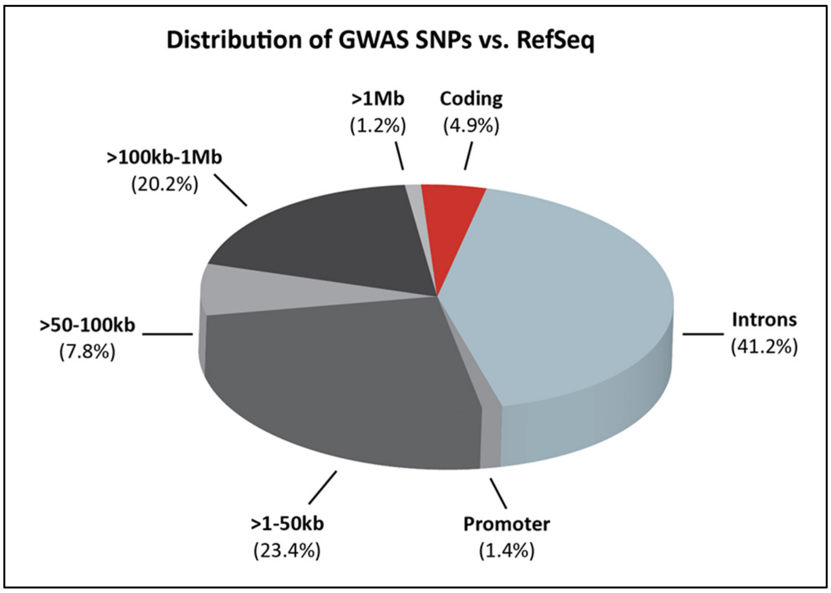


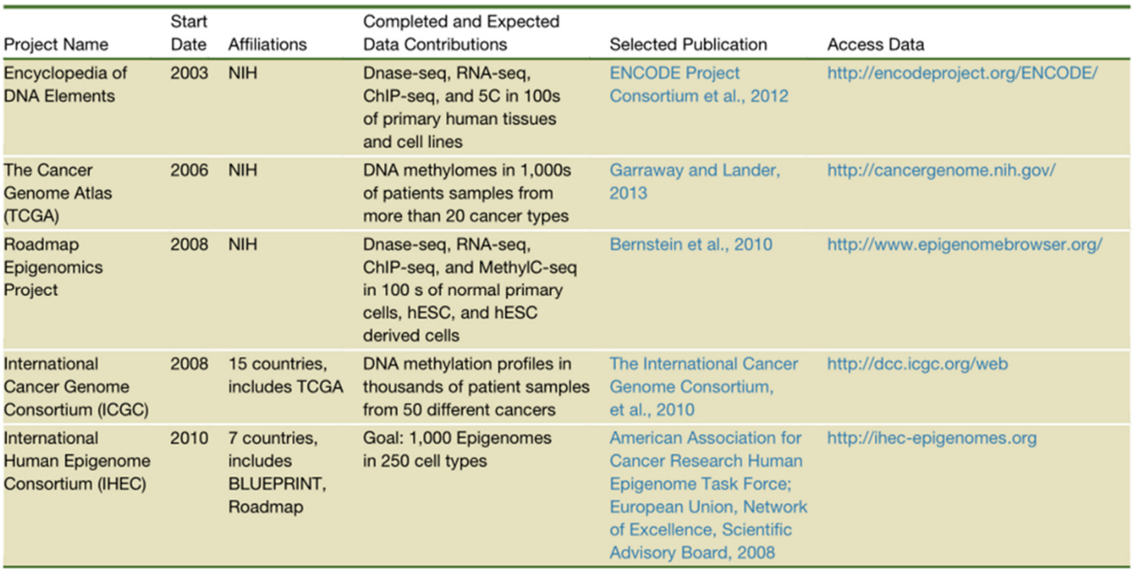
**Figure 1.6 Genomic distribution of GWAS SNPs (Maurano et al., 2012)**

(A) Distribution of GWAS SNPs according to localization on various genomic features, including coding regions, promoters, introns or intergenic regions.

## 1.5 Genome-wide TRE mapping

To functionalize TREs and understand how they are coordinated for gene regulation, as well as the genetic basis of TRE-associated diseases, it is essential to identify and annotate these elements on a genome-wide scale. Unlike TREs at core/proximal promoters, distal TREs are distributed throughout the vast regions of non-coding DNA far from their target genes. In addition, these elements tend to be small and degenerate, and exhibit variable sequence conservation, confounding their identification and annotation. In the past decade, several international consortia, such as ENCODE and the Roadmap Epigenomics Project (Fig 1.7A), have been set up for collaborative epigenome mapping through resource- and data-sharing. The rapid development of NGS technologies and data analysis tools have greatly increased the scope and precision through which nucleotide sequences could be interrogated. As a result, many novel biochemical assays, such as DNase-seq, ChIP-seq and ChIA-PET, were developed and adopted by multiple research labs to leverage on the capabilities of NGS platforms (Fig 1.7B). Using these assays and model cell lines, TREs in 1% of the human genome were mapped in 2007 (Consortium, 2007), and a comprehensive catalogue of TREs in the entire genome was completed in 2012 (Bernstein et al., 2012). These high-throughput genomic mapping tools are discussed in the following sections, with a focus on their specific applications and limitations.

**(A)**



| Project Name | Start Date | Affiliations | Completed and Expected Data Contributions | Selected Publication | Access Data |
|---|---|---|---|---|---|
| Encyclopedia of DNA Elements | 2003 | NIH | Dnase-seq, RNA-seq, ChIP-seq, and 5C in 100s of primary human tissues and cell lines | ENCODE Project Consortium et al., 2012 | http://encodeproject.org/ENCODE/ |
| The Cancer Genome Atlas (TCGA) | 2006 | NIH | DNA methylomes in 1,000s of patients samples from more than 20 cancer types | Garraway and Lander, 2013 | http://cancergenome.nih.gov/ |
| Roadmap Epigenomics Project | 2008 | NIH | Dnase-seq, RNA-seq, ChIP-seq, and MethylC-seq in 100 s of normal primary cells, hESC, and hESC derived cells | Bernstein et al., 2010 | http://www.epigenomebrowser.org/ |
| International Cancer Genome Consortium (ICGC) | 2008 | 15 countries, includes TCGA | DNA methylation profiles in thousands of patient samples from 50 different cancers | The International Cancer Genome Consortium, et al., 2010 | http://dcc.icgc.org/web |
| International Human Epigenome Consortium (IHEC) | 2010 | 7 countries, includes BLUEPRINT, Roadmap | Goal: 1,000 Epigenomes in 250 cell types | American Association for Cancer Research Human Epigenome Task Force; European Union, Network of Excellence, Scientific Advisory Board, 2008 | http://ihec-epigenomes.org |

**(B)**

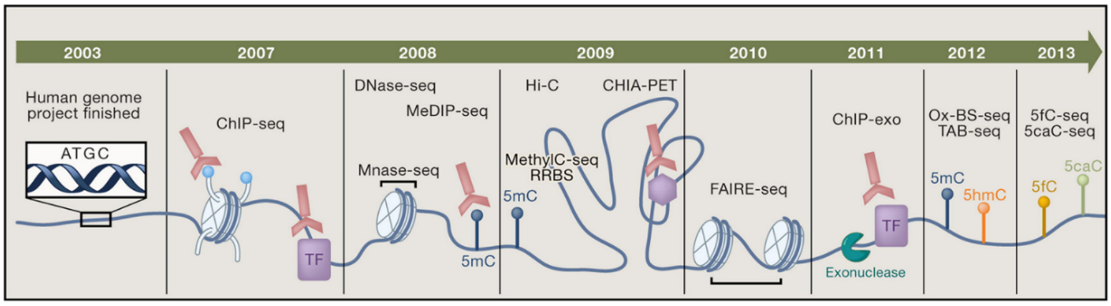


**Figure 1.7 Genome-wide functional element identification.**

(A) Multiple large-scale international consortia were established to focus on epigenomic mapping. (B) High-throughput experimental techniques employed for the identification of various types of genomic elements (Rivera and Ren, 2013).

### 1.5.1 Comparative genomics

Given their central role in mediating gene expression, TREs tend to be evolutionarily conserved due to functional selection. Hence, in principle, TREs can be identified through sequence comparisons between genomes of different organisms and determining regions of high homology (Loots et al., 2000). Such sequence comparisons involve the generation of alignments between orthologous sequence pairs using bioinformatics approaches, taking into account inherent variations such as DNA rearrangements, insertions, deletions, and repeating elements. BLAST (Altschul et al., 1990) represented one of the first alignment tools for pairwise sequence comparison, and has since evolved into a family of related programs for different types of input sequences (Loots, 2008). To overcome the limitations of BLAST in alignment of large genomic regions, newer tools, such as PipMaker and Mulan, were developed which can handle multiple sequence comparisons up to genome-scale size. Visualization of large scale alignment data were facilitated by various web-based database interfaces, such as the UCSC and Ensembl genome browsers, which allow visual navigation along entire genomes. Comparative sequence alignment between human, mouse and fish genomes have been leveraged for genome-wide identification of human TREs which function as tissue-specific enhancers (Pennacchio et al., 2006; Visel et al., 2008) and silencers (Ochi et al., 2012).

One limitation of the comparative genomics approach is the underlying assumption that functional elements under evolutionary constraint exhibit higher similarity, which may not always be true. In a large-scale deletion of over 1,000 constrained elements in mice, no obvious impact on phenotype was observed, suggesting that not all conserved elements are functional (Nóbrega et al., 2004). Conversely, non-conserved elements may be functional. This is illustrated at the *HBB* locus, where well-characterized *HBB* regulatory elements do not exhibit sequence alignment in all the organisms examined (King et al., 2005). On a genome-wide scale, ChIP-seq of mouse heart tissue identified functional enhancers which are poorly conserved (Blow et al., 2010). A second limitation of comparative genomics is the lack of functional details of the identified TREs beyond sequence conservation. Collectively, the limitations of comparative genomics necessitate complementation with other methods for comprehensive TRE annotation.

### 1.5.2 Chromatin accessibility

Eukaryotic genomes are organized as chromatin, a complex of DNA and histone proteins tightly wound into repeating units called "nucleosomes" (Kornberg and Lorch, 1999). Native chromatin, with a condensed nucleosomal structure, occludes DNA-regulatory protein interactions. Regions of chromatin harboring TREs must therefore undergo structural remodeling to increase accessibility of the underlying DNA to regulatory factors. Such physical differences in chromatin accessibility can be used as a proxy for TRE mapping, as nucleosome-free regions exhibit pronounced sensitivity to nuclease digestion compared to native chromatin (Gross and Garrard, 1988). Enzymatic cleavage of chromatin using DNase I or micrococcal nuclease (MNase) can be coupled with high-throughput sequencing (DNase- and MNase-seq) (Boyle et al., 2008; Schones et al., 2008) to achieve high resolution genome-wide mapping of open chromatin structure. An alternative approach, Formaldehyde-Assisted Isolation of Regulatory Elements-sequencing (FAIRE-seq), employs formaldehyde crosslinking and sequencing to identify open chromatin through depletion of histone-bound DNA (Gaulton et al., 2010; Giresi et al., 2007).

By measuring chromatin accessibility in 125 cell and tissue types, a total of 2.9 million DNase-hypersensitive sites (DHSs) have been identified in the human genome, of which ~ 970,000 sites were cell-type-specific (Thurman et al., 2012). The findings that these DHSs were highly correlated with TF binding signals, and that a majority (97.4%) of experimentally-validated *cis*-regulatory elements were found within DHSs, highlighted chromatin accessibility profiling as a powerful tool for comprehensive mapping of regulatory elements. One limitation of DHS-mapping is the large number (50 million) of cells needed (Song and Crawford, 2010), precluding the use of this assay when on clinical samples. To overcome this limitation, an improved assay, ATAC-seq, was developed to leverage on the unique properties of Tn5 transposase for genome fragmentation and sequencing adaptor transposition into accessible chromatin regions for sequencing (Buenrostro et al., 2013). This assay negates multiple intermediate steps, such as gel purification and cross-link reversal, enabling mapping of open chromatin with as little as 500 cells and within clinical timescales. However, like comparative genomics approaches, the lack of functional details for these accessible elements necessitated the integration of additional genomic information, as discussed below, for more comprehensive regulatory element annotation.

### 1.5.3 Regulator binding

The gene regulatory function of TREs are typically mediated by TF binding, hence the identification of TF binding sites represents an alternative method for high-throughput TRE discovery. Among the first genomic technique developed for this purpose is DNA adenine methyltransferase (Dam) identification (DamID) (van Steensel and Henikoff, 2000). In this method, Dam is fused to the DNA-binding protein of interest, resulting in methylation of DNA loci bound by, or in proximity to the protein of interest. Genomic DNA is subsequently digested using methylation-sensitive restriction enzymes followed by ligation of universal adapters. Methylated fragments, representing molecular beacons for protein-bound loci, are then selectively amplified by PCR and detected through microarray hybridization. Using DamID, it is possible to detect indirect and transient binding interactions. However, this assay relies on average DNA methylation patterns over up to 24hrs (Vogel et al., 2007) and is unable to map TF binding changes and TRE activation in dynamic biological processes, such as embryonic development.

An alternative method is ChIP-seq, where bound transcription factors or chromatin proteins are first covalently crosslinked to DNA. Following chromatin fragmentation, DNA fragments are enriched, through their bound proteins, in an immunoprecipitation step. These DNA fragments, representing genomic locations of the bound factor, are purified for adapter ligation and high-throughput sequencing (Fig 1.8). Unlike DamID, ChIP-seq allows genome-wide quantification of all *in vivo* TF binding sites with high resolution and accuracy, making it possible to perform motif predictions to identify non-canonical TF motifs (Johnson et al., 2007). Furthermore, the covalent crosslinking of TFs to chromatin captures a snapshot of binding interactions and allows time-course analysis of binding dynamics (Sandmann et al., 2007; Yáñez-Cuna et al., 2012). A significant improvement of the original ChIP-seq protocol was developed by incorporating an additional nuclease digestion step to trim out unbound and contaminating DNA from protein-binding nucleotides (Rhee and Pugh, 2011). This refinement allowed ChIP-exo to achieve near single-nucleotide resolution of binding peaks with considerably lower false positive rates.

ChIP-seq has been applied to map the binding profiles of more than 100 known TFs in over 70 cell types, revealing hundreds of thousands of putative TREs (Bernstein et al., 2012). Depending on the TF targeted, ChIP-seq can be used to identify specific classes of TREs. For example, global mapping of the insulator protein CTCF revealed cell-type specific barrier insulator elements (Cuddapah et al., 2009), while profiling of the histone acetyltransferase p300 identified active, tissue-specific enhancers (Visel et al., 2009b). Although ChIP-seq can potentially identify all genomic loci bound by a TF, not all these bound sites are functional in regulating gene expression (Fisher et al., 2012; Li et al., 2008). Hence, TREs identified through regulator binding may include a large number of false positive elements, which may be eliminated through assessment of additional chromatin features, such as DNA methylation or histone modifications.

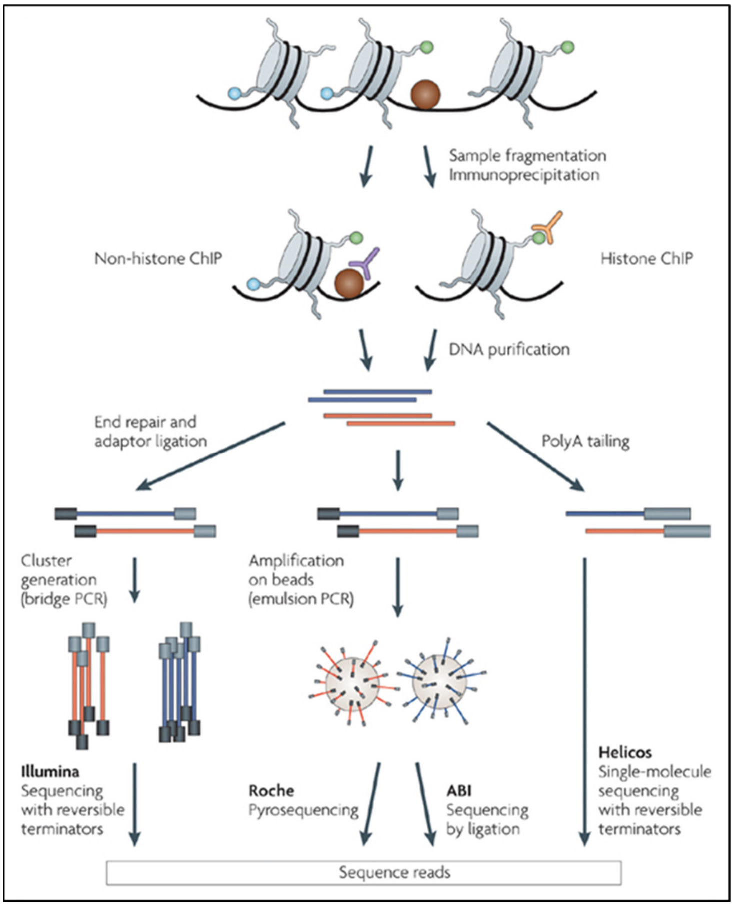


**Figure 1.8 Schematic of the ChIP-seq workflow (Park, 2009).**

### 1.5.4 Chromatin structure

Beyond the primary nucleotide sequence, gene expression regulation by TREs also involves chemical and structural changes to DNA, commonly referred to as epigenetic modifications (Felsenfeld and Groudine, 2003). These modifications can be classified into two main categories: direct methylation of DNA cytosine residues, and post-translational modification of nucleosomal histones. Functionally, these non-genetic changes regulate chromatin packaging as well as genome interpretation by the transcriptional machinery, thus defining TRE identity and function. As such, global profiling of epigenetic modifications represents a powerful method for TRE discovery and annotation. Recently, long noncoding RNAs (lncRNAs) have been demonstrated as key regulators of gene expression and are considered a new class of epigenetic modulators. The roles of lncRNAs in epigenetic regulation are reviewed elsewhere (Mercer and Mattick, 2013).

### 1.5.4.1 DNA methylation

DNA methylation typically occurs on the cytosine residues of CpG dinucleotides in vertebrates (Bird, 2002), and up to 70 – 80% of all CpG dinucleotides are methylated in humans (Ehrlich et al., 1982). Unmethylated CpG dinucleotides often form clusters, known as CpG islands, at the 5' ends of genes (Bird, 1987). The presence of a CpG methyl group can either promote or inhibit binding of transcriptional regulatory proteins, thereby influencing gene expression patterns. For example, methyl-CpG binding domain proteins specifically bind to methylated CpG dinucleotides, facilitating the recruitment of histone deacetylases for chromatin compaction and transcriptional repression (Bird, 2002; Hashimshony et al., 2003). Conversely, methylation of CTCF binding sites at the H19 locus abolished insulator recruitment, resulting in expression of the imprinted Igf2 gene (Bell and Felsenfeld, 2000). Due to its widespread effects on gene expression, DNA methylation have been implicated in diverse processes including X chromosome inactivation, carcinogenesis, development and aging (Das and Singal, 2004; Lister et al., 2013; Mohandas et al., 1981; Oliveira et al., 2012).

Globally, DNA methylation can be profiled by coupling several biochemical assays with high-throughput sequencing. The first method, bisulphite sequencing, involves bisulphite treatment of DNA to convert unmethylated cytosines to uracil (Clark et al., 1994), which is recognized as thymine after PCR amplification and sequencing. A second method involves restriction digestion of DNA, which typically occurs on unmethylated DNA (Bird and Southern, 1978). Alternatively, methylated DNA can be selectively isolated by affinity purification (Cross et al., 1994). The analysis of such genome-wide methylation profiles, or methylomes, revealed that DNA methylation levels on distal TREs outside CpG islands varies as a direct, functional outcome of

TF binding (Feldmann et al., 2013; Stadler et al., 2011). More specifically, these studies revealed that TF binding outside CpG islands is required to generate low-methylated regions (LMRs), which correspond to distal regulatory elements exhibiting cell-type specificity, DNase hypersensitivity and enhancer activities. Importantly, aberrant methylation profiles at these elements have been associated with gene dysregulation in cancer (Aran et al., 2013), suggesting that gene expression control by DNA methylation at distal TREs underlie changes in gene expression contributing to disease.

### 1.5.4.2 Histone modifications

In eukaryotic cells, DNA is folded into nucleosomes, or histone-DNA complexes which constitute a fundamental, repeating unit of chromatin. Each nucleosome consists of 147 bp of DNA wrapped around a histone octamer, consisting of two copies of each of the 4 core histones (H2A, H2B, H3, H4) (Kornberg, 1974). Each core histone consists of an N terminal tail which can be chemically modified, and more than 100 different posttranslational modifications have been identified on the amino-terminal tails, including acetylation, methylation, ubiquitylation and sumoylation (Kouzarides, 2007). Histone modifications play activating and repressive roles in transcription, and generally regulate gene expression through their effects on chromatin accessibility and protein recruitment, yet the detailed mechanisms and functions of a large number of these modifications are not well understood. Genome-wide surveys of histone modification localization has been recognized as an effective method to study their roles in transcription regulation, as well as facilitate detailed mechanistic studies on effects of their deposition and removal on gene expression.

Extensive global profiling of histone modifications have been performed using ChIP-seq (Barski et al., 2007; Mikkelsen et al., 2007; Wang et al., 2008). These landmark studies provided a first glimpse into the complex patterns of modifications, known as the "histone code", at key genomic features, such as promoters, transcribed gene bodies, enhancers and silenced chromatin (Fig 1.9). For example, H3K27ac is found on both active promoters and distal enhancers, while H3K27me3 generally marks repressed or heterochromatic regions (Fig 1.9). Subsequent efforts in epigenome mapping during development revealed a class of 'poised' enhancers which were enriched in H3K4me1 and H3K27me3, and could be differentiated from active enhancers based on H3K27ac enrichment (Creyghton et al., 2010; Rada-Iglesias et al., 2011). The expanding repertoire of publicly available chromatin maps has facilitated computational efforts in integrative analyses. To systematically and comprehensively dissect the functions of the various histone modification combinations, chromatin states from nine human cell lines were segmented into regions of varying combinations using a multivariate hidden Markov model (Ernst et al., 2011). Such integrative analyses of multiple chromatin marks, together with similar work performed on ENCODE chromatin data (Hoffman et al., 2013; Won et al., 2013), provide greater precision and reliability in TRE prediction and condense the large number of chromatin combinations into manageable sets of functional annotations. A key future direction will be to expand histone profiling to include more physiologically-relevant cell types, such as from normal and diseased tissues, to uncover the unique chromatin signatures underlying the different cell states.

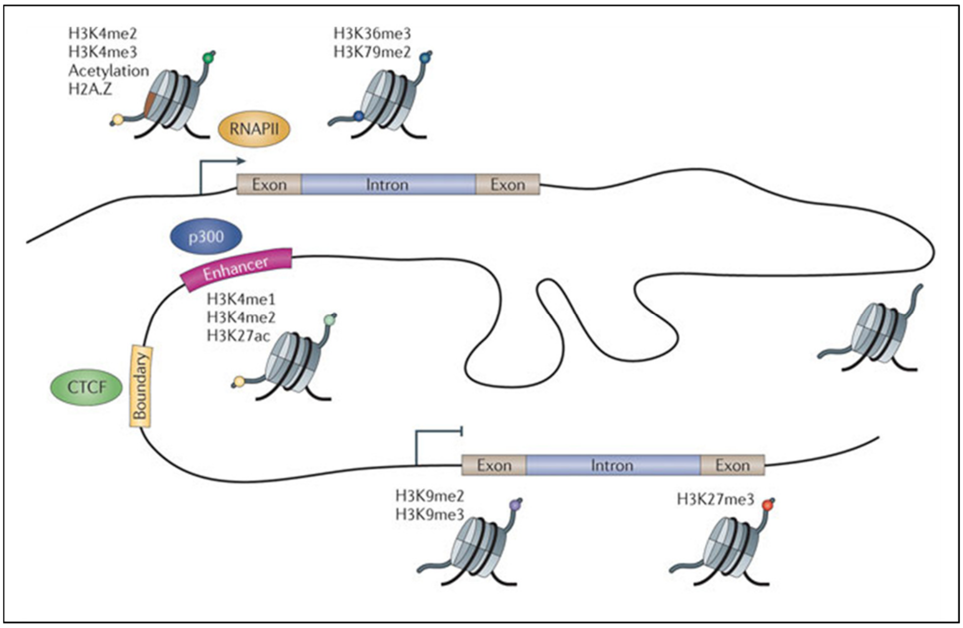


**Figure 1.9 Histone modifications mark functional genomic elements (Zhou et al., 2011)**

Promoters, gene bodies, enhancers, insulators, repressed genes. Histone modifications structurally and functionally define these elements and make it possible to identify them.

## 1.6 Investigating higher-order chromatin organization

The concerted efforts in genome-wide profiling of open chromatin, TF binding sites and chromatin structure over the past few years have greatly facilitated the identification of hundreds of thousands of functional regulatory elements and genetic enhancers (Rivera and Ren, 2013; Xie et al., 2013; Zhu et al., 2013). However, these catalogues of epigenomic data are presented as two-dimensional, linear maps with no information linking regulatory elements to their target gene or loci. The discovery that eukaryotic enhancers can regulate their target genes over megabases of intervening DNA sequences (Lettice et al., 2003; Qin et al., 2004), together with the demonstration that specific DNA regulatory elements are capable of organizing chromatin into domains with distinct expression patterns (Kurukuti et al., 2006; Murrell et al., 2004), suggest a role of topological chromatin organization in transcription regulation. Not only so, defects in higher-order genome organization have also been implicated in gene misregulation leading to human diseases (Misteli, 2010). Hence, large-scale mappings of chromatin conformation and long-range chromatin interactions not only provide a 3D perspective of the genome, but can also inform about the role of chromatin in transcription regulation in health and disease, as well as aid in the functionalization of common disease risk loci.

The past two decades has seen rapid development of various molecular and genomic methods for the investigation of chromatin organization. Among the first genomic technique developed is DNA adenine methyltransferase (Dam) identification (DamID) (van Steensel and Henikoff, 2000). In principle, by fusing Dam to a chromatin-binding protein of interest, target DNA loci bound by, or in proximity to the chromatin protein, will be specifically methylated. Genomic DNA is digested by methylation sensitive restriction enzymes, such as DpnI, followed by ligation of universal adapters. Methylated fragments are then selectively amplified by PCR and detected through microarray hybridization. The DamID technique has been extensively applied in *Drosophila* to study chromatin interactions associated with heterochromatin components, various transcription factors, Polycomb complex proteins and nuclear lamina proteins at kilobase-pair resolution (Moorman et al., 2006; Pickersgill et al., 2006; van Steensel and Henikoff, 2000; Tolhuis et al., 2006). To study the remodeling of genome organization by nuclear lamins during differentiation, Peric Hupkes and colleagues applied DamID to map lamin-associated chromatin contacts in embryonic and terminally differentiated mouse cells, uncovering functional reorganization of hundreds of genes in single transcription units and large, multigene domains (Peric-Hupkes et al., 2010). As DamID relies on average DNA methylation patterns over up to 24 hrs as a readout of proximity (Vogel et al., 2007), it is unable to capture the highly fluid and dynamic changes of lamin scaffold organization previously reported in embryonic stem cells (Bhattacharya et al., 2009). The use of exogenous fusion proteins also precludes DamID experiments on endogenous proteins with native post-translational modifications, such as epigenetic histone signatures, which are intimately tied to chromatin spatial organization. DamID is therefore suitable for genome-wide analysis of chromatin contacts and enhancer-promoter interactions associated with a subset of proteins at low dynamic range.

### 1.6.1 Chromosome Conformation Capture (3C)

A major breakthrough in chromatin organization research resulted from the development of the Chromosome Conformation Capture (3C) assay and its subsequent variants (de Wit and de Laat, 2012) (Fig 1.10). All 3C variants rely on the proximity ligation concept employed in the "nuclear ligation assay" developed in the Seyfred laboratory (Cullen et al., 1993). In 3C, cells are first fixed with formaldehyde to covalently link DNA with their associated proteins. Crosslinked chromatin is then cut using restriction endonucleases, generating chromatin complexes with protruding, sticky-ended DNA strands which are re-ligated under dilute conditions favoring intra-molecular ligations. Interaction frequencies between specific loci, represented by ligation junctions, are then semiquantitatively analyzed through agarose gel electrophoresis (Dekker et al., 2002) or quantitatively through qPCR (Splinter et al., 2006; Vernimmen et al., 2007; Würtele and Chartrand, 2006). In both methods, restriction fragments are amplified using primers specific for predicted ligation junctions. Using this method, Dekker and colleagues constructed a population-averaged contact matrix of yeast chromsome 3, from which a 3D model was developed (Dekker et al., 2002). 3C was subsequently applied to study the β-globin gene locus, uncovering multiple contacts between LCRs with the active β-globin locus in mouse (Tolhuis et al., 2002), consistent with an independent RNA-TRAP analysis of the same locus (Carter et al., 2002). These observations prompted Tolhuis and colleagues to propose that clustering of distal enhancers into an "Active Chromatin Hub" constitutes a key requirement for globin gene expression. Further work employing the 3C assay have demonstrated looping chromatin interactions involving insulator elements (Liu and Garrard, 2005), imprinting control regions (ICR) (Murrell et al., 2004) and trans-interactions from different chromosomes (Dhar et al., 2009; Spilianakis et al., 2005). 3C has also been successfully applied at several noncoding disease risk loci for target gene identification (Bauer et al., 2013; French et al., 2013; Zhang et al., 2012b), providing an important first step towards functionalization of these poorly annotated genomic regions.

### 1.6.2 Circular chromosome conformation capture (4C)

To overcome the limitations of 3C in detecting widely-spaced and novel interactions, several related strategies, collective termed ‘4C’, have been developed to enable whole-genome identification of interaction sites associated with a genomic locus of choice (Lomvardas et al., 2006; Simonis et al., 2006; Zhao et al., 2006). Regardless of the strategy, all 4C methods rely on the formation of a circular DNA template and inverse PCR using primers specific for a genomic locus of interest. Crosslinked chromatin is first cut with four-base cutting restriction enzyme and ligated as in 3C protocols. Subsequent crosslink reversal results in circular DNA molecules which are amenable to inverse PCR amplification. Amplicons representing interacting sites are identified through microarray hybridization or sequencing, revealing global chromatin interactions associated with the H19 ICR and the olfactory receptor enhancer. Frequent four-base cutters are preferred over six-base cutters as four-cutters generate smaller restriction fragments which provide higher resolution and are more reliably amplified by inverse PCR (Ohlsson and Göndör, 2007). Alternatively, a six-cutter can be used in the initial digestion followed by a four-cutter digestion after proximity ligation and crosslink reversal to extract ligation junctions for circularization and inverse PCR (Simonis et al., 2006). Higher circularization efficiency is expected using the latter strategy, as reactions are performed on smaller, naked DNA instead of crosslinked DNA. The second digestion step also ensures that ligation junctions are not too large to be efficiently amplified by inverse PCR, which is possible when multiple DNA fragments are crosslinked and ligated. A modification of this two-step digestion strategy was employed by Ling et al. in their “Associated Chromosome Trap” (ACT) assay (Ling et al., 2006). Instead of circularization and inverse PCR amplification of interacting sequences, PCR adapters are ligated onto DNA fragments after the second digestion and identified through PCR and sequencing. The

ACT assay demonstrated that imprinting control regions (ICRs) can physically colocalise with their target genes through interchromosomal interactions. Like 3C, 4C can be coupled with ChIP and a ChIP-4C technique, termed enhanced-4C (e4C), has been developed to study all transcriptional interactions associated with a locus of interest (Schoenfelder et al., 2009). Through RNAPII immunoprecipitation, proximity ligation and 4C analysis, a cluster of hundreds of transcribed loci were found to be associated with the mouse globin locus, with specific subsets of coassociated gene loci bound and coregulated by a lineage-specific TF, Klf1. To overcome some of the inherent limitations associated restriction digestion of chromatin, a sonication-based 4C strategy was developed (Fullwood et al., 2010). Regardless of the exact strategy used, these studies collectively exemplified 4C as a versatile "one-versus-all" strategy for targeted dissection of global chromatin interaction. Indeed, 4C has been successfully applied to chart the *cis*-regulatory circuitry at the *FTO* locus, unraveling the genetic targets of noncoding variation associated with increased risk of obesity in humans (Smemo et al., 2014).

### 1.6.3 Chromosome conformation capture carbon copy (5C)

5C is an adaptation of 3C for chromatin interaction analysis between multiple genomic elements in parallel. Following proximity ligation and crosslink reversal, 3C templates are subjected to ligation-mediated amplification (LMA) (Landegren et al., 1988) at high levels of primer multiplexing. Each primer pair, specific for one end of a restriction fragment, will anneal adjacent to one another at the targeted ligation junction and are subsequently ligated to form a 5C library. As each primer carry a universal 5' end sequence, the 5C library can be amplified with a pair of universal primers. Ligation junctions can then be quantified using microarray or high-throughput sequencing. Comprehensive 5C analyses of the β-globin locus not only recapitulated known LCR interactions, but also revealed novel interactions between the LCR and the δ- and γ-globin genes (Dostie et al., 2006). The 2D interactions maps generated by 5C can be represented in 3D using an iterative modeling approach, first demonstrated at the HoxA gene cluster to reveal chromatin dynamics during cellular differentiation (Fraser et al., 2009). Subsequent work combining 5C with an integrative structure determination platform (Alber et al., 2007) culminated in higher-resolution 3D models of the α-globin domain, demonstrating the clustering of active genes into chromatin globules (Baù et al., 2011). This strategy was subsequently applied to uncovering the basic folding principles of a bacteria genome (Umbarger et al., 2011). Several inherent limitations preclude the use of 5C technology for modeling larger genomes. When measuring interaction frequencies over many Mbs in complex genomes, the large numbers of ligation junctions, coupled with high levels of multiplexing primers, confounds accurate ligation junction quantification. Although the use of universal primers in 5C largely avoids amplification efficiency issues present in 3C, annealing efficiencies of 5C primers can vary and require ligation templates to control for such differences. The preparation of control templates, consisting of numerous minimally-overlapping BAC clones, can be laborious for large genomes. 5C is thus suited for comprehensive analyses of long range interactions within specific gene clusters or small genomes within a couple of Mbs.

### 1.6.4 Hi-C

Hi-C was developed in 2009 to leverage on the scale of massively parallel DNA sequencing technologies to address global chromatin interactions at a genome-wide scale. In Hi-C, crosslinked chromatin is fragmented through restriction digestion, leaving overhangs which are filled in with biotin-labeled nucleotides. Blunt-ended DNA fragments are ligated under dilute proximity ligation conditions and purified after crosslink reversal. DNA is further sheared and ligation junctions are pulled down using streptavidin beads. High throughput sequencing of ligation junctions following by reference genome mapping will identify interaction between all pairs of restriction fragments. Using this catalogue of interacting fragments, Lieberman *et al* constructed the first genome-wide matrix of chromatin contacts in human cells at Mb-scale, confirming known principles of genomic organization and revealing distinct spatial compartmentalization of open, gene-rich chromatin from inactive and closed domains (Lieberman-Aiden et al., 2009). These domains were proposed to be territorially organized in a highly compact, knot-free configuration within the nucleus. A subsequent Hi-C study using human and mouse cells revealed high pervasiveness and conservation of chromatin domains within and between species (Dixon et al., 2012), suggesting evolutionary conservation of topological domains as a structural feature of mammalian genomes. In *Drosophila*, a higher resolution (kilobase-pair-scale) Hi-C map has likewise uncovered distinct physical domains demarcating genomic activity (Sexton et al., 2012). Notably, analysis of *Drosophila* chromatin interactions at higher resolution (10-100 kb) suggests a hierarchical organization involving discrete physical chromatin modules, distinct from the fractal globule conformation proposed using the human Hi-C map.

One variant of Hi-C has been independently developed to address chromosomal configurations in yeast (Duan et al., 2010). In this assay, ligation junctions are circularized following two restriction digestion steps as in 4C. A further restriction digestion step allows ligation of adapters containing an EcoP15I restriction site. Subsequent EcoP15I digestion releases paired-end tags for high throughput sequencing. This strategy was successfully applied to map global chromatin interactions at kb-resolution in yeast, leading to a proposed 3D model consistent with the Rabl-orientation reported in yeast nuclei (Jin et al., 1998). A second strategy, tethered chromosome capture (TCC) (Kalhor et al., 2012), was developed to minimize random inter-complex ligations by immobilizing chromatin complexes on beads prior to proximity ligation. The improved sensitivity allowed the authors to define specific chromosomal regions and quantify their propensity to from interchromosomal interactions.

The Hi-C assay can be combined with other genomic techniques to study the link between chromatin organization and specific biological processes. As discussed by Wijchers and colleague (Wijchers and de Laat, 2011) and substantiated by *in silico* analysis of human Hi-C data (Fudenberg et al., 2011), 3D genome organization may be the primary dictator of the targets of cancer-causing chromosomal recombination, suggesting that chromatin conformation directly reflects gene expression patterns underlying normal and diseased biological processes. Furthermore, by combining Hi-C with genome-wide translocation sequencing, it was demonstrated that a direct relationship exist between intra- and interchromosomal translocation loci and their proximity in spatial space (Chiarle et al., 2011; Zhang et al., 2012c). A further study subsequently revealed a direct relationship between Hi-C chromatin proximity maps and genome-wide replication timing profiles (Ryba et al., 2010), suggesting domain compartmentalization as a possible mechanism for regulating DNA replication. Collectively, these studies highlight the broad applicability of Hi-C with other genomic techniques to address functional relationships between chromatin architecture and specific biological processes

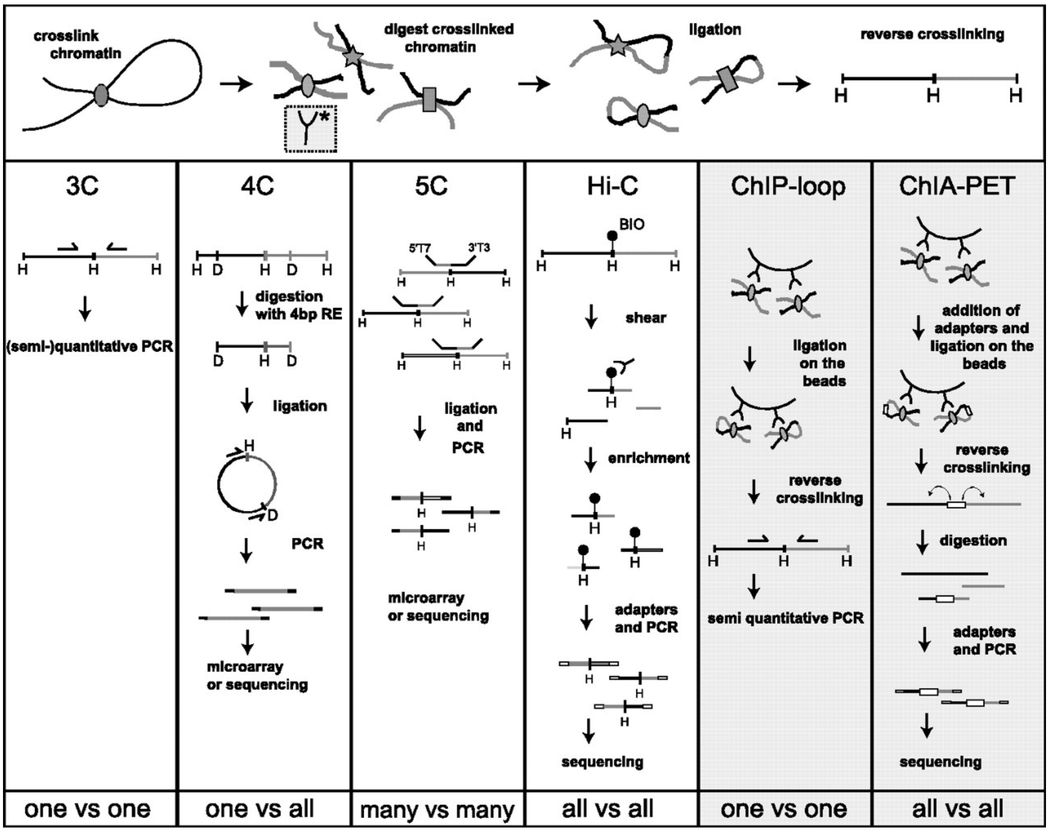


**Figure 1.10 Summary of 3C-based methods (de Wit and de Laat, 2012)**

### 1.6.5 Chromatin Interaction Analysis by Paired-End Tag Sequencing (ChIA-PET)

The determination of 3D chromatin organizations represents an exciting first step towards complete functional annotation of the genome. To understand how DNA sequence and associated proteins specify genome organization, ChIA-PET was developed (Fullwood et al., 2009). In this assay (Fig 1.11), crosslinked chromatin is sheared by sonication and enriched for specific protein factors through immunoprecipitation. Biotinylated half-linkers containing MmeI restriction sites are added to blunt-ended DNA fragments to facilitate proximity ligation. Following crosslink reversal, DNA fragments are restriction digested to release ligation junctions, which are isolated using streptavidin-coupled magnetic beads. Through PCR amplification, high-throughput sequencing and reference genome mapping of ligation junctions, ChIA-PET allows the detection of DNA loci tethered together by the protein factor of interest and connected through proximity ligation. In a proof-of-concept study, ChIA-PET has been applied to analyze chromatin interactions associated with oestrogen receptor α in human breast adenocarcinoma cells (Fullwood et al., 2009). The identification of all ERα-associated chromatin interactions allowed the mapping of an extensive network of intrachromosomal interactions involving gene promoters, suggesting a global mechanism through which ERα organizes higher-order chromatin structure to direct target gene expression.

The role of chromosomal organization in global transcriptional regulation was elucidated in a subsequent ChIA-PET study which mapped long-range chromatin interactions associated with RNA polymerase II (RNAPII) in five human cell lines (Li et al., 2012). Widespread promoter-centered chromatin contacts were uncovered which further aggregated into multigene complexes, suggesting a chromatin structural framework for coordinated transcription, in support of the postulated “transcription factory” model (Cook, 1999). Furthermore, in an independent study of H3K4me2-associated chromatin interactions (Chepelev et al., 2012), a large proportion (56%) of interactions were formed between promoters exhibiting coexpression, suggesting that subnuclear chromatin organization provides a structural framework for transcription regulation.

Insights into chromatin organization beyond the 30 nm fiber have been provided by the identification of various scaffolding proteins, such as SATB, CTCF, condensin  and the mediator complex (Cai et al., 2003; Gartenberg and Merkenschlager, 2008; Kagey et al., 2010; Phillips and Corces, 2009). Among these factors, CTCF has been recognized as the “master” regulator of chromatin architecture. ChIA-PET has been applied to map the CTCF chromatin interactome in mouse embryonic stem cells (Handoko et al., 2011), revealing novel intra- and interchromosomal interactions. Through association of CTCF-associated chromatin loops with their underlying epigenetic signatures, five distinct interaction domains were uncovered, each specifying a unique transcriptional status, suggesting interplay between architectural proteins and transcriptional machinery for functional chromatin compaction. An investigation of enhancer-promoter interactions through cohesin-associated chromosome structures further revealed that transcription of cell identity genes, driven by super-enhancers, are control by local chromosome structure (Dowen et al., 2014).

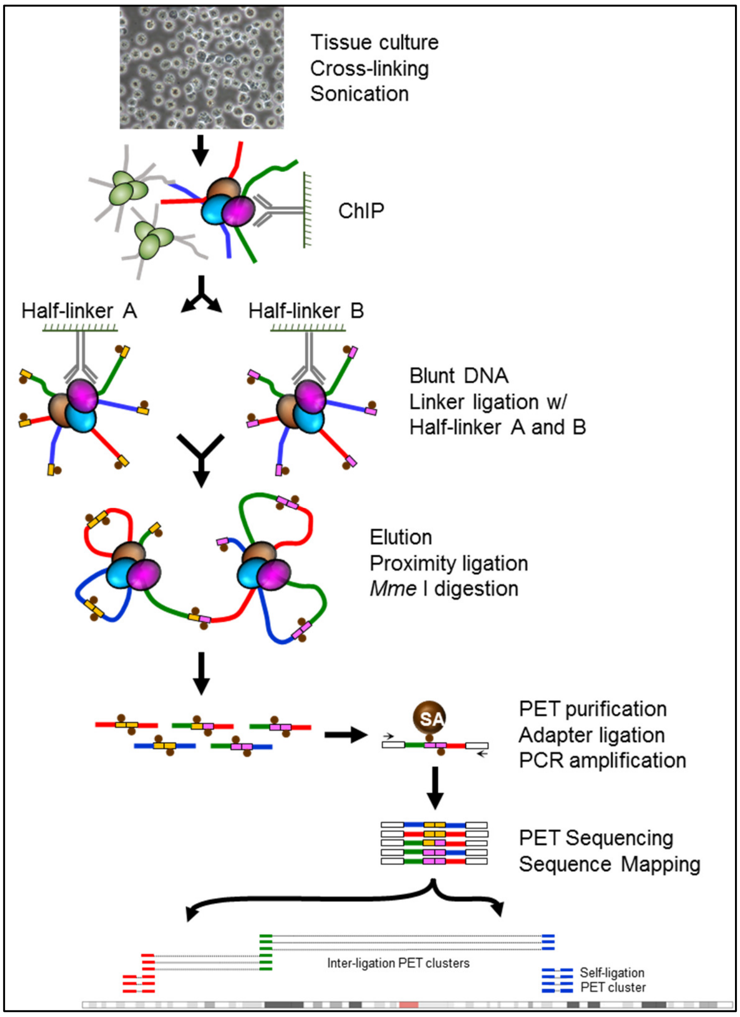


**Figure 1.11 Schematic of the ChIA-PET methodology** (Zhang et al., 2012a)

## 1.7 Embryonic endoderm development

### 1.7.1 Multiple signaling levels regulate endoderm formation

The definitive endoderm (DE) is one of the three primary germ layers which is formed during gastrulation (Lu et al., 2001). DE is a physiologically important germ layer as it gives rise to a diverse array of cell types, including epithelial cells of the digestive and respiratory systems, as well as multiple organs such as the liver, pancreas, thymus, thyroid and lungs. As a result of this versatility in organ specification, disruptions in endodermal organ functions lead to multiple human diseases afflicting millions of individuals each year. A thorough understanding of endoderm organ formation, from descriptive embryology, signaling pathways, transcription factor networks and *cis*-regulatory systems is needed to understand the genetic basis of human congenital diseases, as well as to facilitate efforts in endoderm tissue engineering for transplantation based therapies.

Fate-mapping and embryological experiments in model organisms such as mouse, chick *Xenopus* and zebrafish have revealed the cellular aspects of endoderm emergence (Fig 1.12). Despite differences in morphological arrangements of the prospective germ layers, the underlying molecular pathways culminating in a primitive gut tube formation are highly conserved. In all amniote embryos, gastrulation begins when a cluster of cells at the posterior end of the epiblast ingress to form the primitive streak (PS), which marks the site of further cellular migration culminating in mesoderm and endoderm formation (Mikawa et al., 2004). Specifically, epiblast cells at the anterior end of the primitive streak transit through an “epithelial-mesenchymal” process and form the middle (mesoderm) or outer (definitive endoderm) layer of the gastrula (Shook and Keller, 2003). Definitive endoderm is then patterned along the anteroposterior axis by signals arising from the surrounding mesoderm, which specify the naïve endoderm to distinct foregut, midgut and hindgut domains (Zorn and Wells, 2009). These committed domains are further specified into various organ-specific lineages, ultimately contributing to the formation of organs along the anterior-posterior body axis, such as the esophagus, lungs, stomach, liver, pancreas and intestines (Grapin-Botton and Melton, 2000).

*In vivo* genetic perturbation and explant studies have provided much insights into the extrinsic signals required for endoderm lineage specification (Zorn and Wells, 2009). The Nodal signaling pathway is central to body axes specification as well as mesoderm and endoderm formation in frog, zebrafish, mouse and chick (Schier, 2003). Nodal signaling ligands belong to the TFBβ family of secreted growth factors, and signal through the transmembrane serine-threonine kinase receptors Alk4 or Alk7. Phosphorylation by Alk4/7 results in activation of the cytosolic proteins Smad2/3 which, together with Smad4, lead to stimulation of endodermal genes such as Foxh1 or Mixl1. As a result, loss of function components of the Nodal signaling pathway result in compromised endoderm formation (Dunn et al., 2004; Liu et al., 2004; Shen, 2007). Consistent with these studies, inactivation of the Nodal antagonist Lefty2 results in excessive endoderm formation (Meno et al., 1999). Early studies investigating signals required for Nodal expression revealed that mouse embryos lacking Nodal or β-catenin do not form primitive streak, suggesting an involvement of the canonical Wnt pathway in endoderm induction (Huelsken et al., 2000). Subsequent studies revealed that Wnt signaling is required to maintain Nodal expression and promote endoderm formation (Ben-Haim et al., 2006; Fan et al., 2007).

Nodal signaling induces the endodermal lineage through the expression of a network of transcription factors, including Sox17, Eomes, Gata4/6 and the Mix-like homeobox genes(Zorn and Wells, 2007). Collectively, these genes result in a commitment to an endodermal fate while separating endoderm and mesoderm lineages. Global expression profiling of endoderm precursor cells suggested that these genes are wired within multiple complex feedback loops, influencing each other's expression and harboring distinct and overlapping target genes (Brown et al., 2008; Tamplin et al., 2008).

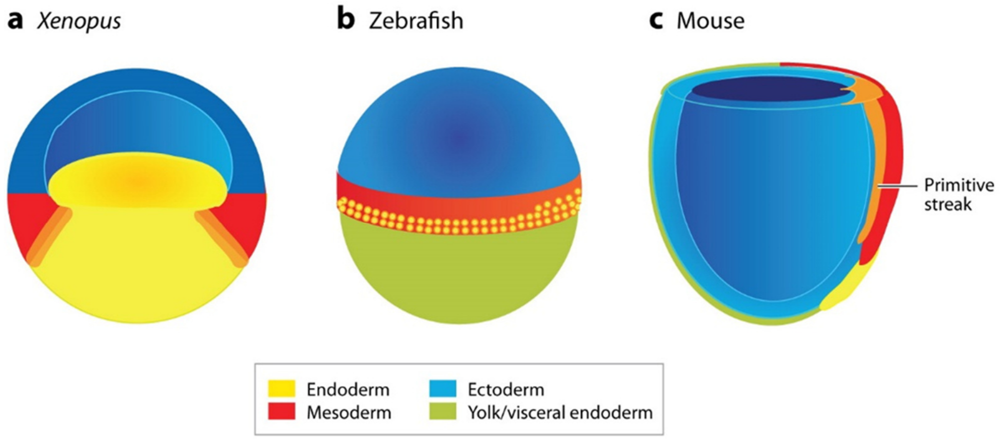


**Figure 1.12 Fate maps of *Xenopus*, zebrafish and mouse (Zorn and Wells, 2009)**

Fate mapping studies in (a) *Xenopus*, (b) zebrafish and (c) mouse have delineated the cell types in the embryo giving rise to endoderm.

### 1.7.2 Endoderm patterning

Following lineage commitment, the flat sheet of endodermal germ cells undergoes a series of morphogenetic movements, originating from the anterior and posterior ends, which converge at the yolk stalk to form a gut tube (Grapin-Botton and Melton, 2000). Studies in *Xenopus* and mice revealed that by the end of gastrulation, maternal Wnt/β-catenin signals at the anterior domain induces expression of the anterior transcription factors Hhex (Zorn et al., 1999), Mixl1 (Hart et al., 2002) and Foxa2 (Dufort et al., 1998). These TFs are required for foregut development and their differential expression at the anterior domain result in the establishment of a broad anterior-posterior domain boundary which specifies foregut development at the anterior end. This initial boundary is further reinforced and maintained by graded and overlapping patterning signals, including FGF, Wnt and BMP ligands, which are secreted from the surrounding mesodermal tissues (Tam et al., 2007). At the transcriptional level, FGF4 signaling induces *Cdx2* expression, required for hindgut specification, while repressing *Hhex* and *Foxa2* (Dessimoz et al., 2006). Similarly, experiments in *Xenopus* and zebrafish established the role of Wnt signaling in promoting hindgut development and blocking foregut fate (Goessling et al., 2008; McLin et al., 2007). Conversely, Wnt signaling must be repressed in the anterior endoderm by the Wnt antagonist Sfrp5 to maintain a foregut identity (Fig 1.13). At this stage, each anterior-posterior domain has already acquired distinct developmental potentials, evidenced by observations that the foregut has a higher potential to form hepatic tissues compared to hindgut cells (Fukuda-Taira, 1981), while the hindgut is incompetent to respond to posterior foregut pancreatic induction by retinoid acid (Kinkel et al., 2008). Cellular signals and TFs required for domain induction are well-established, it is unclear how this process is regulated at the transcription regulation level, and current knowledge of the *cis*-regulatory elements driving this process are lacking. Given that transplantation-based therapies based on *in vitro* derivation of endoderm organs have great therapeutic potential, and that human congenital malformations frequently afflict multiple endodermal organs simultaneously, there are growing interests in unraveling the molecular mechanisms underlying endoderm tissue and organ development.

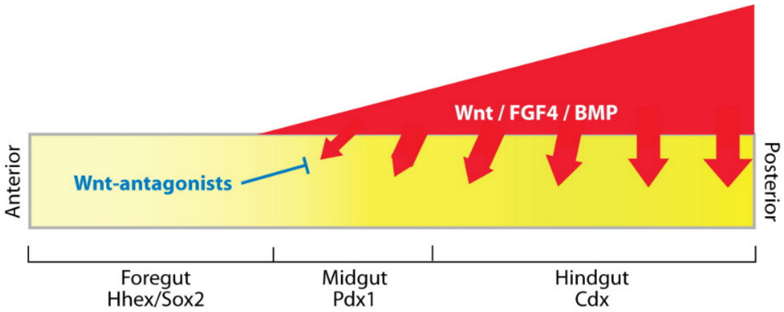


**Figure 1.13 Schematic illustration of endoderm patterning (Zorn and Wells, 2009)**

Towards the end of gastrulation, endoderm domains expressing distinct TFs are patterned by differential signaling pathways, specifying foregut, midgut and hindgut progenitor domains.

### 1.7.3 Endoderm induction from embryonic stem cells

Human embryonic stem cells (hESCs), with their unlimited potential for self-renewal and ability to form all human cell types, has been widely touted as a potential cure for a wide range of diseases, including those affecting endoderm-derived organs such as diabetes, biliary atresia and ulcerative colitis. Efforts focusing on differentiation of hESCs into endoderm have relied on lessons learnt through *in vivo* genetic perturbations and explants approaches using model vertebrate organisms (Bernardo et al., 2011; Tam and Loebel, 2007; Zorn and Wells, 2009). TGFβ signaling through Nodal/Activin could induce an endodermal cell fate from hESCs just as they do in mouse and *Xenopus* embryos (D'Amour et al., 2005; Hudson et al., 1997; Tada et al., 2005). Strikingly, the timing of ESC-directed endoderm differentiation by Activin, as revealed by molecular marker expression, closely mimics that in mouse gastrulation. Furthermore, as in Xenopus embryos, low levels of Activin (1-10ng/ml) induced a mesoderm fate, while high (10-100ng/ml) doses induced DE formation. These observations suggest that a deep understanding of endoderm development in model organisms can help guide efficient differentiation of hESCs through each developmental juncture into endoderm. Based on this principle, several *in vitro* differentiation methods have been developed to varying degrees of success based on induction efficiencies, yielding a mixture of DE and other cell lineages (Cheng et al., 2012; D'Amour et al., 2005; Touboul et al., 2010). These mixed induction outcomes may arise from a lack of understanding of the precise involvement of each signal in DE induction, leading to inefficient induction or suppression of alternative lineages at each developmental branchpoint. For example, although BMP, FGF, Wnt and VEGF have been used, in addition to TFGβ pathway activation, for DE differentiation, these signals are also involved in mesoderm formation (Cheng et al., 2012; Green et al., 2011; Touboul et al., 2010). As a heterogeneous population of DE cells compromise subsequent generation of endoderm organs (McKnight et al., 2010), it is imperative that signals at each step of endoderm formation and patterning be accurately defined. Through systematic provision and blockage of these signals in hESCs, considerable progress have been achieved in endoderm specification and patterning (Loh et al., 2014). These advances have enabled in-depth studies of global chromatin structure, as well as gene expression and TF occupany in these transient developmental cell states (Loh et al., 2014).

## 1.8 Aim and objectives of thesis

As discussed above, enhancers represent the most abundant and cell-specific elements among all known TRE classes. These properties, coupled with recent findings that a majority of genetic variants associated with human complex diseases reside on enhancers (Maurano et al., 2012), fueled much motivation and interest in enhancer discovery and annotation. These pursuits are further facilitated by the recent availability of next-generation sequencing technologies and novel molecular assays, such as RNA-seq, ChIP-seq and DNase-seq. These assays make it possible for global investigations of epigenomic profiles and identification of transcriptional enhancers.

In this thesis, we aim to identify transcriptional enhancers in a genome-wide and cell-type specific manner, and make use of these predicted enhancers to study the role of *cis*-regulation in human endoderm development and disease. To achieve this, the main objective involves genome-wide profiling of the transcriptome and histone modifications in human endoderm at various stages of development. As vertebrate endoderm formation and patterning constitutes one of the earliest events during human gestation, and are largely inaccessible for detailed molecular studies, we will use an *in vitro* hESC differentiation model which recapitulates marker gene expression characteristic of endoderm. With this, we will examine transcriptomic profiles and underlying chromatin states as hESCs transit through defined developmental stages, giving rise to endoderm and downstream regional antecedents of endodermal organs. Although much progress have been made in elucidating the signaling and TF network regulating endoderm formation, how these signals and TFs are coordinated for  endoderm specification are largely unknown. As enhancers function as integrated binding platforms for lineage-specifying TFs and signaling effectors, we will also investigate the coordination between these elements and factors in regulating endoderm development. We also aim to investigate whether the predicted enhancers are enriched in variants contributing to complex diseases afflicting endodermal organs. Lastly, we aim to set up chromatin interaction analysis assays to study the general principles underlying enhancer-promoter interactions, which will aid in elucidating target genes of enhancers harboring disease-causing risk variants.

During the course of this work, two independent groups have similarly performed human endoderm epigenome mappings (Gifford et al., 2013; Xie et al., 2013). Unlike this study, Xie *et al* profiled promoter chromatin remodeling, revealing the role of repressive Polycomb group proteins in endoderm specification. Gifford *et al* claimed to comprehensively map promoter and enhancers from all three hESC-derived germ layers, including endoderm. However, the reported endoderm enhancers were enriched for neural (ectoderm-derived) functions (Loh et al., 2014), suggesting that those elements were not endoderm-specific. This study therefore represented the first successful effort in high-quality endoderm enhancer discovery. Using these predicted enhancers we sought to understand the role of distal regulatory elements in integrating extrinsic signals and master regulatory TFs for endoderm lineage specification. We also aimed to test whether these elements are enriched for causal variants of complex diseases afflicting endoderm-derived organs. A crucial problem in enhancer research involves target gene identification, as enhancer function is typically quite insensitive to distance to their target loci. The final objective of this work is to facilitate enhancer target gene identification through the mapping of all promoter-associated chromatin interactions in human cells.

# CHAPTER 2 MATERIALS AND METHODS

## 2.1 Cell culture

hESCs were propagated on matrigel- (BD Biosciences) coated plates in mTeSR1 media (Invitrogen). mTeSR1 media was refreshed every 24 hrs, and cultures were visually inspected daily for spontaneous differentiation. Differentiated cell clumps were removed thorough scraping. During serial passaging, culture media was aspirated and cells were incubated with collagenase IV at 37°C for 5 min. Subsequently, cells were scraped into clumps and transferred at a split ratio of 1:3 onto new coated plates. PS, DE, AFG, PFG and MHG cells were differentiated from hESCs (Loh et al., 2014) and were contributed by Kyle Loh and Dr. Lay Teng Ang (Genome Institute of Singapore). Details of the differentiation protocol is described in Chapter 2.2.

MCF7 and K562 cells were obtained from the American Type Culture Collection (ATCC). MCF7 cells were grown in DMEM/F12 media (Invitrogen) supplemented with 5% FBS (Invitrogen), 10 U/ml pencillin (Invitrogen), 100 µg/ml streptomycin (Invitrogen) and 4 mM L-glutamine (Invitrogen). Cells were maintained at 37°C, 5% $CO_2$ and passaged at 80% confluency. During passaging, cells were detached from culture vessel using 0.25% trypsin/0.03% EDTA solution at 37°C for 5 min, rinsed with fresh growth media and seeded into new culture vessels at a split ratio of 1:5. K562 cells were grown in RPMI 1640 media (Invitrogen) supplemented with 5% FBS (Invitrogen), 10 U/ml pencillin (Invitrogen), 100 µg/ml streptomycin (Invitrogen) and 4mM L-glutamine (Invitrogen). Cells were maintained at 37°C, 5% $CO_2$ and passaged when confluent ($8\times10^5$ cells/ml). During passaging, cells were centrifuged at 100 *g* for 5min and resuspended to $2\times10^5$ cells/ml in culture media.

## 2.2 DE specification and anterioposterior patterning from hESCs

Confluent hESC cultures were passaged as small clumps with collagenase and seeded onto Matrigel coated plates. After 1-2 days of recovery in mTeSR1, hESCs were washed with F12 (Gibco) and were treated for 24 hrs with Activin A (100 ng/mL, R&D Systems), CHIR99021 (2 μM, Stemgent), and PI-103 (50 nM, Tocris) in CDM2 to specify APS. Afterwards, cells were washed (F12), then treated for 48 hours with Activin A (100 ng/mL), LDN-193189/DM3189 (250 nM, Stemgent), and FGF2 (10 ng/mL, Invitrogen) in CDM2 to generate DE by day 3. Media was refreshed every 24 hrs.

Day 3 DE was patterned into AFG, PFG, or MHG by 4 days of continued differentiation in CDM2. DE was washed (F12), then differentiated as follows: AFG, A-83-01 (1 μM, Tocris) and DM3189 (250 nM); PFG, RA (2 μM, Sigma) and DM3189 (250 nM); MHG, BMP4 (10 ng/mL, R&D Systems), CHIR99021 (3 μM), and FGF2 (100 ng/mL), yielding day 7 anterioposterior domains. To derive specific PFG organ domains (in CDM2 + KnockOut Serum Replacement (KOSR, 10% v/v, Gibco)), DE was washed, treated with DM3189 (250 nM), IWP2 (4 μM, Stemgent), PD0325901 (500 nM, Tocris), and RA (2 μM) for 1 day, washed (F12), and then differentiated 3 more days for pancreatic or hepatic induction. Pancreatic differentiation: Activin A (10 ng/mL), DM3189 (250 nM), IWP2 (4 μM), PD0325901 (500 nM), RA (2 μM), and SANT1 (150 nM, Tocris). Hepatic differentiation: A-83-01 (1 μM), BMP4 (10 ng/mL), IWP2 (4 μM), and RA (2 μM). Day 7 pancreatic progenitors were further differentiated towards endocrine progenitors by treatment with RA (2 μM), DM3189 (250 nM), and DAPT (10 μM, Tocris) for 2 days in CDM2 + KOSR (10%). Media was refreshed every 24 hours for all conditions.

CDM2 media consist of the following: 50% IMDM (Gibco) and 50% F12 (Gibco), supplemented with 5 mg/mL polyvinyl alcohol (Sigma, A1470 or Europa Bioproducts, EQBAC62), 1% v/v chemically-defined lipid concentrate (Gibco, 11905-031), 450 μM monothioglycerol (Sigma, M6145), 0.7 μg/mL insulin (Roche, 1376497), and 15 μg/mL transferrin (Roche, 652202).

### 2.3 RNA-seq and data analysis

Prior to RNA extraction, cell culture media was aspirated and cells washed with PBS. Total RNA was extracted using an RNeasy Mini Kit (Qiagen) coupled with on-column DNase (Qiagen) treatment according to the manufacturer's protocols. RNA concentration and quality was assessed using a Bioanalyzer (Agilent). 1μg total RNA with RIN > 9.0 was used for RNA-seq library construction with a TruSeq RNA Library Preparation Kit (Illumina) according to the manufacturer's instructions. Briefly, mRNA purified from total RNA was fragmented to a size range of 300 to 500 bp and reverse-transcribed. The resulting cDNA fragments were blunt-ended and 3'-adenylated to allow sequencing adapter ligation. Following adapter ligation, libraries were PCR amplified for 15 cycles. Library size and concentration were validated by on-chip electrophoresis (Bioanalyzer, Agilent) and qPCR (Lightcycler, Roche). High-throughput sequencing was performed on a Hi-Seq2000 machine (Illumina) for 1×36+7 cycles (single read, 36 bp for insert, 7 bp for adapter barcode). Preliminary data analyses were performed using the Illumina sequencing analysis software CASAVA 1.8, including image analysis, base calling, cluster filtering, library demultiplexing and sequence alignment. Aligned reads were used as input into DESeq (Anders and Huber, 2010) for normalization and differential expression analysis. Gene expression levels were normalized as RPKM (reads per kilobase of exon per million mapped reads) to account for gene length differences, and genes with RPKM < 1 in all cell types were regarded as not expressed and excluded from subsequent analyses. For each gene, RPKM values were normalized as percentage of the highest value across all cell types to allow expression profile comparison among genes with widely differing expression levels. Normalized expression values were clustered using the HOPACH package in Bioconductor (van der Laan and Pollard, 2003) to identify stage-specific transcripts. Expression heatmaps were generated using Java Treeview (Saldanha, 2004) and Genepattern software (Broad Institute).

## 2.4 Reverse-transcription quantitative-PCR

Cells were harvested and total RNA extracted using an RNeasy Mini Kit (Qiagen) as described. 1 µg total RNA was reverse-transcribed into cDNA using Superscript III (Invitrogen). cDNA levels for specific genes were quantified by quantitative PCR using gene-specific primers (Appendix I) and detected on a LightCycler 480 instrument (Roche). Gene expression levels were normalized to the housekeeping gene *PBGD* for comparison between different cDNA samples.

## 2.5 Chromatin Immunoprecipitation (ChIP) and sequencing

Adherent cells were washed with PBS and cross-linked in 1% formaldehyde for 10 mins at RT. Cross-linking is quenched with 0.2 M glycine for 5mins, and cells were collected by scraping and centrifugation. Cells were further washed twice with ice-cold PBS supplemented with Complete Protease Inhibitor (Roche), pelleted and stored at -80°C. Prior to immunoprecipitation, cell pellets were thawed on ice, lysed in 1% SDS lysis buffer (50 mM HEPES-KOH pH 7.5, 150 mM NaCl, 1 mM EDTA, 1% Triton X-100, 0.1% Na-deoxycholate, 1% SDS with 1X Complete Protease Inhibitor) twice and sonicated for 10 cycles at high intensity (30 s on, 60 s off) using a Bioruptor sonicator (Diagenode). To assess sonication efficiency, a small aliquot of sonicated chromatin was digested with Proteinase K (1 hr, 50°C), column-purified and electrophoresed to ensure a fragment size of 100 – 300 bp. Prior to immunoprecipitation, sonicated chromatin was diluted ten times in ChIP dilution buffer (0.01% SDS, 1.1% Triton X-100, 1.2 mM EDTA, 16.7 mM Tris-HCl pH 8.1, and 167 mM NaCl) to 0.1% SDS concentration. Diluted chromatin was centrifuged (13,200 rpm, 10 min) to remove cellular debris and pre-cleared overnight with Protein G Dynabeads (Invitrogen). Concurrently, for each individual ChIP, 100 µL of Protein G Dynabeads was washed twice (PBS + 0.1% Triton X-100), complexed with ChIP-qualified antibody overnight at 4°C, and washed thrice to yield antibody-bead complexes. Pre-cleared chromatin was combined with antibody-bead complexes and incubated overnight (4°C) for immunoprecipitation. Antigen-antibody-bead complexes were washed twice respectively in low salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris pH 8.0, 150 mM NaCl), high salt wash buffer (0.1% SDS, 1% Triton X-100, 2 mM EDTA, 20 mM Tris pH 8.0, 500 mM NaCl), LiCl wash buffer (10 mM Tris pH 8.0, 1 mM EDTA, 0.25 M LiCl, 1% Nonidet P-40) and TE buffer. Antibodies were eluted from beads and formaldehyde cross-linking was reversed overnight by mild heating (65°C), and chromatin was sequentially treated with RNase and Proteinase K before final column purification. The final concentration of immunoprecipitated chromatin was quantified by a PicoGreen assay (Invitrogen). To ensure specificity of the ChIP protocol, enrichment of ChIP DNA was quantified by ChIP-qPCR with gene-specific primers and detected on a LightCycler 480 instrument (Roche). Antibodies used for ChIP are anti-H3K4me2 (ab32356), anti-H3K27ac (ab4729), anti-H3K4me3 (ab8580) from Abcam, and anti-H3K27me3 (07-449) from Millipore/Upstate.

10 ng of ChIP DNA was used for sequencing library construction using a ChIP-seq DNA sample Prep Kit (Illumina) with modifications. Briefly, DNA was end-repaired, 3'-adenylated, ligated with Illumina adapters and PCR amplified (15 cycles) with Phusion High Fidelity DNA polymerase (Finnzymes). Amplified fragments were purified and size-selected using AMpure beads (Beckman Coulter). Library size was verified by on-chip electrophoresis (Agilent Bioanalyzer) and library concentration quantified by qPCR. High-throughput sequencing was performed on a Hi-Seq 2000 system (Illumina) for 1 x 36+7 cycles (single read, 36 bp of insert of a multiplexed library, 7 bp for adapter barcode identification). Each library was sequenced to a depth of at least 37 million reads.

## 2.6 ChIP-seq data analysis

Sequenced reads were mapped to the hg19 human reference genome using Bowtie (Langmead et al., 2009), allowing up to 3 bp mismatches and discarding reads mapping to more than 1 genomic locus. Each aligned fragment was extended by 200 bp and input-normalization and peak calling was performed using HOMER (Heinz et al., 2010). Histone modification peaks were called using a 2 kb peak width, without subtraction of local surrounding background. Enrichment fold changes were computed against input controls and peaks were identified based on > 4-fold enrichment. Enhancer peaks were considered gene-distal based a minimum 2 kb distance from annotated TSS. Enhancer centers were defined using the "Nucleosome Free Regions" (nfr) feature for peak calling in HOMER, which centralizes enhancer peaks on local depletions (dips) in histone modification, signifying nucleosome-free regions which facilitates TF binding. Cell type-specific peaks were identified using *k*-means clustering with Cluster 3.0 software (de Hoon et al., 2004) based on Euclidean distance. Histone modifications on promoters were evaluated by calculating ChIP-seq tag counts within a 2 kb window, normalized to corresponding tag counts in the input dataset. Motif enrichment was performed by comparing cell-type-specific peak sequences with 50,000 genomic fragments with same length and GC content of the peaks. Enrichment of gene ontology terms was calculated using GREAT (McLean et al., 2010) using the "basal plus extension" association rule, with maximal extension length of 200kb. ChIP-seq tracks were visualized using the Integrative Genomics Viewer from the Broad Institute (Thorvaldsdottir et al., 2012).

### 2.7 Gene Set Enrichment Analysis (GSEA)

The GSEA software (Subramanian et al., 2005) was downloaded from http://www.broadinstitute.org/gsea. GSEA accepts as input 1. Gene expression profile obtained from microarray or RNA-seq and 2. A gene set obtained based on prior biological knowledge. The GSEA algorithm will determine whether genes in the provided gene set are randomly distributed throughput the supplied gene expression profile, or preferentially enriched in the top or bottom in terms of expression level. Here, we used as gene expression profile microarray data of hESCs induced towards endodermal differentiation by Activin treatment (Bernardo et al., 2011). The gene set consisted of the genes proximal to the top 500 putative DE enhancers by H3k27ac enrichment levels. These two inputs were used for GSEA analysis using default settings. Briefly, the gene set is ranked using the Signal2Noise metric, which ranks each gene based on the difference of the means scaled by standard deviation:

**Signal2noise(gene) =** $\frac{\mu_A - \mu_B}{\sigma_A + \sigma_B}$

Where $\mu_A$ represents the mean of the expression level of gene g of all replicates under Activin treatment, and $\mu_B$ represents the mean of expression level of gene g of all replicates in untreated cells. Likewise, $\sigma_A$ represents standard deviation of gene g among all replicates under Activin treatment, while $\sigma_B$ represents standard deviation in untreated cells.

With these two inputs, GSEA ranks the gene list by expression level, and computes an enrichment score (ES) which reflects the extent to which putative DE enhancer-proximal genes is overrepresented at the extremes (highest and lowest) of the gene list. ES is calculated through a running-sum statistic through the gene list, increasing when a DE enhancer-proximal gene is encountered in the gene list, and decreasing when not (Subramanian et al., 2005). A *p* value is then calculated using a phenotype-based permutation test method. Specifically, genes are randomly assigned to either "Activin-treated" or "untreated", re-ranked and an ES calculated. This process is repeated for 1,000 permutations, generating a histogram of permutated, null ESs. P value of the observed ES is then estimated relative to this null ES distribution.

## 2.8 Putative functional variant identification

To look for associations between MHG enhancers and known ulcerative colitis risk polymorphisms, 108 ulcerative colitis risk SNP IDs, together with their respective genomic coordinates and population of origin, were obtained from the NHGRI GWAS catalogue (http://www.genome.gov/gwastudies) (Annex III) and used as input in the FunciSNP software (Coetzee et al., 2012). All known variants within a window size of 500 kb from each ulcerative colitis risk SNP were extracted from the relevant population from the 1000 Genomes database and merged with MHG-specific enhancers to identify overlaps. $R^2$ and D' values were calculated for each SNP residing within an MHG enhancer.

## 2.9 ChIA-PET library construction

RNAPII and CTCF ChIP was performed as described above in K562 and hESC respectively. Antibodies used were RNAPII 8WG16 monoclonal (Covance mms-126R) and anti-CTCF polyclonal (Millipore 07-729). For each library, 500 ng of ChIP-enriched chromatin fragments, tethered to antibody-conjugated Protein G Dynal beads, were end-repaired using T4 DNA polymerase (Promega) and washed thrice with ChIA-PET wash buffer (10 mM Tris-HCl, 1 mM EDTA, 500 mM NaCl). Each sample was divided into two equal aliquots and ligated separately with Linkers A and B respectively using T4 DNA ligase (Fermentas). These two linkers share the same nucleotide sequences except for 4 nucleotides in the middle (Linker A - TAAG; Linker B - ATGT) which serve as a nucleotide barcode for linker identification. Following overnight ligation at 16°C, the two aliquots were washed, 5'-phosphorylated using DNA polynucleotide kinase (NEB) and eluted in elution buffer (10 mM Tris-HCl, 1 mM EDTA, 1% SDS). To minimise inter-ligation between different chromatin complexes, proximity ligation is carried out in dilute conditions (1.2 ml chromatin in 10 ml ligation reaction). ChIA-PET constructs were extracted from ligation products by *Mme*I digestion, amplified using PCR (18 cycles), and purified by phenol-chloroform extraction and isopropanol precipitation. To check for successful library amplification, purified PCR products were electrophoresed in a 6% TBE gel (Invitrogen) and stained with SYBR Green I for 10min. The 223 bp band, representing amplified ChIA-PET ligation fragments, was gel purified, and sequenced using a Hi-Seq 2000 instrument (Illumina). ChIA-PET data analysis was performed by Dr Li Guoliang using the ChIA-PET tool (Li et al., 2010).

## 2.10 Fluorescence In Situ Hybridization (FISH)

MCF7 cells were harvested by trypsinization, transferred to 15ml falcon tubes and swelled in hypotonic conditions (0.75 M KCl) for 15 min at 37°C. Cells were fixed with methanol/acetic acid (3/1), dropped onto glass slides and aged overnight at 37°C. Following overnight propagation of *E.Coli* harbouring BAC DNA of interest, plasmid DNA was extracted (Nucleobond PC500 Plasmid Purification Kit, Macherey-Nagel) and labelled by nick translation (Nick Translation Labelling Kit, ENZO Life Sciences) in the presence of biotin-16-dUTP or digoxigenin-11-dUTP (Roche). In the presence of 1mg/ml of Human Cot1 and Salmon sperm DNA (Invitrogen), labelled BAC clones were resuspended to 5 ng/ml in hybridization buffer (PBS, 2×SSC, 10% dextran sulphate, 50% formamide). Prior to hybridization, MCF7 nuclei on slides were digested with (0.005% pepsin (Sigma), 0.01 M HCl) at 37°C for 3min followed by fixation with 1% formaldehyde (Merck - Calbiochem) and dehydrated through a 70%-80%-100% ethanol series. Labelled probes were denatured at 75°C for 5 min and hybridized to aged slides (described above) at 37°C overnight. Post-hybridization washes were performed twice at 45°C in 2×SSC, 50% formamide for 7 min each followed by 2 washes in 2×SSC at 45°C for 7 min each. Slides were revealed with avidin-conjugated fluorescein isothiocyanate (FITC) (Vector Laboratories, CA) for biotinylated probes and anti-digoxigenin-Rhodamine for digoxigeninlabeled probes (Roche). After washing, slides were mounted with vectashield (Vector Laboratories) and observed under an epifluorescence microscope (ZEISS, Image.Z2) and under 63× lens magnification. Between 300 – 800 interphase nuclei were analyzed for each probe mix. Probe signals were visualized and analysed using metafer4 software. A center-to-center distance of 1μm was used as cut-off for colocalization analysis. BACs used are listed in APPENDIX II.

## 2.11 RNAPII immunofluorescence staining combined with DNA FISH

MCF7 cells were grown to 70% confluency in hybridization chambers and fixed with 4% paraformaldehyde (PFA) (Sigma-Aldrich) for 15min at room temperature (RT) followed by permeabilization with 0.04% Triton X (Promega) for 30min at RT. Prior to staining, cells were blocked with 10% normal donkey serum (Millipore) for 1 hr at RT and incubated with mouse RNA polymerase II 8WG16 monoclonal antibody (Covance, 1:1000) overnight at 4°C. Cells were incubated with Cy3-conjugated donkey anti-mouse IgG polyclonal antibody (Millipore, 1:1000) for 1hr at RT, after which slides were mounted with ProLong Gold antifade reagent containing DAPI (Invitrogen). For combined RNAPII staining-DNA FISH, RNAPII-stained cells were post-fixed in 4% PFA for 10 min at RT after secondary antibody incubation and permeabilized in 0.5% Triton X for 10 min at RT. Cells are subsequently dehydrated through a 70%-80%–100% ethanol series, rehydrated with 2×SSC and denatured in 2×SSC/50% formamide at 80°C for 40 min. Prior to FISH probe hybridization, cells were incubated in 2×SSC for 5min at 4°C and permeabilized in 2×SSC/0.5% Triton X for 5 min at 4°C. Probe preparation and hybridization for FISH were performed as described for DNA-FISH. Association of RNAPII foci with promoter-promoter interaction loci was visualized with a Carl Zeiss Meta LSM confocal microscope with 63x optical lens and interslice distance (Z-axis) of 0.4mm. Data was analysed using LSM image browser and percentage overlap was determined by manual counting.

## 2.12 Zebrafish enhancer reporter assay

Putative enhancer elements were PCR-amplified from 100ng human genomic DNA (Promega) using Phusion High-Fidelity DNA polymerase master mix (Thermo Scientific) and enhancer-specific PCR amplification primers (APPENDIX I). Amplification products were electrophoresed in a 1% agarose gel for size verification, and PCR conditions were optimized to ensure only a single PCR band per amplification reaction was observed. Successfully amplified enhancers were cloned into pENTR-TOPO vectors (Invitrogen) according to manufacturer's instructions, and purified using a plasmid elution kit (Qiagen Miniprep kit). Positive transformants were screened using colony PCR with Taq DNA polymerase (NEB). pENTR-TOPO vectors contain recombination sites derived from bacteriophage λ to facilitate recombination with the ZED vector through the Gateway® Technology (Landy, 1989). ZED vector was a kind gift from Dr. José Bessa (Bessa et al., 2009). LR recombination was performed using Clonase enzyme mix (Clontech) and positive recombinants were identified through BglII restriction digestion and further verified by single-pass sequencing. Tol2 transposase mRNA was prepared by *in vitro* transcription using the mMESSAGE mMACHINE kit (Life Technologies) using Tol2 cDNA. Tol2 cDNA was a kind gift from Dr. Cathleen Teh (IMCB, A*STAR Singapore).

For embryo microinjection, injection needles were prepared by pulling borosilicate glass capillaries (O.D.: 1.0 mm, I.D.: 0.58 mm) using a micropipette puller (P-97, Sutter Instruments) with the following settings: Heat=560, Pull=30, Vel=90, Time=80. The injection mix was prepared by mixing 2.5 ul enhancer DNA (80 ng/ul), 5 ul Tol2 transposase mRNA (50 ng/ul) and 2.5 ul phenol red solution (0.1% in PBS, Sigma). 1nl of the injection mix was injected into each zebrafish embryo during the 1-cell stage. For each enhancer element, a total of 100 – 120 embryos were injected. Embryos were observed by fluorescence microscopy between 16-36 hours-post-fertilization (hpf) for reporter activity.

# CHAPTER 3 TRANSCRIPTIONAL AND EPIGENOMIC PROFILING DURING HUMAN ENDODERM DEVELOPMENT

## 3.1 Transcriptomic profiling of endoderm differentiation and patterning

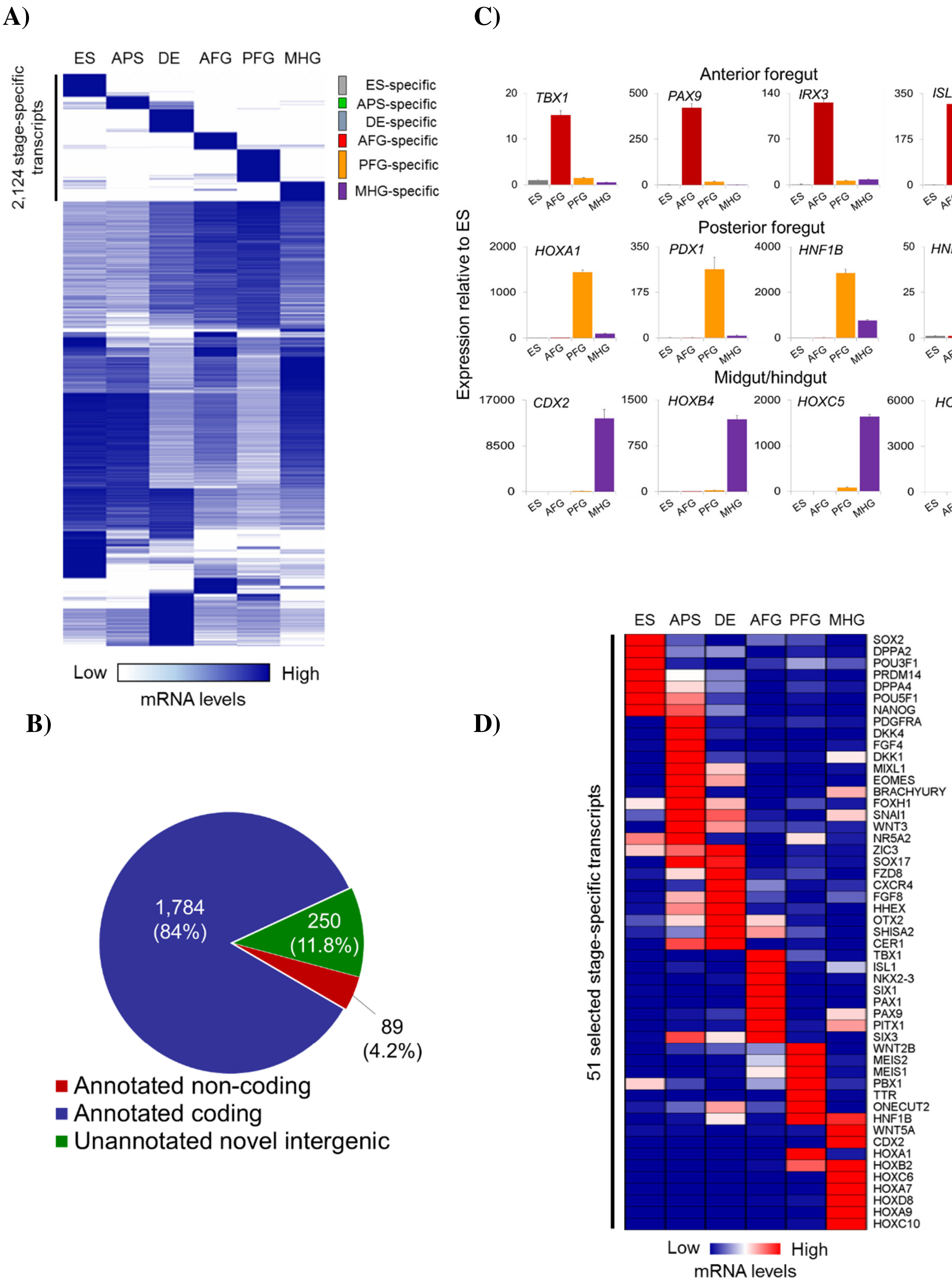
To recapitulate human endoderm formation and subsequent derivatives along the anteroposterior axis, an *in vitro* differentiation model was employed in which hESCs are progressively induced to differentiate into primitive streak (PS) and definitive endoderm (DE), then subsequently patterned into anterior foregut (AFG), posterior foregut (PFG) and mid/hindgut (MHG) progenitor cells. To capture the dynamics of global gene expression patterns during this process, levels of polyadenylated transcripts from each of these cell types were quantified by RNA-seq. Following total RNA extraction, mRNA purification, library construction and high-throughput sequencing, 74 – 95 million raw reads per library was obtained (Table 2). These reads were filtered and mapped using the standard Illumina analysis pipeline (described in Materials and Methods), generating an average of 68 million uniquely mapped reads per library for downstream analyses. A transcript was considered as expressed if it presented an [RPKM] > 1 in all six cell types. Using these criteria, a total of 16,781 transcripts were detected over the course of differentiation and patterning. To study cell-type specific expression, all transcripts were clustered based on expression levels using the HOPACH algorithm (van der Laan and Pollard, 2003) (Fig 3.1A). Among the 2,124 cell-type specific transcripts identified, 1,784 (84%) mapped to coding regions, 89 (4.2%) consisted of non-coding RNAs, while a further 250 (11.8%) transcripts mapped to unannotated genomic regions (Fig 3.1B). The expression levels of 12 cell-specific transcripts were further verified by qPCR, which confirmed their stage-specific expression patterns (Fig 3.1C).

To investigate whether these clusters are truly representative of cell-type and region-specific identity, expression levels of 51 key regulatory genes in endoderm development and patterning were presented as a heatmap (Fig 3.1D). Characteristic of an undifferentiated state, genes specifically expressed in hESCs include the pluripotency markers *SOX2*, *PRDM14* and *DPPA4*, while *NANOG* and *OCT4* expression persists up to PS stage, consistent with previous observation that expression of these two genes persists until streak formation (Teo et al., 2011). Endoderm markers, including *SOX17*, *CXCR4* and *CER1*, were specifically enriched in the DE cluster. Likewise, transcripts for the anterior foregut markers *TBX1*, *PAX1* and *PAX9* (Green et al., 2011) were enriched in AFG. Consistently, *PDX1*, *MEIS1* and *MEI2* (Fujitani et al., 2006; von Burstin et al., 2010) exhibited PFG-specific expression, and multiple posterior *HOX* transcripts (Illig et al., 2013) were enriched in MHG. Gene Ontology analysis was performed on genes enriched in each stage following differentiation (Fig 3.1E) (McLean et al., 2010). The PS and DE-specific cluster was enriched for the Gene Ontology term "endoderm development" ($p = 6.3\times10^{-5}$) and "primary germ layer formation" ($p = 4.4\times10^{-4}$); AFG cluster for "pattern specification process" ($p = 1.5\times10^{-7}$); PFG cluster for "liver development" ($p = 5.2\times10^{-7}$); MHG cluster for "tube development" ($p = 8.2\times10^{-6}$) and "urogenital system development" ($p = 2.6\times10^{-4}$). These distinct gene signature classes suggest that key endoderm developmental programs are regulated in a stage-specific manner, and further demonstrated that the *in vitro* hESC-differentiation model employed can accurately recapitulate early stages of human embryological development. Our global transcriptomic data represent the first comprehensive gene catalogues of human endoderm differentiation, and further revealed a panel of non-coding RNAs and unannotated transcripts whose roles in endoderm development are poorly characterized and understood. These data hence represent a novel and valuable resource for future investigation of lineage determinants driving human endoderm development.

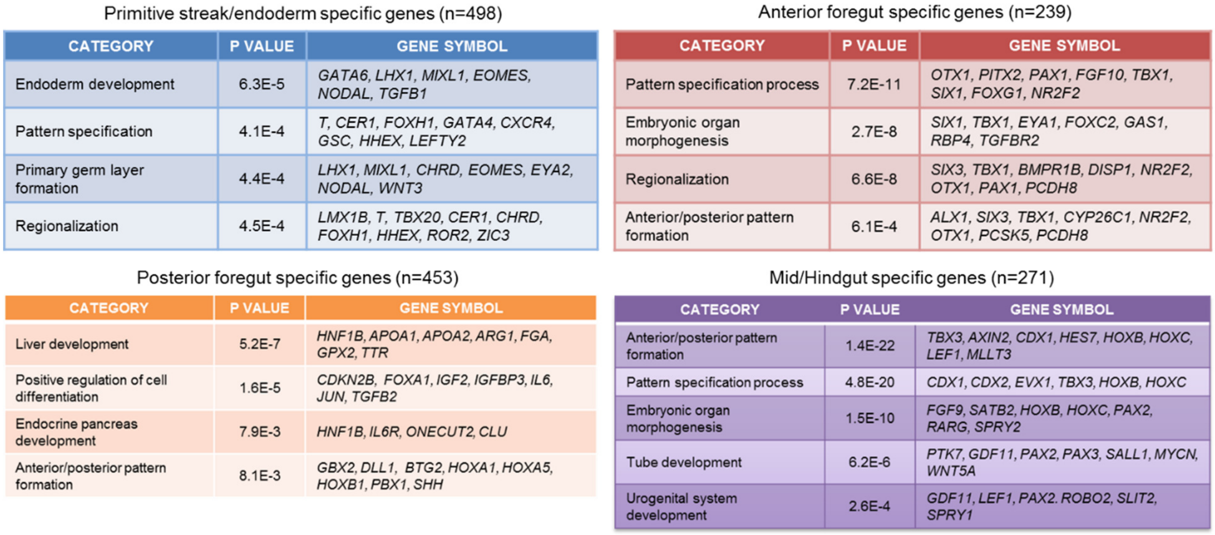
**Table 2 Statistics of sequenced and aligned RNA-seq reads**

| Sample | Sample Yield (Mb) | Clusters (raw) | % PF Clusters | Uniquely Aligned Clusters | % Align (PF) |
|---|---|---|---|---|---|
| ES | 2,473 | 74,549,297 | 94.77 | 54,097,899 | 76.57 |
| PS | 3,152 | 95,345,326 | 94.45 | 70,418,857 | 78.2 |
| DE | 2,990 | 91,244,264 | 93.63 | 68,458,887 | 80.13 |
| AFG | 3,090 | 94,498,635 | 93.44 | 71,416,623 | 80.88 |
| PFG | 3,117 | 95,569,242 | 93.18 | 73,263,188 | 82.27 |
| MHG | 3,099 | 95,183,586 | 93.01 | 68,860,031 | 77.78 |
| Average | 2,986.83 | 91,065,058 | 93.75 | 67,752,581 | 79.31 |

PF: pass-filter


A)
ES APS DE AFG PFG MHG
2,124 stage-specific transcripts
ES-specific
APS-specific
DE-specific
AFG-specific
PFG-specific
MHG-specific
Low High
mRNA levels
B)
1,784 (84%)
250 (11.8%)
89 (4.2%)
Annotated non-coding
Annotated coding
Unannotated novel intergenic
C)
Expression relative to ES
Anterior foregut
TBX1 20 10 0
PAX9 500 250 0
IRX3 140 70 0
ISL1 350 175 0
Posterior foregut
HOXA1 2000 1000 0
PDX1 350 175 0
HNF1B 4000 2000 0
HNF4A 50 25 0
Midgut/hindgut
CDX2 17000 8500 0
HOXB4 1500 750 0
HOXC5 2000 1000 0
HOXC6 6000 3000 0
ES AFG PFG MHG
D)
ES APS DE AFG PFG MHG
51 selected stage-specific transcripts
SOX2
DPPA2
POU3F1
PRDM14
DPPA4
POU5F1
NANOG
PDGFRA
DKK4
FGF4
DKK1
MIXL1
EOMES
BRACHYURY
FOXH1
SNAI1
WNT3
NR5A2
ZIC3
SOX17
FZD8
CXCR4
FGF8
HHEX
OTX2
SHISA2
CER1
TBX1
ISL1
NKX2-3
SIX1
PAX1
PAX9
PITX1
SIX3
WNT2B
MEIS2
MEIS1
PBX1
TTR
ONECUT2
HNF1B
WNT5A
CDX2
HOXA1
HOXB2
HOXC6
HOXA7
HOXD8
HOXA9
HOXC10
Low High
mRNA levels

**E)**



Primitive streak/endoderm specific genes (n=498)

| CATEGORY | P VALUE | GENE SYMBOL |
|---|---|---|
| Endoderm development | 6.3E-5 | *GATA6, LHX1, MIXL1, EOMES, NODAL, TGFB1* |
| Pattern specification | 4.1E-4 | *T, CER1, FOXH1, GATA4, CXCR4, GSC, HHEX, LEFTY2* |
| Primary germ layer formation | 4.4E-4 | *LHX1, MIXL1, CHRD, EOMES, EYA2, NODAL, WNT3* |
| Regionalization | 4.5E-4 | *LMX1B, T, TBX20, CER1, CHRD, FOXH1, HHEX, ROR2, ZIC3* |

Anterior foregut specific genes (n=239)

| CATEGORY | P VALUE | GENE SYMBOL |
|---|---|---|
| Pattern specification process | 7.2E-11 | *OTX1, PITX2, PAX1, FGF10, TBX1, SIX1, FOXG1, NR2F2* |
| Embryonic organ morphogenesis | 2.7E-8 | *SIX1, TBX1, EYA1, FOXC2, GAS1, RBP4, TGFBR2* |
| Regionalization | 6.6E-8 | *SIX3, TBX1, BMPR1B, DISP1, NR2F2, OTX1, PAX1, PCDH8* |
| Anterior/posterior pattern formation | 6.1E-4 | *ALX1, SIX3, TBX1, CYP26C1, NR2F2, OTX1, PCSK5, PCDH8* |

Posterior foregut specific genes (n=453)

| CATEGORY | P VALUE | GENE SYMBOL |
|---|---|---|
| Liver development | 5.2E-7 | *HNF1B, APOA1, APOA2, ARG1, FGA, GPX2, TTR* |
| Positive regulation of cell differentiation | 1.6E-5 | *CDKN2B, FOXA1, IGF2, IGFBP3, IL6, JUN, TGFB2* |
| Endocrine pancreas development | 7.9E-3 | *HNF1B, IL6R, ONECUT2, CLU* |
| Anterior/posterior pattern formation | 8.1E-3 | *GBX2, DLL1, BTG2, HOXA1, HOXA5, HOXB1, PBX1, SHH* |

Mid/Hindgut specific genes (n=271)

| CATEGORY | P VALUE | GENE SYMBOL |
|---|---|---|
| Anterior/posterior pattern formation | 1.4E-22 | *TBX3, AXIN2, CDX1, HES7, HOXB, HOXC, LEF1, MLLT3* |
| Pattern specification process | 4.8E-20 | *CDX1, CDX2, EVX1, TBX3, HOXB, HOXC* |
| Embryonic organ morphogenesis | 1.5E-10 | *FGF9, SATB2, HOXB, HOXC, PAX2, RARG, SPRY2* |
| Tube development | 6.2E-6 | *PTK7, GDF11, PAX2, PAX3, SALL1, MYCN, WNT5A* |
| Urogenital system development | 2.6E-4 | *GDF11, LEF1, PAX2. ROBO2, SLIT2, SPRY1* |

**Figure 3.1 RNA-seq analysis reveals cell-type specific transcripts during endoderm differentiation and patterning**

(A) HOPACH clustering of RNA-seq reads to identify cell-specific transcripts according to expression levels. (B) Proportion of transcripts sorted according to annotation and coding characteristics. (C) qPCR validation of expression levels of 12 cell-type specific transcripts. (D) Heatmap showing expression patterns of 51 key endoderm regulatory genes. (E) Gene Ontology analysis of the identified transcript classes.

## 3.2 Chromatin state profiling during endoderm formation

The precise and complex spatial-temporal gene expression patterns during early embryonic development are controlled by TREs found throughout the genome (Maston et al., 2006). We hypothesized that comprehensive profiling of chromatin states during endoderm lineage commitment will enable identification and annotation of such regulatory elements. ChIP-seq, with its ability for global unbiased mapping of protein binding or localization, was selected as the method of choice to profile the distribution of a panel of histone modifications (H3K4me2, H3K4me3, H3K27ac, and H3K27me3) during defined stages of endoderm differentiation. Prior to ChIP-seq library construction, efficiency of the ChIP process was first evaluated to ensure sufficient and specific enrichment for the appropriate histone modification of interest. One key factor affecting this process is antibody quality, which may exhibit substantial lot-to-lot variability and affect data reproducibility and biological relevance (Egelhofer et al., 2011). Due to a lack of understanding and available chromatin data in endoderm and subsequent derivatives, ChIP-qPCR was performed for TSS of well-characterized genes in hESCs (Fig 3.2). These genes include pluripotency markers (*SOX2*, *DPPA4*, *PRDM14*), general transcription/translation factors (*BTF3*, *EIF3B*), and housekeeping genes (*GAPDH*, *ACTB*), all of which are well-characterized and actively transcribed in hESC. In addition, chromatin states at TSS of three differentiation markers (*WNT8B*, *FOXP4*, *NODAL*), which are not expressed in hESC, were also assayed. We observed that the TSS of actively transcribed genes were strongly enriched with the activating marks H3K4me2 (112 – 372 fold), H3K4me3 (11 – 59 fold) and H3K27ac (8 – 102 fold), while relatively depleted of repressive H3K27me3 (1.6 – 15 fold) (Fig 3.2). Conversely, TSS of differentiation markers were highly marked with H3K27me3 (26 – 76 fold) but not H3K4me3 (2.4 – 3.5) and H3K27ac (2.6 – 8.4 fold). Notably, H3K4me2 was highly enriched (97

– 101 fold) at TSS of differentiation markers, consistent with previous observations that H3K4me2 marks developmentally poised genes (Orford et al., 2008). In sum, these data informed that the ChIP protocol and antibodies were efficient and specific in enriching for the appropriate histone marks and associated genomic DNA.

The validated ChIP conditions were subsequently applied on ES, PS and DE cells, and ChIP-enriched DNA from each experiment was used for sequencing library preparation. In total, 15 ChIP-seq libraries were constructed and sequenced (Table 3), culminating in a range of 37 – 57 million raw sequencing reads per library. Low quality reads were filtered through the standard Illumina pipeline, and pass-filtered reads were mapped to the human reference genome (hg19) using Bowtie (Langmead et al., 2009). In total, 23 to 38 million reads aligned uniquely per library, while non-specific reads mapping to more than one genomic locus were discarded from subsequent analyses.

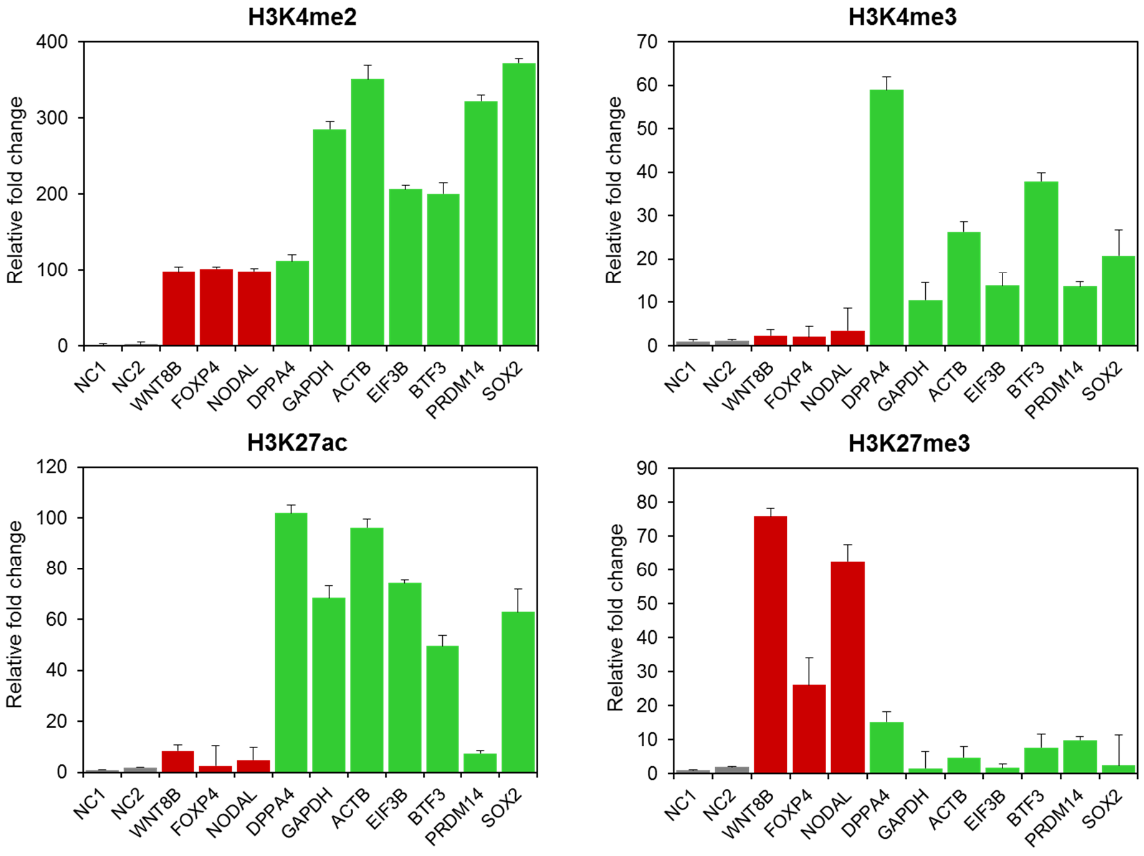


**Figure 3.2 ChIP-qPCR validation of histone modifications in ES cells**

ChIP DNA enriched with H3K4me2, H3K4me3, H3K27ac and H3K27me3 were amplified using TSS-specific primers. Expression levels were normalized against negative control regions (NC1 and NC2). Fold change values and standard deviations were determined from triplicate samples. Green columns represent genes which are actively transcribed in hESCs; red columns represent differentiation marker genes which are not transcribed in hESCs.

**Table 3 Statistics of ChIP-seq sequencing and mapping during endoderm differentiation**

| Cell Type | Sample | Sample Yield (Mbases) | Clusters (raw) | Clusters (PF) | % PF Clusters | Uniquely Aligned Clusters | % Align (PF) |
|---|---|---|---|---|---|---|---|
| ES | Input | 1,643 | 57,979,888 | 46,953,326 | 80.98 | 37,999,327 | 80.93 |
| | H3K4me2 | 1,134 | 40,221,845 | 32,407,688 | 80.57 | 28,071,539 | 86.62 |
| | H3K4me3 | 1,160 | 40,691,323 | 33,151,167 | 81.47 | 23,968,294 | 72.3 |
| | H3K27ac | 1,529 | 53,796,546 | 43,680,960 | 81.2 | 37,421,478 | 85.67 |
| | H3K27me3 | 1,384 | 48,853,675 | 39,539,641 | 80.93 | 28,812,536 | 72.87 |
| PS | Input | 1,423 | 48,428,268 | 40,668,117 | 83.98 | 32,778,502 | 80.6 |
| | H3K4me2 | 1,235 | 42,634,762 | 35,277,317 | 82.74 | 30,405,520 | 86.19 |
| | H3K4me3 | 1,177 | 39,937,664 | 33,621,173 | 84.18 | 24,852,771 | 73.92 |
| | H3K27ac | 1,299 | 44,327,988 | 37,109,028 | 83.71 | 31,761,617 | 85.59 |
| | H3K27me3 | 1,288 | 44,036,094 | 36,807,047 | 83.58 | 28,289,896 | 76.86 |
| DE | Input | 1,684 | 57,605,100 | 48,121,636 | 83.54 | 38,684,983 | 80.39 |
| | H3K4me2 | 1,254 | 42,599,651 | 35,819,126 | 84.08 | 30,360,291 | 84.76 |
| | H3K4me3 | 1,128 | 37,717,885 | 32,226,238 | 85.44 | 24,024,660 | 74.55 |
| | H3K27ac | 1,483 | 50,508,880 | 42,379,490 | 83.91 | 36,162,419 | 85.33 |
| | H3K27me3 | 1,215 | 40,966,068 | 34,701,839 | 84.71 | 27,362,400 | 78.85 |

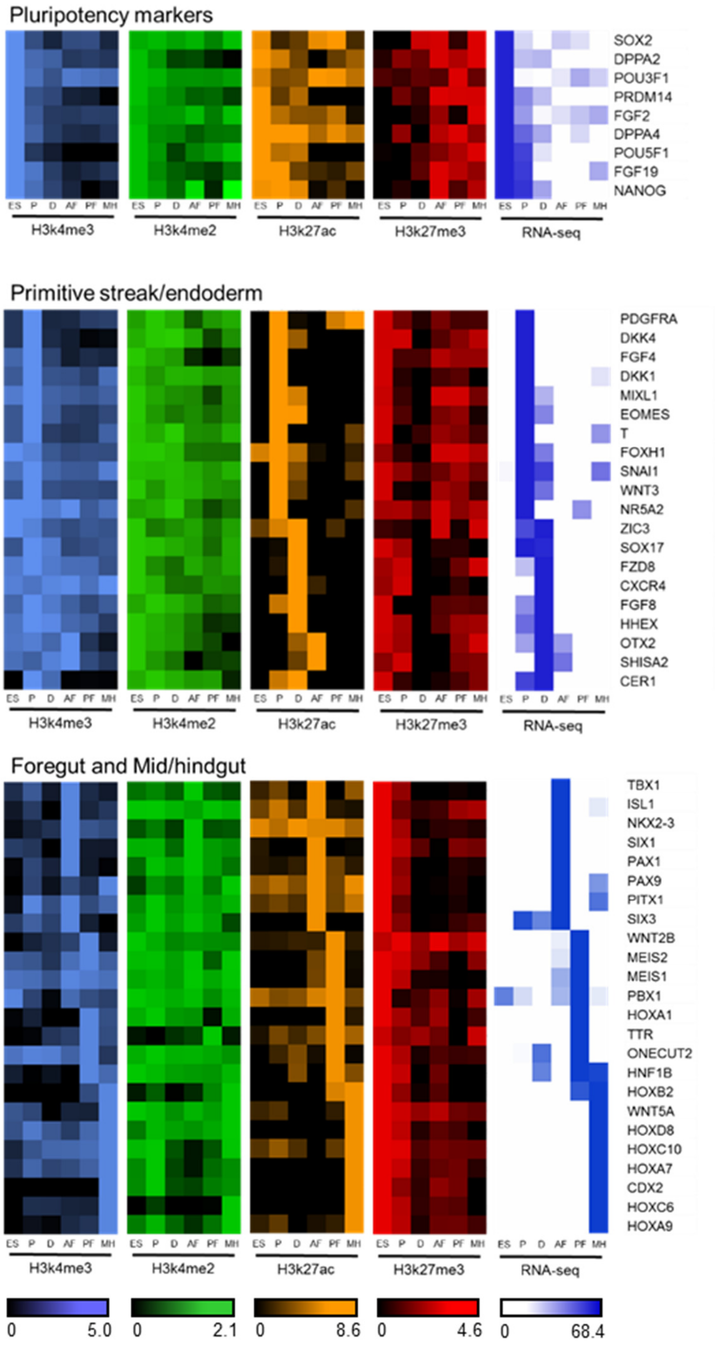
## 3.3 Promoter chromatin signatures for key regulators of endoderm differentiation

To investigate promoter chromatin states underlying the dynamic, stage-specific gene expression patterns observed during endoderm differentiation, histone modification levels were examined at the promoters of 51 stage-specific genes previously identified by RNA-seq. ChIP-seq tag counts within +/- 2 kb of these promoters were calculated and input-normalized for all 6 cell types (Fig 3.3A). Generally, mRNA expression at each developmental stage is associated with promoter enrichment of H3K4me3 and H3K27ac, with depletion of H3K27me3, as illustrated at the *PRDM14* locus (Fig 3.3B). Notably, H3K4me2 enrichment on these promoters was largely invariant with expression levels. Although H3K4me2 and H3K4me3 often co-localize on active promoters, H3K4me2 also marks developmentally poised genes (Orford et al., 2008). H3K4me2 enrichment on promoters during early developmental stages may serve to facilitate subsequent reactivation of these genes in later developmental stages.

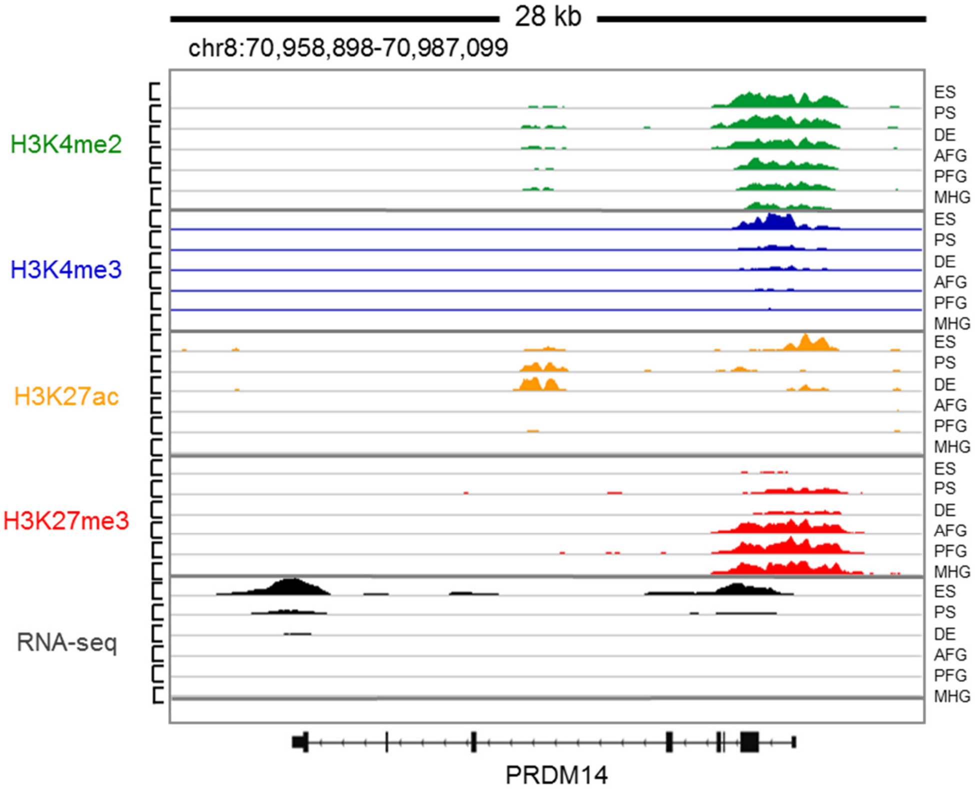
Developmental genes exhibit a bivalent chromatin structure in ES cell stage (Bernstein et al., 2006), yet it is unclear how genes activated during different developmental stages differ in their poising status. To investigate this, we analyzed the chromatin states of PS/DE and gut progenitor genes (AFG, PFG and MHG) in the ES cell stage. 80% of PS/DE genes, which were activated within 24 hours of differentiation, were bivalent, compared to only 25% of gut progenitor marker genes (Fig 3.3C). In contrast, 75% of these gut progenitor genes, which were activated only after 4 days of differentiation, were solely marked by repressive H3K27me3. These observations suggest that promoter bivalency is established in a developmental-stage specific manner, and that epigenetic mechanisms exist to maintain a ‘chromatin hierarchy’ in ES cells, ensuring a permissive state for early developmental genes, while maintaining a repressive state for later developmental genes.

Our integrative analysis of chromatin states in relation to expression patterns further revealed that a subset of developmental genes with similar expression patterns exhibit distinct chromatin profiles. For example, both *CXCR4* and *CER1* exhibit similar expression profiles at the DE stage, yet H3K4me3 persisted in gut progenitors at the *CXCR4* promoter, while *CER1* promoter lost H3K4me3 and regained repressive H3K27me3 (Fig 3.3D,E). *CXCR4* encodes a chemokine receptor involved in diverse developmental processes from regional specification of endoderm and pancreas (Katsumoto and Kume, 2013), to intestinal epithelial development (Zimmerman et al., 2011) and neuronal development (Stumm et al., 2003). Conversely, targeted disruption of *Cer1*, a TGF-β signaling inhibitor in anterior endoderm, showed that the gene is not required for later development (Stanley et al., 2000). Hence, the differences in chromatin states between these two genes may reflect their varying roles and expression in later development stages, suggesting that distinct epigenetic mechanisms exist to regulate developmental gene expression.

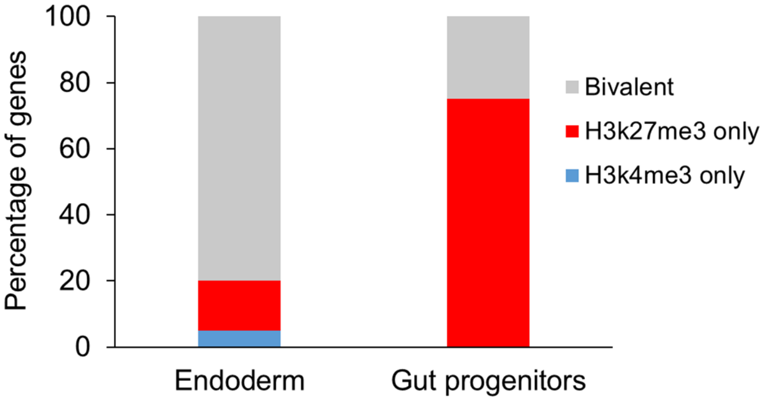
A)


Pluripotency markers
SOX2
DPPA2
POU3F1
PRDM14
FGF2
DPPA4
POU5F1
FGF19
NANOG
ES P D AF PF MH
H3k4me3
H3k4me2
H3k27ac
H3k27me3
RNA-seq
Primitive streak/endoderm
PDGFRA
DKK4
FGF4
DKK1
MIXL1
EOMES
T
FOXH1
SNAI1
WNT3
NR5A2
ZIC3
SOX17
FZD8
CXCR4
FGF8
HHEX
OTX2
SHISA2
CER1
Foregut and Mid/hindgut
TBX1
ISL1
NKX2-3
SIX1
PAX1
PAX9
PITX1
SIX3
WNT2B
MEIS2
MEIS1
PBX1
HOXA1
TTR
ONECUT2
HNF1B
HOXB2
WNT5A
HOXD8
HOXC10
HOXA7
CDX2
HOXC6
HOXA9
0 5.0
0 2.1
0 8.6
0 4.6
0 68.4

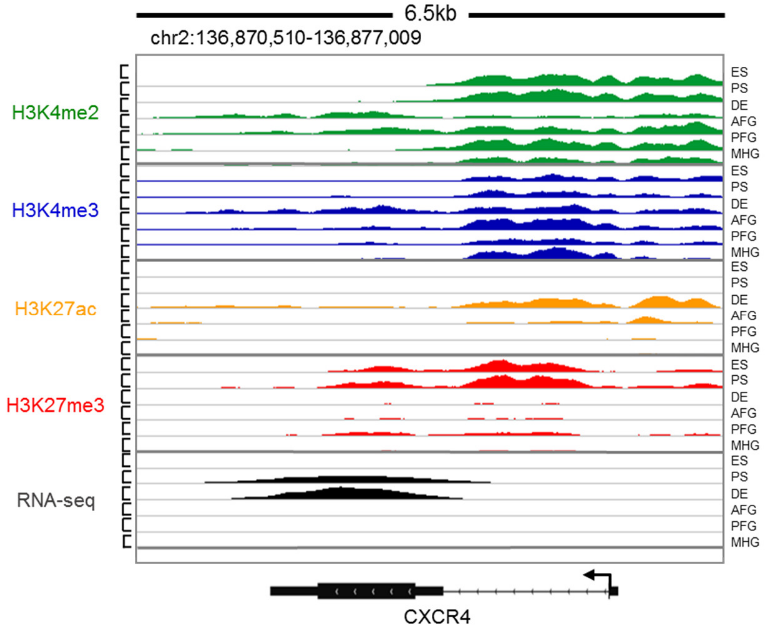
B)


28 kb
chr8:70,958,898-70,987,099
H3K4me2
H3K4me3
H3K27ac
H3K27me3
RNA-seq
ES
PS
DE
AFG
PFG
MHG
PRDM14

C)


100
80
60
40
20
0
Percentage of genes
Endoderm
Gut progenitors
Bivalent
H3k27me3 only
H3k4me3 only

D)



E)

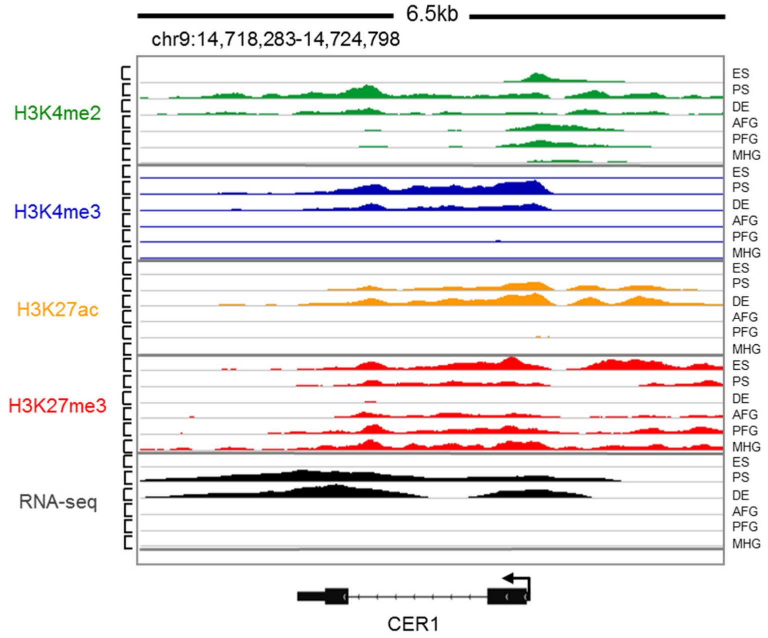


**Figure 3.3 Histone modification signatures around promoters of lineage specific genes during endoderm differentiation**

(A) Heatmaps of histone modification and RNA-seq read enrichment at TSS of 51 stage-specific transcripts. (B) Chromatin signatures at the PRDM14 locus (C) Proportion of genes in endoderm and gut progenitors with promoter enrichment of H3K4me3, H3K27me3 or both. (D-E) Chromatin signatures at the *CXCR4* and *CER1* locus.

## 3.4 Identification and epigenomic profiling of endoderm enhancers

Although promoters are important in regulating transcription initiation, enhancers are the key drivers of cell-type specific gene expression programs (Bulger and Groudine, 2011; Heintzman et al., 2009). Several genome-wide studies have successfully leveraged on H3K27ac as a tissue- and developmental stage-specific chromatin marker for active enhancer elements (Cotney et al., 2012; Creyghton et al., 2010). Furthermore, the dynamics of H3K27ac enrichment are associated with gene expression changes during the transition from pluripotency to cell specification (Bogdanović et al., 2012). We therefore focused on H3K27ac profiling for endoderm enhancer discovery.

### 3.4.1 H3K27ac peak profiling identifies putative endoderm enhancers

To predict active enhancer elements and study their activation dynamics during endoderm differentiation, mapped H3K27ac reads were used to identify enhancer peaks. In total, 8,653, 11,510 and 13,993 H3K27ac peaks were identified in ES, PS and DE respectively. Consistent with the association of H3K27ac with active promoters and enhancers (Bonn et al., 2012; Ernst et al., 2011; Heintzman et al., 2009), we observed that for all 3 cell types, ~45% of H3K27ac+ elements were TSS-proximal (0 – 2kb), while the remaining were TSS-distal ( > 2kb) (Fig 3.4A). We considered TSS-distal, H3K27ac+ loci as putative enhancers and identified 13,367 such peaks, which were further subjected to *k*-means clustering to obtain cell-type specific peaks. From this analysis, 2,052, 2,155 and 4,320 H3K37ac peaks were identified in ES, PS/DE and DE respectively (Fig 3.4B). We hypothesized that peaks found in PS/DE and DE contain active enhancers which drive an endoderm transcriptional program, and merged them for subsequent analyses (thereafter referred to as 'putative DE enhancers'). Compared to flanking regions, putative DE enhancers were more evolutionary-constrained (Fig 3.4C), suggesting functional conservation.

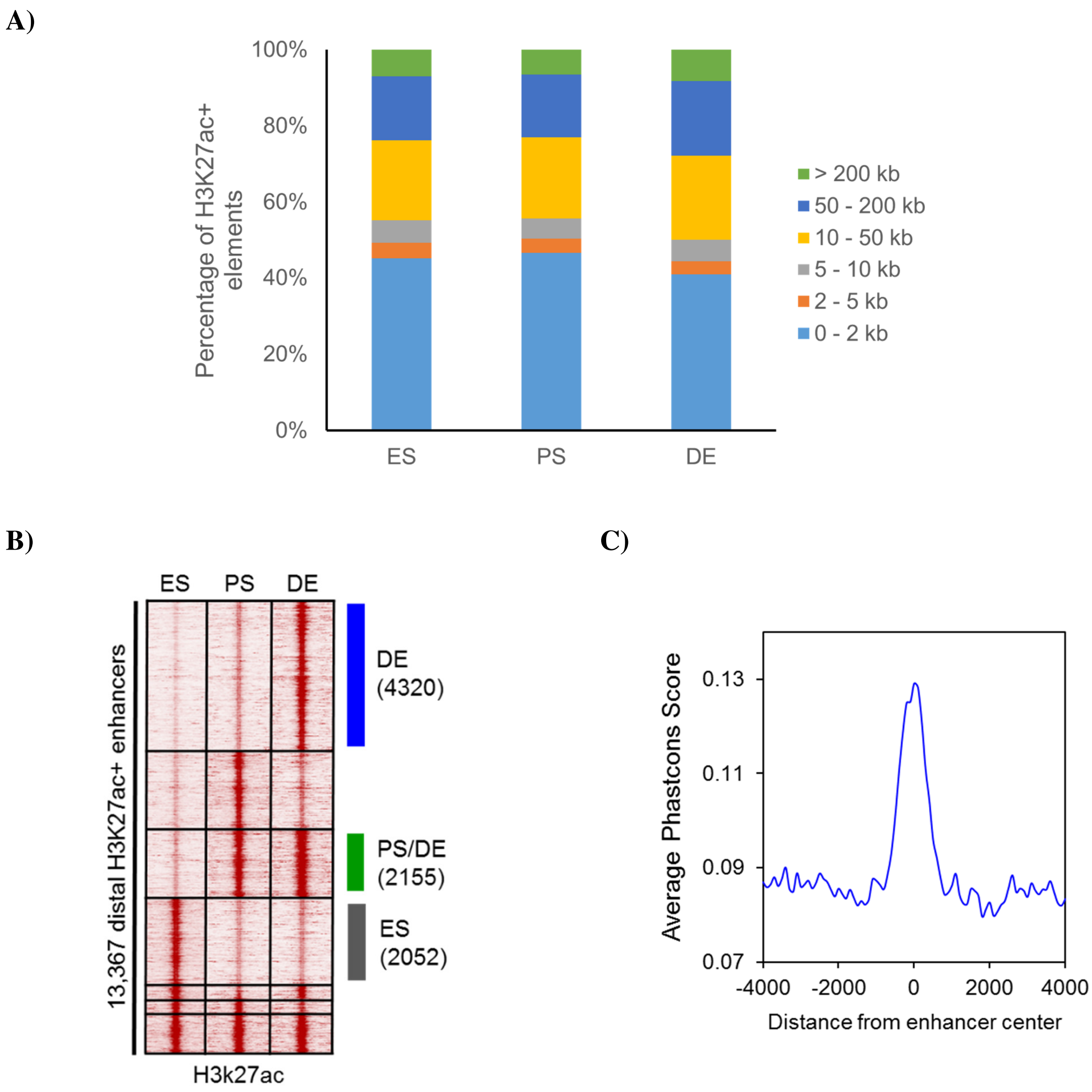


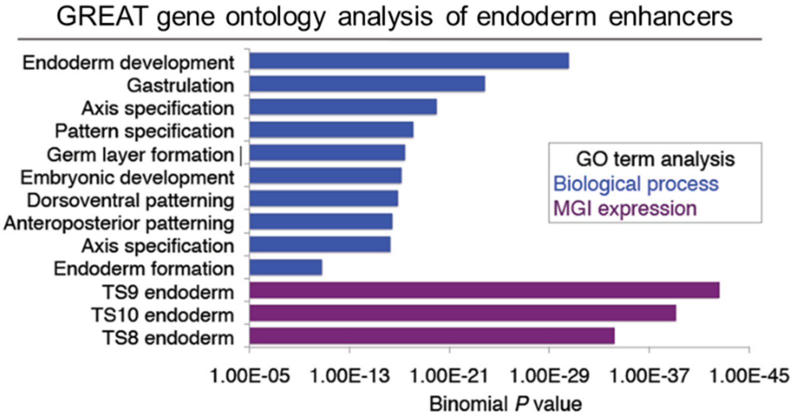
**Figure 3.4 Epigenomic profiling to identify DE enhancers**

(A) Number of H3K27ac elements identified in ES, PS and DE, and their proportion according to distance away from the nearest TSS. (B) Heatmap illustrating *k*-means clustering (*k*=7) of all distal H3K27ac+ elements. (C) Phastcons score of DE enhancers.

### 3.4.2 Functional annotation of putative DE enhancers

Having identified a list of putative DE enhancers, we next asked whether these elements resided near genes involved in endoderm development. An unsupervised gene ontology statistical analysis, GREAT (McLean et al., 2010), was performed to assign each peak to annotated genes, followed by an analysis of functional gene annotation enrichment among the assigned genes. Notably, gene ontology analysis associated these peaks significantly with genes involved in endoderm development ($p=4.53\times10^{-30}$), gastrulation ($p=1.07\times10^{-25}$), and axis specification ($p=5.82\times10^{-18}$) (Fig 3.5A). Furthermore, these genes exhibit robust *in vivo* expression in mouse endoderm according to expression patterns recorded in the Mouse Genome Database (Bult et al., 2008). To test whether these peaks correlate positively with proximal gene expression in endoderm, expression levels of transcripts located within 50kb from each putative DE enhancer were measured and compared in each cell-type (Fig 3.5B). Importantly, genes proximal to putative DE enhancers exhibited significantly higher expression in DE, demonstrating a positive correlation between tissue-specific enhancer activation and gene expression patterns in the appropriate cell-type. Collectively, these data suggest that the H3K27ac+ peaks identified through global chromatin state profiling of *in vitro* differentiation cultures represent bona fide enhancers relevant to endoderm development.

A)



B)

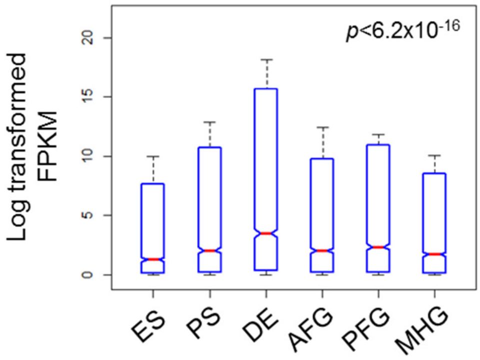


**Figure 3.5 Functional annotation of putative DE enhancers**

(A) Gene ontology terms associated with DE enhancers. Blue columns represents biological processes, while purple columns refers to Mouse Genome Informatics (MGI) data of curated mouse expression information. (B) Expression levels of transcripts within 50 kb of putative DE enhancers measured in all 6 cell-types. The p-value stated represents the least significant p-value of all pairwise comparisons with respect to expression levels in DE.

## 3.5 Convergence of endoderm lineage-specifying TFs with TGF-β signaling effectors for enhancer activation

Genome-wide analyses of enhancer function revealed that in addition to core TFs, enhancers are also bound by DNA-binding effectors of various signaling pathways, leading to the recognition of enhancers as integration hubs between extrinsic cell signaling, genomic sequence and epigenetic information (Calo and Wysocka, 2013; Chen, 2008; Mullen et al., 2011). We sought to investigate whether our putative DE enhancers bind core endoderm TFs or signaling effectors, and if so, how these different classes of factors are coordinated for the enhancer activation.

### 3.5.1 Binding site analysis of putative DE enhancers

We hypothesized that DNA motif enrichment at putative DE enhancers will reveal TFs and signaling effectors driving endoderm enhancer activation for the initiation of downstream transcriptional programs. Binding site analysis revealed significant enrichment of motifs for GATA4, NANOG, EOMES, FOXA2, and AP-1, as well as the NODAL signaling effectors SMAD2/3, SMAD4 and FOXH1 on putative DE enhancers (Fig 3.6A). Notably, GATA4, NANOG, EOMES and FOXA2 were all reported to play key regulatory roles in endoderm development (Rojas et al., 2010; Teo et al., 2011). The enrichment of motif sequences of the AP-1 transcriptional complex is consistent with the role of JNK kinases in regulating endoderm development (Loebel et al., 2011; Xu and Davis, 2010), illustrating a novel link between the JNK-AP1 signal transduction pathway and the active enhancer landscape in regulating endoderm lineage specification. Taken together, these observations demonstrate the association of putative DE enhancers with multiple core TFs of the TGF-β signaling pathway, as well as master regulatory TFs in endoderm development, supporting a function of these elements as enhancers in regulating endoderm gene expression.

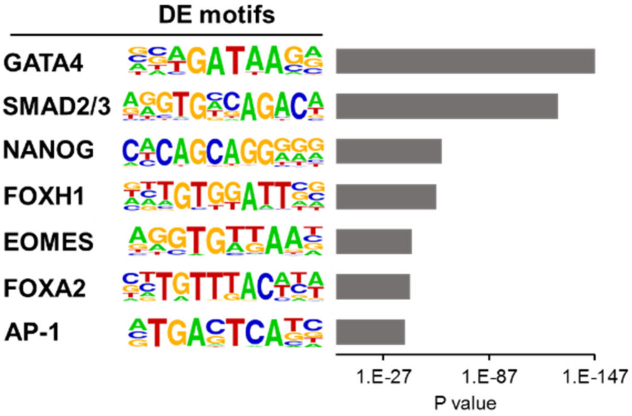


**Figure 3.6 Overrepresented DNA sequence motifs enriched at putative DE enhancers**

Position weight matrix logo and p-value of the top sequence motifs enriched at putative DE enhancers.

### 3.5.2 Association of putative DE enhancers with endoderm-specifying TFs

To validate the motif analyses data and evaluate whether putative DE enhancers are broadly associated with lineage-specifying TFs/signaling effectors, we adopted an integrative genomics approach using SMAD2/3, SMAD4, FOXH1 (Kim et al., 2011), FOXA2 (Gifford et al., 2013) and EOMES (Teo et al., 2011) binding profiles in endoderm, as well as our DE chromatin modifications. We first calculated the correlation of tag counts of these factors with chromatin modifications in ES, PS and DE (Fig 3.7A). EOMES and FOXH1 peaks co-clustered with H3K27ac in PS and DE (Cluster 1), while SMAD2/3/4 and FOXA2 peaks co-clustered with H3K4me2 in DE (Cluster 2). Notably, both clusters are characterized by enhancer-associated histone modifications. These two clusters formed a super-cluster in relation to clusters 3 and 4, which are characterized by promoter (H3K4me3) and repressive (H3K27me3) chromatin modifications, indicating that endodermal TFs broadly converged onto putative DE enhancers. To investigate the cell-type specificity of this overlap, we compared TF ChIP-seq intensities at putative DE enhancers with neural crest enhancers (Rada-Iglesias et al., 2012). Notably, binding of all 5 TFs enriched significantly at DE enhancers over neural crest enhancers ($p < 2.2 \times 10^{-4}$) (Fig 3.7B). To further characterize the degree of association between endoderm-specifying TFs and putative DE enhancers, we calculated the extent of overlap between their respective binding peaks. Out of 6475 H3K27ac+ peaks in DE, 5342 (82.5%) overlapped with both SMAD2/3 and EOMES (Fig 3.7C), indicating extensive convergence between putative DE enhancers and endoderm TFs.

The observation that multiple NODAL signaling effectors converged onto putative DE enhancers raised the question of whether putative DE enhancers modulate NODAL signaling transcriptional output. To investigate this, gene set enrichment analysis (GSEA) (Subramanian et al., 2005) was performed to query putative DE enhancer-proximal genes against differentially-expressed genes during endoderm differentiation induced by Activin treatment on hESCs (Bernardo et al., 2011) (Fig 3.7D). Strikingly, the top 500 putative DE enhancer-associated genes were significantly enriched in the gene set positively regulated by Activin (ES = 0.51, $p$ value < 0.001), suggesting that putative DE enhancers positively regulate the expression of endoderm differentiation genes induced by NODAL. Collectively, these data suggest that endoderm-specifying TFs drive endoderm transcriptional program through binding to, and subsequent activation of, putative DE enhancers.

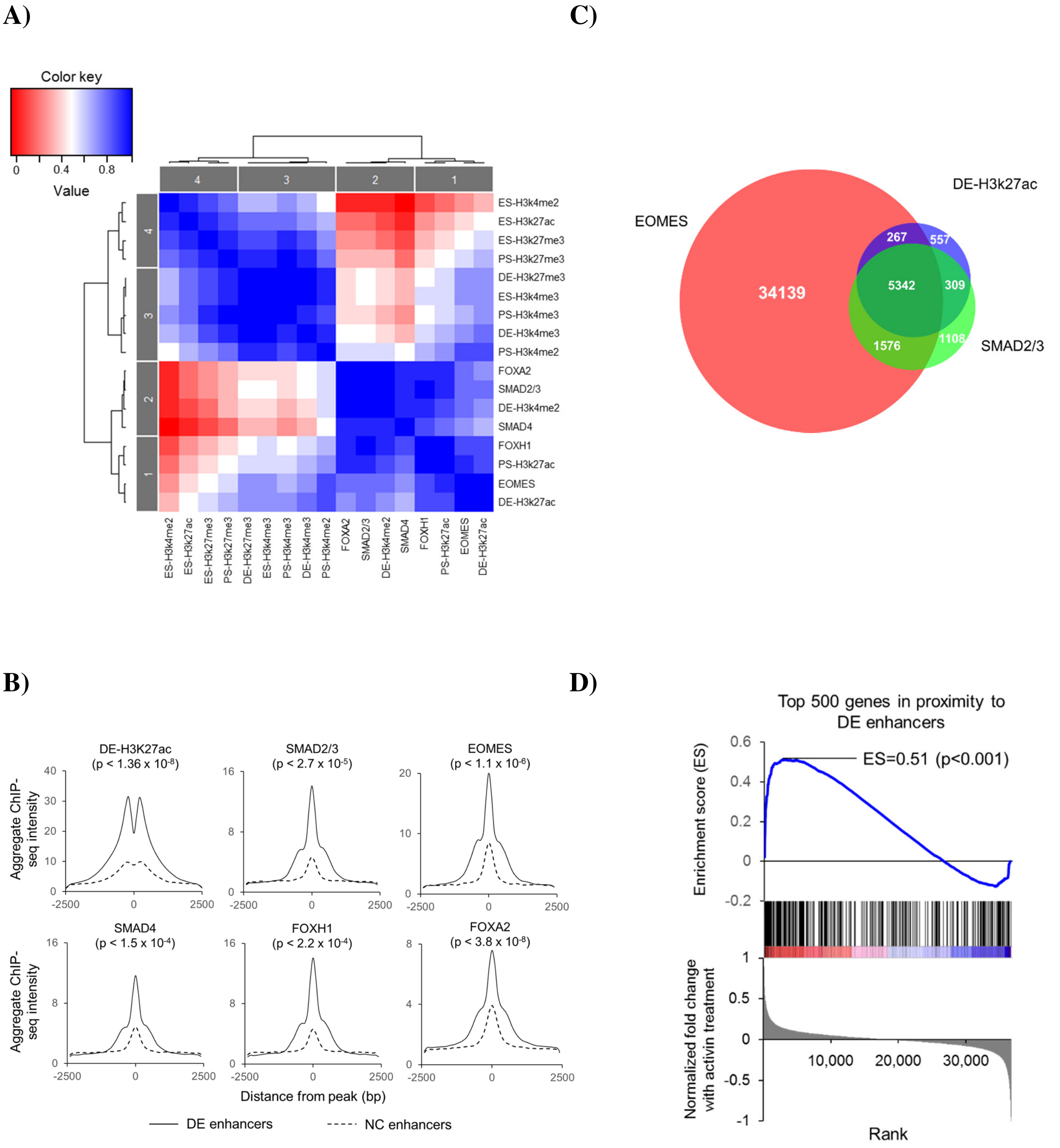


**Figure 3.7 TGF-β signaling effectors and endoderm-specifying TFs bind at putative DE enhancers *in vivo***

(A) Hierarchical clustering of TF ChIP-seq datasets with ES, PS and DE chromatin modifications. (B) ChIP-seq profiles of SMAD2/3, SMAD4, EOMES, FOXA2 and FOXH1 and H3K27ac at DE enhancers. (C) ChIP-seq peak counts for EOMES, SMAD2/3 and H3K27ac in

DE, and the extent of overlap between these peaks. (D) GSEA analysis of top 500 putative DE enhancer proximal genes within the gene set positively regulated by Activin (Bernardo et al., 2011). Top panel: ES generated through a running-sum statistic, as each gene in the ranked expression profile is considered one by one. ES is increased when a DE enhancer proximal gene is observed, and decreased when a DE enhancer proximal gene is not observed. Each black vertical line below the ES score represents a single gene in the ranked expression profile, which may be increased (red) or decreased (blue) by Activin treatment compared to untreated controls. Bottom panel: ranked gene list, consisting of 36,704 genes and their corresponding mean expression value.

### 3.5.3 Coordination between TGF-β effectors and EOMES for enhancer activation

An interestingly observation from the analyses described above was that only a small fraction of all EOMES binding peaks (12.9%: 5342 out of 41324) coincided with both SMAD2/3 and putative DE enhancers (Fig 3.7C), prompting us to question the function of the remaining peaks. To study this, we examined the association between EOMES and TGF-β signaling effectors with H3K27ac enrichment in DE. Binding profiles of these TFs were examined in relation to H3K27ac in DE using a non-supervised clustering approach (Fig 3.8A). EOMES binding alone correlated with weak H3K27ac in DE (Class I), as per TGF-β effectors in the absence of EOMES (Class II). Notably, maximal DE enhancer acetylation was observed only with co-localization of EOMES with all 3 NODAL effectors (Class III) (p value < $10^{-84}$, Fig 3.8B), indicating that endoderm-specifying effects of TFG-β signaling is determined by binding of effectors to a small subset of EOMES binding sites.

The large proportion of Class I (36704/53902, 68.1%) binding sites suggested that these regulatory loci may be associated with developmental programs driven by EOMES independent of TGF-β signaling. EOMES has been reported to regulate neurogenesis (Arnold et al., 2008a; Hodge et al., 2012), trophoblast and mesoderm development, as well as endoderm specification (Arnold and Robertson, 2009; Arnold et al., 2008b; Russ et al., 2000), and is thus considered a master regulator of various mesodermal- and endodermal-derived cell lineages. We hypothesized that EOMES-bound sites which do not overlap DE enhancers may represent regulatory elements driving non-endodermal developmental lineages. Consistently, gene ontology analysis of Class I EOMES binding sites revealed that these loci were significantly associated with genes involved in neurogenesis ($p$=1.7 × $10^{-19}$), mesoderm morphogenesis ($p$=1.8 × $10^{-11}$) and heart development

($p$=1.4 × $10^{-10}$) (Fig 3.8C). The integration of TF binding datasets with developmental stage-specific chromatin profiles allowed the genetic dissection of TF loci involved in transcriptional programs of different developmental lineages.

In sum, our data suggest that convergence of extrinsic signaling pathways and lineage-specifying TFs onto multiple TF-binding enhancer loci controls the endodermal cell fate, possibly through regulating key endoderm lineage specifiers and/or markers, such as *SOX17* and *CER1* (Fig 3.8D). Multiple TF-bound loci have previously been reported in mouse ES cells and heart to drive tissue-specific gene expression (Chen et al., 2008; He et al., 2011), and our analyses have provided a comprehensive catalogue of such multiple TF-bound loci in early human endoderm development.

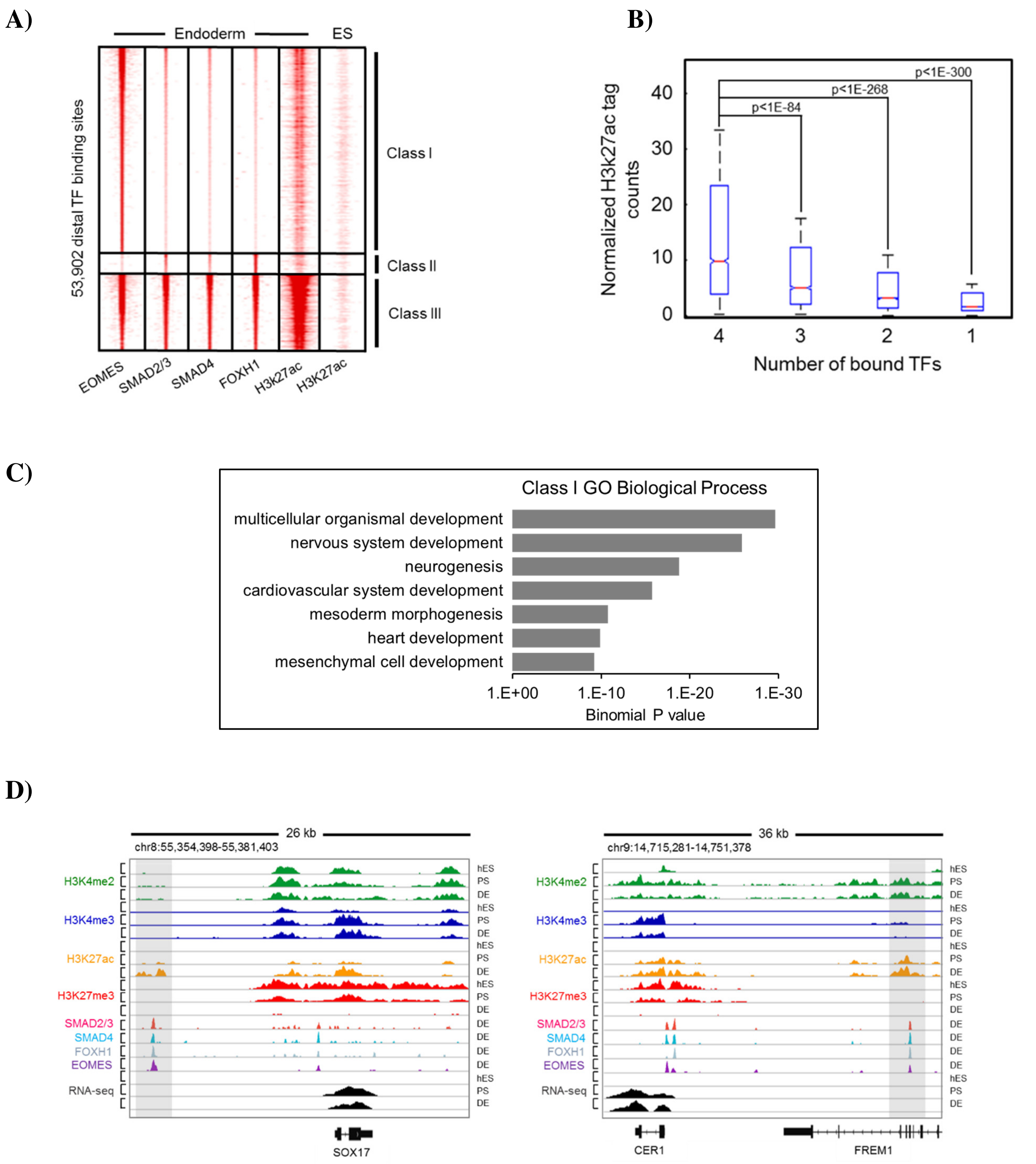


**Figure 3.8 Genetic dissection of endoderm TF binding sites**

(A) Clustering of endoderm TF binding sites with H3K27ac signals in DE and ES. (B) Average H3K27ac tag counts in DE computed based on number of endoderm TFs bound. TFs included in the analysis are SMAD2/3, SMAD4, EOMES and FOXH1. (C) GO terms associated with Class I

EOMES binding sites identified from the clustering analysis in (A). (D) Examples of key endoderm marker gene loci regulated by multiple TF-bound putative DE enhancers.

## 3.6 Endoderm enhancer priming

We next focused on examining the mechanisms facilitating enhancer activation during DE differentiation. TF binding (Zaret and Carroll, 2011) and chromatin structure (Calo and Wysocka, 2013; Rada-Iglesias et al., 2011) have been proposed to pre-mark distal regulatory elements for rapid target gene activation. Although SMAD2/3, SMAD4 and FOXH1 extensively occupied DE enhancers following differentiation, they are significantly less enriched at the same loci in ES cells (Fig. 3.9A). Pre-marking with H3K4me1 has been observed to correlate with developmental enhancer poising (Calo and Wysocka, 2013; Rada-Iglesias et al., 2011), and we hypothesized that additional priming mechanisms may exist. To comprehensively investigate the relationship between chromatin structure and enhancer poising, we overlapped putative DE enhancers with 17 histone modifications and chromatin modifiers in human ES cells (Ernst et al., 2011). Surprisingly, unsupervised clustering revealed that only ~30% of putative DE enhancers were pre-marked with H3K4me1, indicating that this histone modification alone does not comprehensively identify all poised developmental enhancers (Fig 3.9B). We identified repressive H3K9me3 (cluster 2) as a poising mark on distal regulatory elements, extending a previously reported function of H3K9me3 in pre-marking promoters of transcriptional targets of the NODAL/TGF-β pathway during endoderm formation (Xi et al., 2011). In addition, we discovered an enhancer cluster pre-marked predominantly by H4k20me1 (cluster 6). Interestingly, H4k20me1 also plays a central role in regulating RNAPII promoter-proximal pausing (Kapoor-Vazirani and Vertino, 2014), suggesting that a common epigenetic mechanism may regulate both enhancer poising and RNAPII pausing for rapid transcriptional output. We further observed ~25% of DE enhancers pre-marked solely by H2A.Z (cluster 1). H2A.Z is a variant of the H2A histone subunit involved in transcription regulation through its effects on chromatin structure (Goldman et al., 2010; Thambirajah et al., 2006), and was recently demonstrated to be enriched at promoters and enhancers in hESCs during ES self-renewal and differentiation (Hu et al., 2013). Importantly, H2A.Z decreases nucleosome stability and occupancy, promoting nucleosome displacement by TFs (Jin et al., 2009). Hence, H2A.Z may facilitate rapid and preferential endoderm TF binding and enhancer activation. Indeed, upon ES differentiation, H2A.Z-marked pre-enhancers were significantly more enriched in SMAD2/3, SMAD4, FOXH1 and EOMES compared to unmarked, latent pre-enhancers (cluster 5) (*p* value $< 10^{-13}$) (Fig 3.9C). Collectively, our analyses revealed novel and diverse combinations of chromatin priming modifications, including H3K9me3, H4k20me1, and H2A.Z, which complements and expands our current understanding of enhancer priming mechanisms.

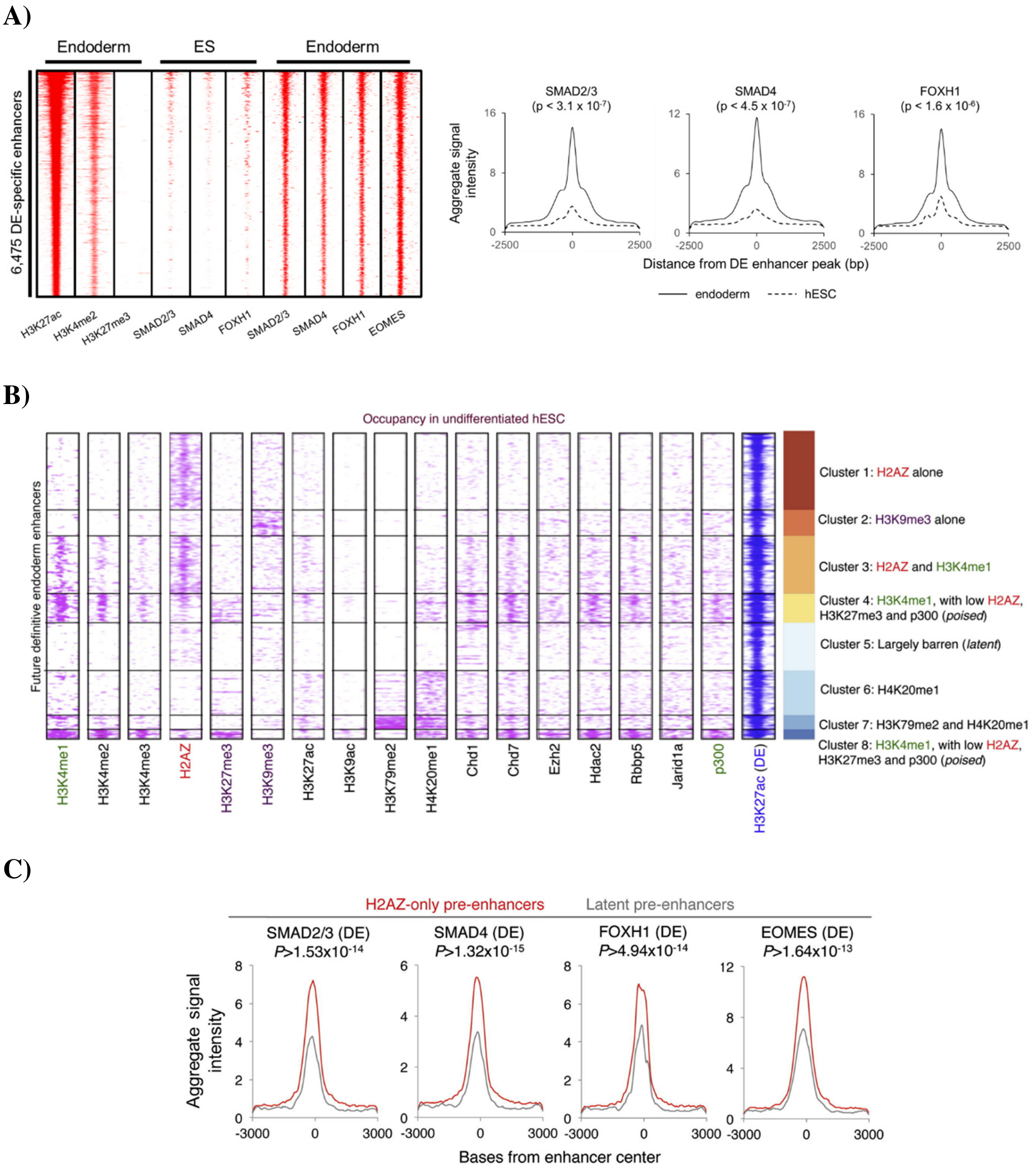


**Figure 3.9 A diversity of DE enhancer priming states**

(A) (left) ChIP-seq profiles of endoderm-specifying TFs (SMAD2/3, SMAD4, FOXH1 and EOMES) in ES cells and endoderm at 6,475 active DE enhancers. EOMES is not expressed in ES and no binding data is available. (right) Average signals of SMAD2/3, SMAD4 and FOXH1 centered at DE enhancers in endoderm and ES cells. (B) Heatmap illustrating ChIP-seq enrichment profiles for 17 chromatin modifications and epigenetic modifiers in hESCs at DE enhancers. (C) Average TF binding signals at H2AZ-primed enhancers (red) in comparison to latent pre-enhancers (grey). Significance of difference between these two classes were obtained using the two-sample Wilcoxon test on the sum of reads within a 6 kb window.

# CHAPTER 4 ENHANCER DISCOVERY, ANNOTATION AND VALIDATION DURING ENDODERM PATTERNING

## 4.1 Enhancer discovery and epigenome profiling during endoderm patterning

We performed epigenome mapping of H3K4me2, H3K4me3, H3K27ac and H3K27me3 through ChIP-seq in AFG, PFG and MHG progenitor cells. A total of 15 libraries were successfully sequenced and mapped, generating 25 to 35 million reads per library (Table 4). We identified a total of 16,158 gene-distal H3K27ac+ peaks from all three cell types and performed *k*-means clustering to identify cell-specific peaks, resulting in 4,071, 5,163 and 2,620 peaks representing putative enhancers in AFG, PFG and MHG respectively (Fig 4.1A). Consistent with an active enhancer signature, these H3K27ac+ elements were also strongly enriched in H3K4me2, while relatively depleted in H3K4me3 and H3K27me3 (Fig 4.1B). Notably, the average H3K27ac and H3K4me2 signals at these peaks exhibited a characteristic nucleosomal displacement "dip" suggestive of trans-factor binding. Furthermore, these peaks resided on genomic loci which were generally evolutionarily constrained, exhibiting a higher conservation score than flanking background regions (Fig 4.1C) suggesting functional conservation.

**Table 4 Statistics of ChIP-seq sequencing and mapping during endoderm patterning**

| Cell Type | Sample | Sample Yield (Mbases) | Clusters (raw) | Clusters (PF) | % PF Clusters | Uniquely Aligned Clusters | % Align (PF) |
|---|---|---|---|---|---|---|---|
| AFG | Input | 1,485 | 52,995,835 | 42,425,219 | 80.05 | 34,203,212 | 80.62 |
|  | H3K4me2 | 1,261 | 45,215,583 | 36,016,254 | 79.65 | 31,503,417 | 87.47 |
|  | H3K4me3 | 1,200 | 42,481,592 | 34,289,338 | 80.72 | 25,343,250 | 73.91 |
|  | H3K27ac | 1,451 | 51,943,097 | 41,460,711 | 79.82 | 36,427,381 | 87.86 |
|  | H3K27me3 | 1,281 | 45,439,398 | 36,586,138 | 80.52 | 28,987,197 | 79.23 |
| PFG | Input | 1,370 | 47,407,892 | 39,138,045 | 82.56 | 31,423,936 | 80.29 |
|  | H3K4me2 | 1,147 | 40,098,063 | 32,781,185 | 81.75 | 28,808,105 | 87.88 |
|  | H3K4me3 | 1,224 | 42,103,096 | 34,958,923 | 83.03 | 26,369,516 | 75.43 |
|  | H3K27ac | 1,328 | 45,984,718 | 37,947,007 | 82.52 | 33,321,267 | 87.81 |
|  | H3K27me3 | 1,356 | 46,984,836 | 38,752,066 | 82.48 | 31,017,154 | 80.04 |
| MHG | Input | 1,528 | 52,945,040 | 43,654,736 | 82.45 | 35,469,473 | 81.25 |
|  | H3K4me2 | 1,073 | 37,199,292 | 30,663,675 | 82.43 | 26,708,061 | 87.1 |
|  | H3K4me3 | 1,184 | 40,429,661 | 33,831,958 | 83.68 | 25,238,641 | 74.6 |
|  | H3K27ac | 1,309 | 45,449,581 | 37,394,612 | 82.28 | 32,219,198 | 86.16 |
|  | H3K27me3 | 1,237 | 42,719,651 | 35,329,148 | 82.7 | 28,295,115 | 80.09 |

PF = pass-filter

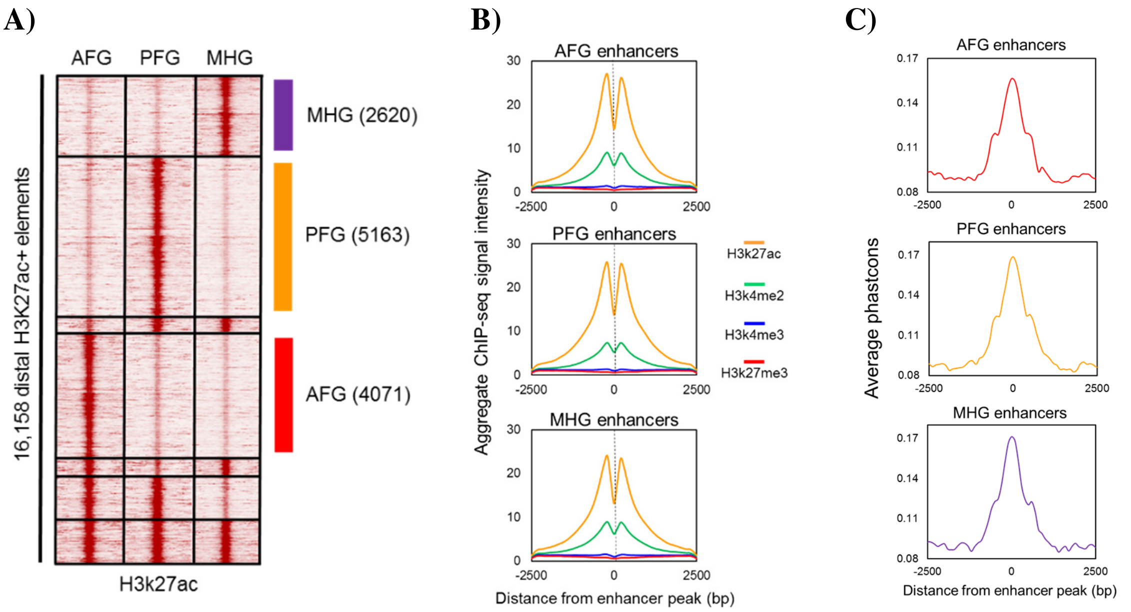


**Figure 4.1 Identification of cell-specific H3K27ac+ peaks following endoderm patterning**

(A) Heatmap of H3K27ac profiles grouped into cell-specific classes using *k*-means clustering. (B) Average signals of H3K4me2, H3K4me3, H3K27ac and H3K27me3 at AFG, PFG and MHG enhancers identified through clustering. (C) Sequence conservation (phastCons score) of AFG, PFG and MHG enhancers.

## 4.2 Functional annotation of enhancers

### 4.2.1 Gene ontology and RNA expression analysis

To assess whether the putative enhancers in AFG, PFG and MHG correlate with expression of proximal genes, transcript levels of genes within 50kb from each group of H3K27ac+ elements were computed from RNA-seq datasets (Fig 4.2A). Notably, these elements correlated significantly with increased proximal transcript levels within their respective domains. Functional annotation using GREAT (McLean et al., 2010) revealed associations of the identified elements with annotated developmental genes (Fig 4.2B). Specifically, AFG elements were associated with genes required for foregut morphogenesis ($p < 4.8\times10^{-12}$); PFG elements with genes involved in foregut ($p < 1.7\times10^{-12}$) and pancreas development ($p < 1.0\times10^{-8}$); MHG elements with genes relevant to mid- ($p < 7.1 \times10^{-14}$) and hindgut morphogenesis ($p < 3.7\times10^{-12}$). Importantly, these putative enhancers were also strongly associated with genes linked to endodermal developmental defects identified through mouse deletion studies (Bult et al., 2008) (Fig 4.2C), highlighting the potential of these putative enhancer catalogues as comprehensive resources for understanding the molecular etiology of human endodermal congenital diseases.

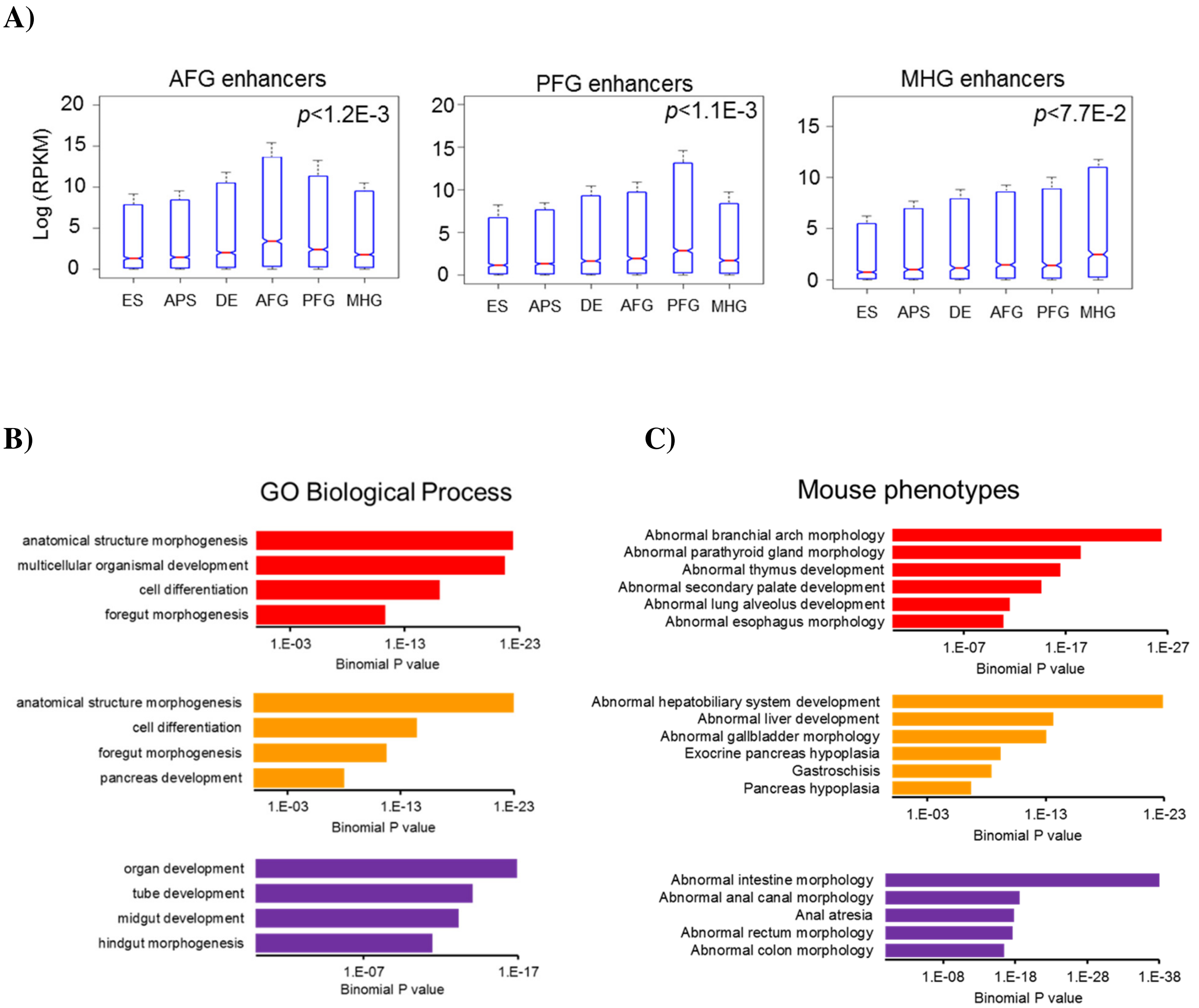


**Figure 4.2 Functional annotation of putative AFG, PFG and MHG enhancers**

(A) RNA-seq expression levels of transcripts proximal to each enhancer class. P values represent the least significant value between all pairwise comparisons involving the putative enhancers of interest. P values were calculated using paired t-test. (B) Gene ontology biological processes associated with each enhancer class. (C) Developmental defects associated with genes in proximity to enhancers from each enhancer class.

### 4.2.2 Motif analysis of cell-type-specific enhancers

To further establish cell-type-specific activities of these putative enhancers, we performed TF binding site analysis (Fig 4.3A). Among the most highly overrepresented AFG motifs included OTX2 ($p = 10^{-45}$) and SIX1 ($p = 10^{-29}$). Otx2 is a homeodomain-containing TF required for anterior patterning and is restricted in expression to the anterior endoderm (Ang et al., 1994; Jin et al., 2001). Six1 is a homeobox protein required for anterior organ development, including the thymus, lung and inner ear (Bricaud and Collazo, 2006; El-hashash et al., 2011; Laclef et al., 2003). Enriched PFG motifs included hepatic progenitor specifiers HNF1B ($p = 10^{-90}$), GATA6 ($p = 10^{-66}$) and FOXA2 ($p = 10^{-16}$), and pancreas differentiation factor PDX1 ($p = 10^{-50}$). Importantly, all these TFs play significant roles in organ-specification, specifically liver and pancreas, from the posterior foregut endoderm (Lee et al., 2005; Lokmane et al., 2008; Zhao et al., 2005) (Ahlgren et al., 1996). MHG enhancers were enriched in motifs for CDX2 ($p = 10^{-118}$), a master regulator of intestinal specification expressed in hindgut endoderm (Gao et al., 2009), as well as the WNT signaling effector TCF4 ($p = 10^{-47}$), required for proper hindgut formation and subsequent intestine development and homeostasis (van Es et al., 2012; Gregorieff et al., 2004; Verzi et al., 2010). To check whether the motif enrichment was specific for each type of progenitor cell, we sampled two motifs from each cell-type and calculated their motif densities. All motifs exhibited a sharp density peak corresponding to the cell-type they were enriched in (Fig 4.3B). In sum, TF binding site analyses corroborated foregut, midgut and hindgut progenitor-specific activities of the identified elements suggested by RNA-seq and Gene Ontology analyses.

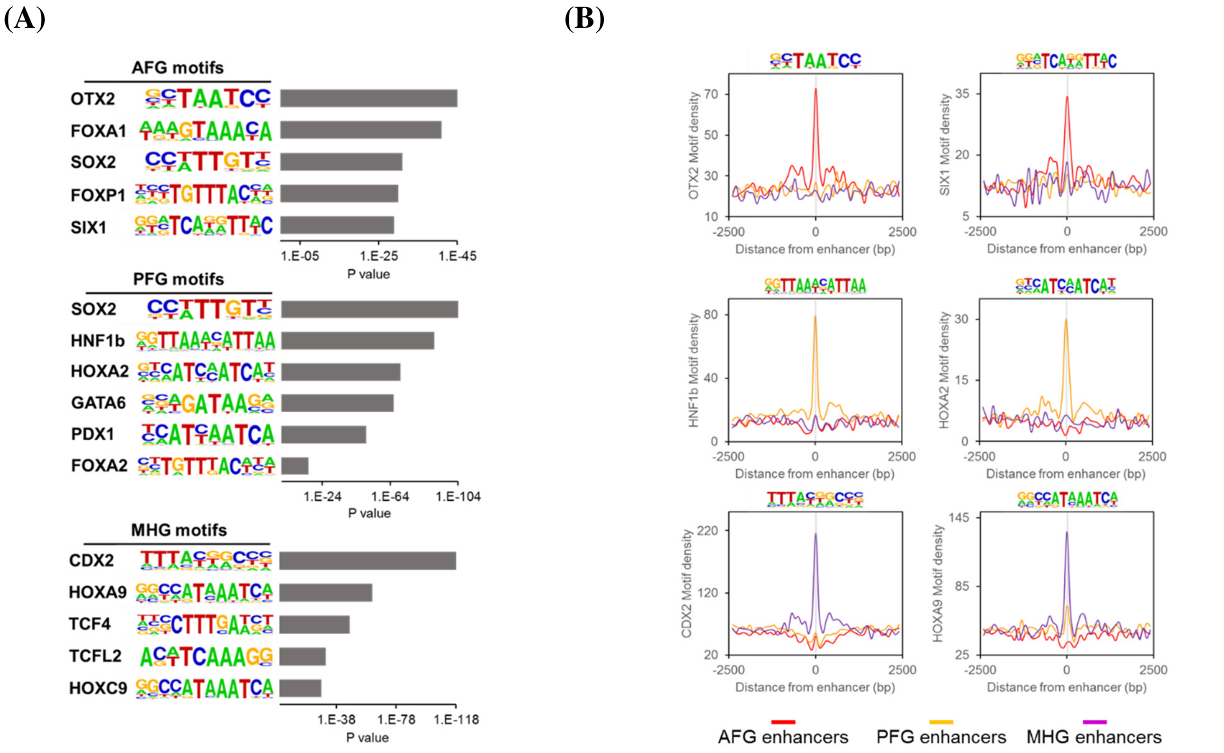


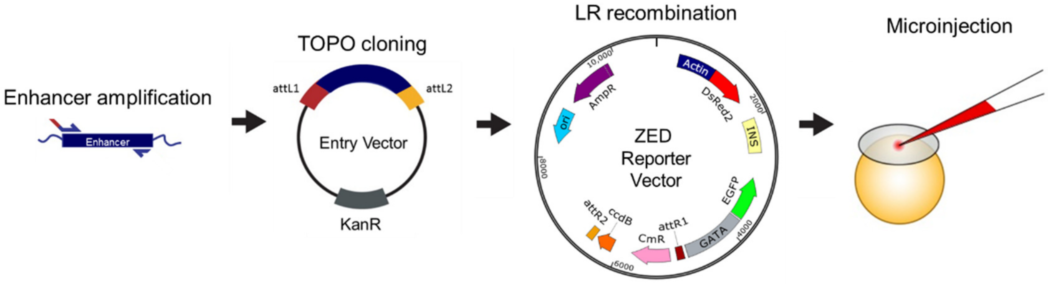
**Figure 4.3 TF binding site analysis on putative enhancers in AFG, PFG and MHG**

(A) TF motifs and their enrichment scores for each enhancer class. (B) Two TFs were selected from each enhancer class (AFG: OTX2 and SIX1; PFG: HNF1B and HOXA2; MHG: CDX2 and HOXA9) and their motif densities measured across all three enhancer groups.

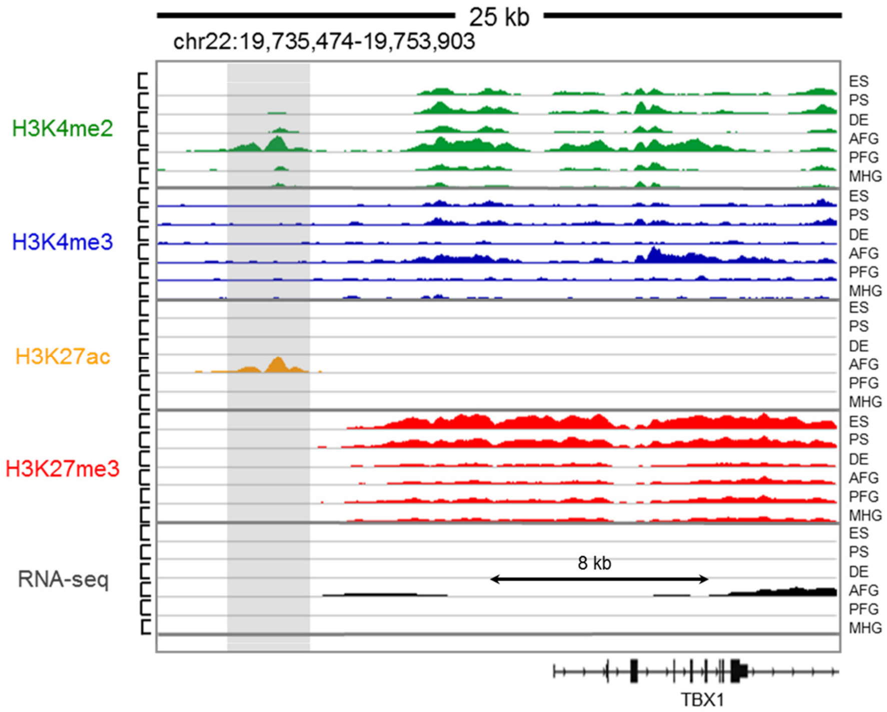
## 4.3 *In vivo* validation using a transgenic zebrafish model

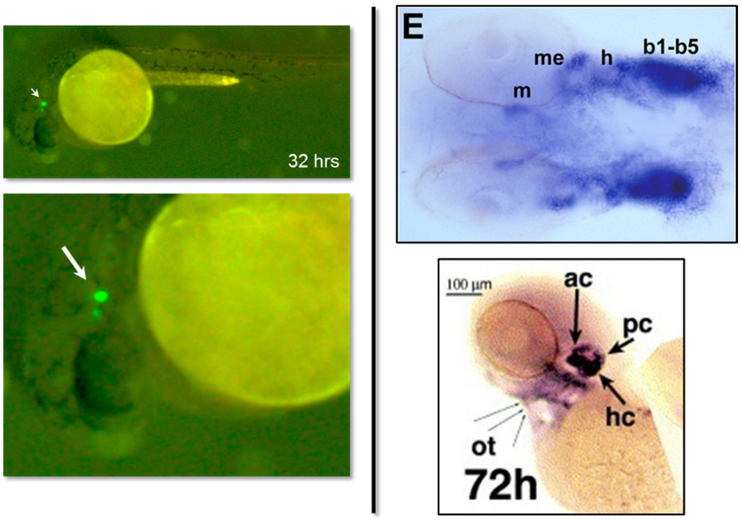
To validate the spatiotemporal activities of the predicted enhancers, we utilized a Tol2 zebrafish reporter assay system (Kawakami, 2007). This system has been demonstrated to exhibit reproducible germline transmission from $G_0$ embryos to $G_1$, with minimal $G_0$ mosaicism (Fisher et al., 2006). Since then, $G_0$ embryos have been successfully utilized for validation of human heart enhancers (Narlikar et al., 2010) and enhancers from mouse embryonic stem cells (Zhang et al., 2013b). We therefore set up a reporter vector system (Fig 4.4A) using the Tol2 ZED vector (Bessa et al., 2009) and analyzed reporter gene expression in $G_0$ zebrafish embryos. Three elements exhibiting cell-type-specific H3K27ac enrichment were tested for their ability to drive expression of a GFP reporter in zebrafish embryos. All tested loci corresponded to previously-uncharacterized genomic regions, and were proximal to *TBX1*, *HNF1B*, and *CDX2* respectively. The *TBX1*-proximal AFG element drove GFP expression in zebrafish embryos (Fig 4.4B). GFP expression broadly corresponded with anterior tbx1 expression revealed by *in situ* hybridization (Hong et al., 2008; Kochilas et al., 2003), suggesting that the tested element is a *tbx1* enhancer in anterior foregut progenitor cells. The PFG enhancer ~35 kb downstream of *HNF1B* drove GFP expression in zebrafish embryos which broadly corresponded to insulin-GFP signals (Fig 4.4C) (Song et al., 2007), suggesting H3K27ac enrichment predicted an enhancer driving *HNF1B* expression. This observation supports the hypothesis that the element is an enhancer in posterior foregut progenitors and is consistent with the role of *HNF1B* in pancreas beta cell development (De Vas et al., 2015). Finally, we tested a H3K27ac-enriched element within an intron of *WNT5B* in MHG, and observed GFP expression in zebrafish embryos which broadly overlapped with wnt5b expression (Fig 4.4D) (Freisinger et al., 2010; Kudoh et al., 2001). This observation suggest that the tested element is an enhancer driving WNT5B expression, consistent with the epigenome profile of this region suggestive of an MHG enhancer.

A)

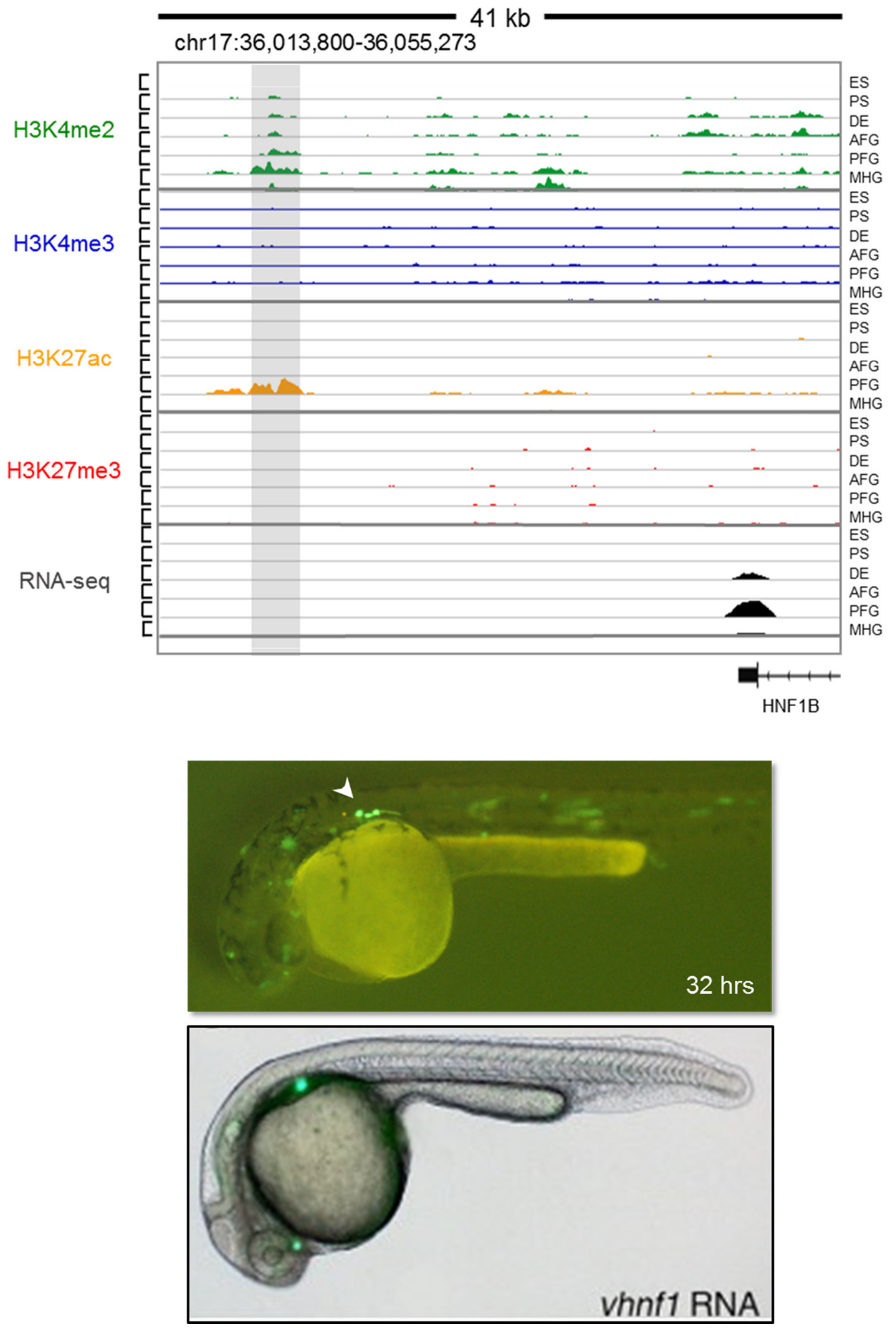

Enhancer amplification
Enhancer
TOPO cloning
attL1
attL2
Entry Vector
KanR
LR recombination
ZED Reporter Vector
Actin
DsRed2
INS
EGFP
GATA
attR1
CmR
ccdB
attR2
ori
AmpR
10,000
2000
4000
6000
8000
Microinjection

B)


25 kb
chr22:19,735,474-19,753,903
H3K4me2
H3K4me3
H3K27ac
H3K27me3
RNA-seq
ES
PS
DE
AFG
PFG
MHG
8 kb
TBX1


32 hrs
E
me
h
b1-b5
m
100 µm
ac
pc
hc
ot
72h

**C)**


41 kb
chr17:36,013,800-36,055,273
H3K4me2
H3K4me3
H3K27ac
H3K27me3
RNA-seq
ES
PS
DE
AFG
PFG
MHG
HNF1B
32 hrs
*vhnf1* RNA

D)

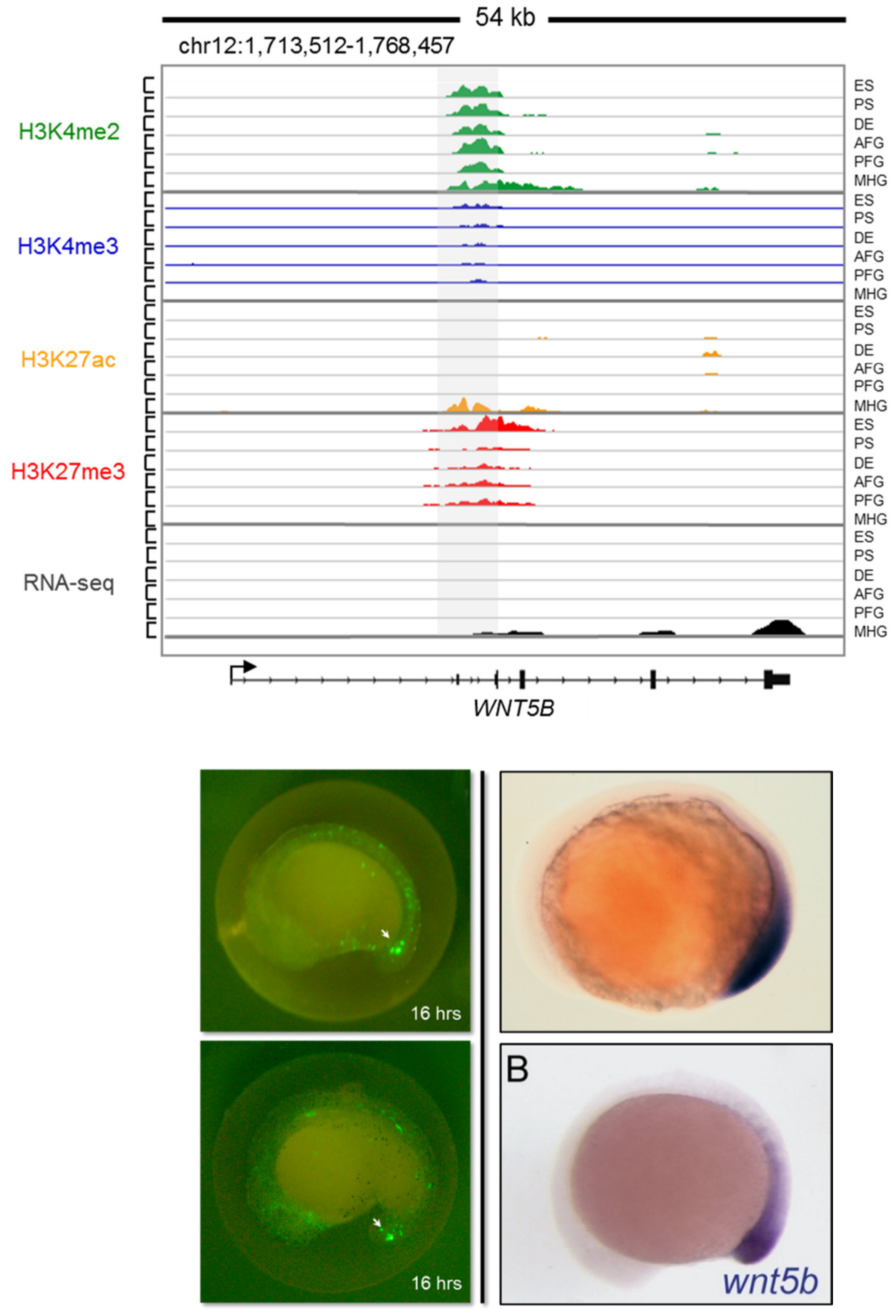

54 kb
chr12:1,713,512-1,768,457
H3K4me2
H3K4me3
H3K27ac
H3K27me3
RNA-seq
ES
PS
DE
AFG
PFG
MHG
WNT5B
16 hrs
16 hrs
B
wnt5b

**Figure 4.4 *In vivo* enhancer validation in zebrafish embryos**

(A) Schematic representation of the *in vivo* enhancer validation process. Putative enhancers were PCR amplified from genomic DNA, TOPO-cloned into a transfer vector and transferred into the ZED reporter vector (Bessa et al., 2009) through recombination. Each enhancer-ZED vector was injected into ~100 zebrafish embryos during 1-cell stage. (B) AFG enhancer activity upstream of *TBX1*, compared to *tbx1 in situ* hybridization data. Right, top: (Hong et al., 2008); Right, bottom: (Kochilas et al., 2003). (C) PFG enhancer activities downstream of *HNF1B*, compared with insulin-GFP fluorescence (Right, bottom) (Song et al., 2007). (D) Intragenic enhancer at *WNT5B* locus, compared to *wnt5b in situ* hybridization data. Right, top: (Kudoh et al., 2001); Right, bottom: (Freisinger et al., 2010).

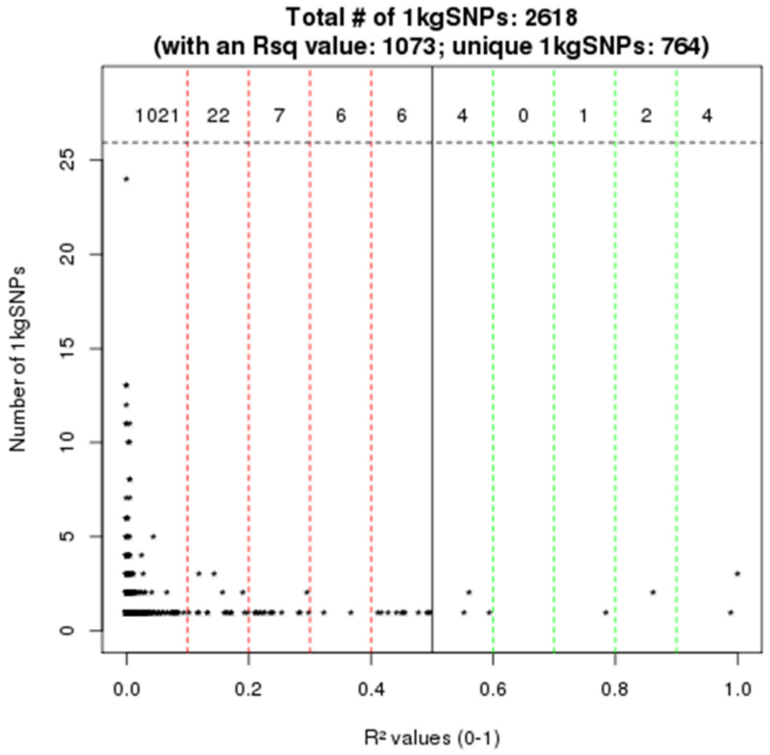
## 4.4 Identification of a putative causative variant for ulcerative colitis

GWAS studies have been conducted for diverse human traits and diseases over the past few years, and have successfully identified thousands of associated genetic variants. A major limitation for GWAS studies is the difficulty in interpreting the biological relevance of the susceptibility loci, as genotyped SNPs are designed to tag genome linkage structure and not disease risk. Identification of the causal variant(s) from associated SNPs is often not straightforward, as many risk loci fall in non-coding regions far from known genes, where annotation information remains limited (Maurano et al., 2012). Furthermore, many risk SNPs exist in strong linkage disequilibrium with other variants in proximity, further expanding the search space for the causal variant. Large-scale open chromatin mapping in multiple human cells and tissues revealed enrichment of risk variants at DNaseI-hypersensitive sites, implicating TREs in common human disease risk (Maurano et al., 2012). Our GREAT analysis of endoderm domain-specific enhancers revealed associations of these elements to various developmental disorders (Fig 4.2C), prompting us to question whether these enhancers may harbor causal variants of human complex diseases involving endoderm organs, such as the intestines.

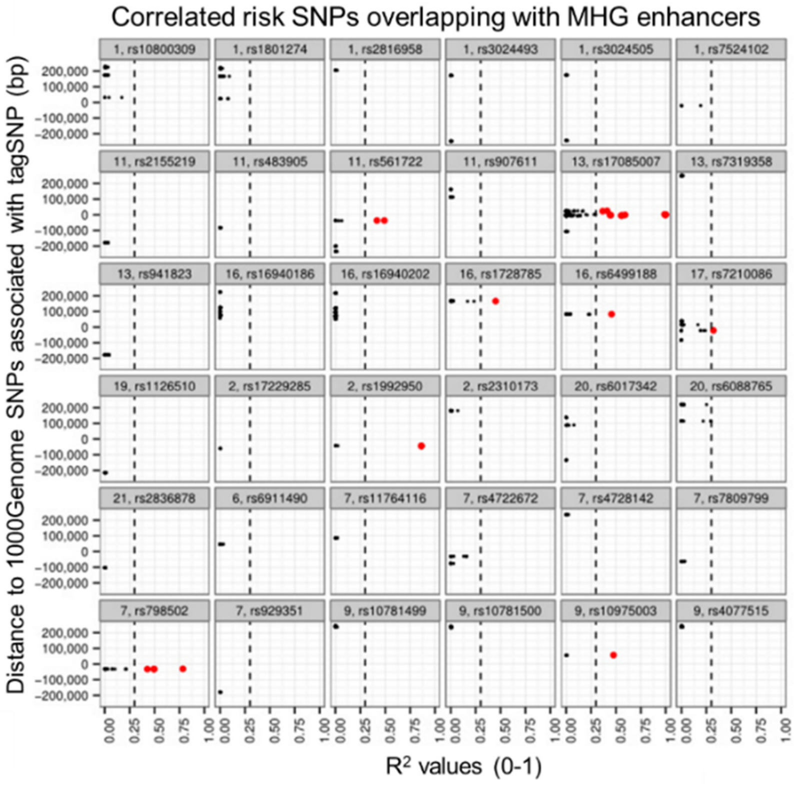
Ulcerative colitis is a complex disease characterized by inflammation and ulceration of the large intestine epithelium lining. The exact cause of this disorder is unknown, but likely involves multiple contributing factors including genetics, stress and diet. More than 100 risk loci have been reported to be associated with the disease (Anderson et al., 2011; Jostins et al., 2012), yet how these loci affect disease phenotype is largely unknown. Although treated as an autoimmune disorder, multiple studies have highlighted intestinal epithelial dysfunction as a contributing factor to disease pathogenesis (Hering et al., 2012; Laukoetter et al., 2008; McGuckin et al., 2009; Salim and Söderholm, 2011; Shen et al., 2009). Because MHG enhancers were associated with genes regulating hindgut morphogenesis and proper intestine and colon development (Fig 4.2B), we set out to test whether these enhancers may include causal variants contributing to ulcerative colitis.

To investigate this we obtained a list of 108 published ulcerative colitis-associated SNPs (tagSNP) from the NHGRI GWAS catalogue (www.genome.gov/gwastudies) and overlapped all known SNP variants from the 1000 Genomes Project within 250kb from each tagSNP with 2,620 MHG enhancers. A total of 2,618 SNPs overlapped MHG enhancers, of which 1,073 were linked to a tagSNP ($r^2 > 0$) (Fig 4.5A). We refer to these linked SNPs as "putative causal SNPs" and sorted them according to linkage strength and distance from tagSNPs. Out of all 108 tagSNPs, 36 were in linkage with at least one putative causal SNP ($r^2>0$). A further 8 out of these 36 tagSNPs were in relatively strong linkage ($r^2 > 0.3$) with at least one putative causal SNP (Fig 4.5B) (Table 5). Strikingly, tagSNP rs17085007 was linked to 12 putative causal SNPs, of which 3 were in absolute linkage ($R^2 = 1.0$) (Table 5). rs17085007 has been identified as a UC risk locus in Japanese (Asano et al., 2009), European (Jostins et al., 2012) and Korean (Yang et al., 2013) populations. It resides on chr13q12.13, a non-coding region approximately 110kb downstream of *USP12*. Like most non-coding risk variants, the biological significance of disease association underlying this SNP is unknown. Fine mapping revealed a 74kb linkage block surrounding rs17085007 (Asano et al., 2009), which overlapped an MHG enhancer cluster (Fig 4.5C). ChIP-seq data from an additional panel of ENCODE cell types failed to reveal any histone modification enrichment across this SNP locus, even though the risk locus mapped to regions of high sequence conservation (Fig 4.5D). Disruption of an/multiple MHG enhancer(s) may account for ulcerative colitis risk associated with the 13q12.13 genetic locus.

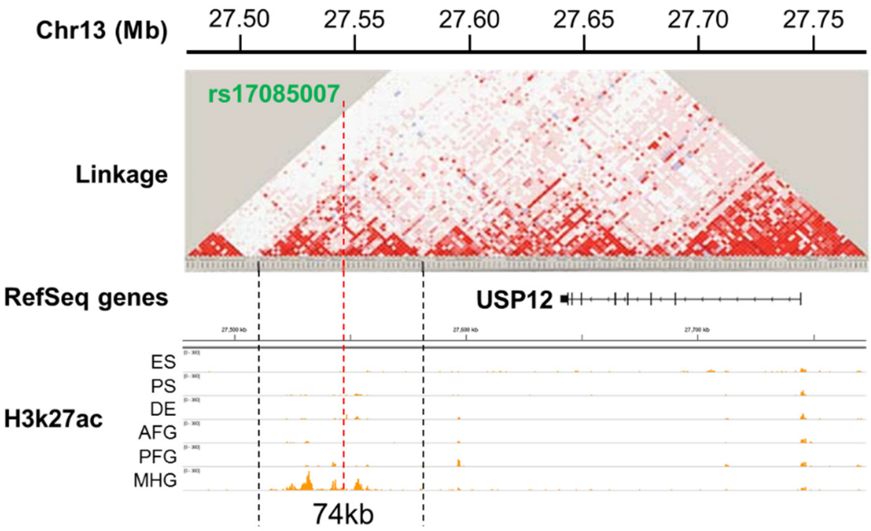
A)


Total # of 1kgSNPs: 2618
(with an Rsq value: 1073; unique 1kgSNPs: 764)
1021
22
7
6
6
4
0
1
2
4
Number of 1kgSNPs
0
5
10
15
20
25
0.0
0.2
0.4
0.6
0.8
1.0
R² values (0-1)

B)


Correlated risk SNPs overlapping with MHG enhancers
Distance to 1000Genome SNPs associated with tagSNP (bp)
200,000
100,000
0
-100,000
-200,000
1, rs10800309
1, rs1801274
1, rs2816958
1, rs3024493
1, rs3024505
1, rs7524102
11, rs2155219
11, rs483905
11, rs561722
11, rs907611
13, rs17085007
13, rs7319358
13, rs941823
16, rs16940186
16, rs16940202
16, rs1728785
16, rs6499188
17, rs7210086
19, rs1126510
2, rs17229285
2, rs1992950
2, rs2310173
20, rs6017342
20, rs6088765
21, rs2836878
6, rs6911490
7, rs11764116
7, rs4722672
7, rs4728142
7, rs7809799
7, rs798502
7, rs929351
9, rs10781499
9, rs10781500
9, rs10975003
9, rs4077515
0.00
0.25
0.50
0.75
1.00
R² values (0-1)

C)


Chr13 (Mb)
27.50
27.55
27.60
27.65
27.70
27.75
rs17085007
Linkage
RefSeq genes
USP12
27,500 kb
27,600 kb
27,700 kb
H3k27ac
ES
PS
DE
AFG
PFG
MHG
74kb

D)

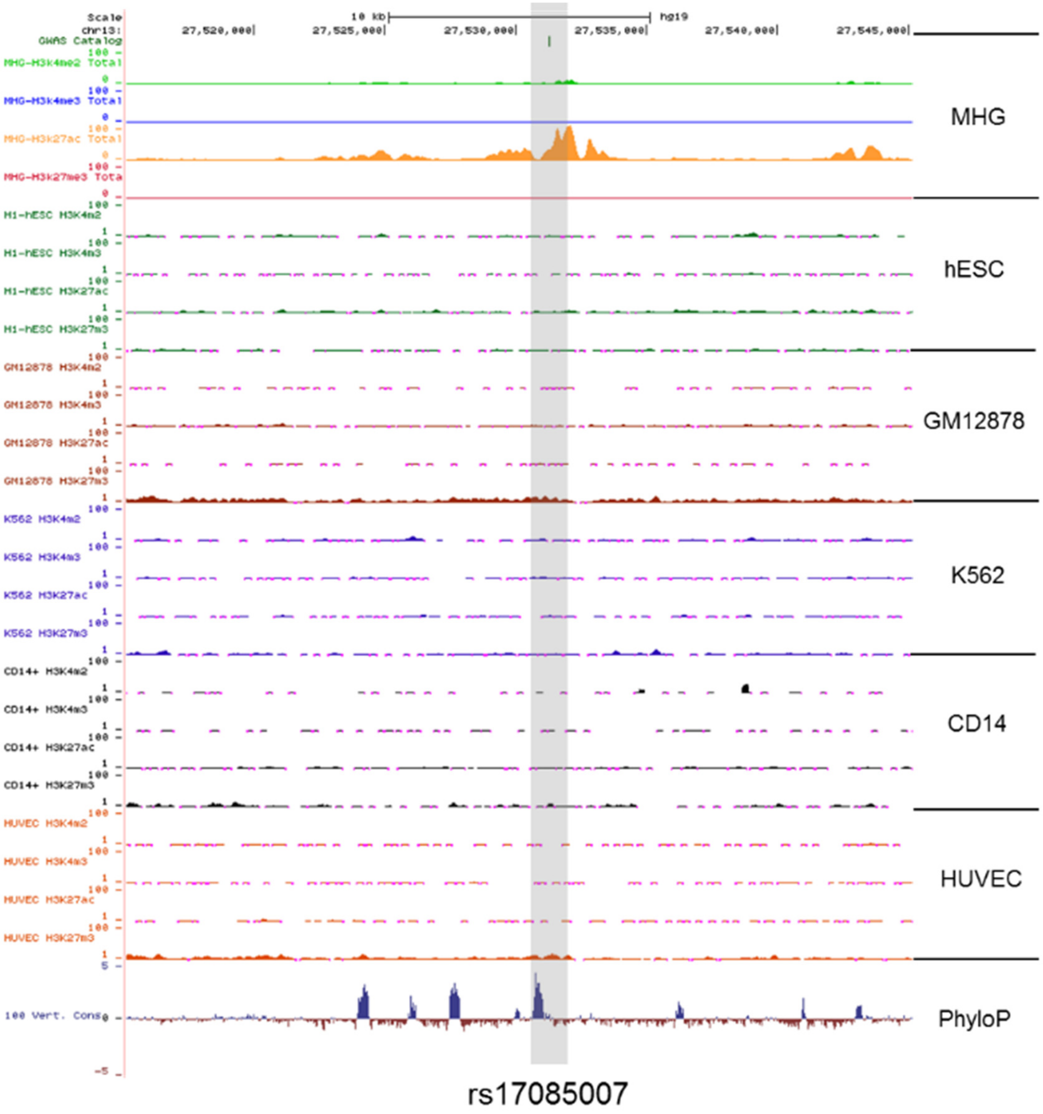


**Figure 4.5 Ulcerative colitis causal variant identification using MHG enhancers**

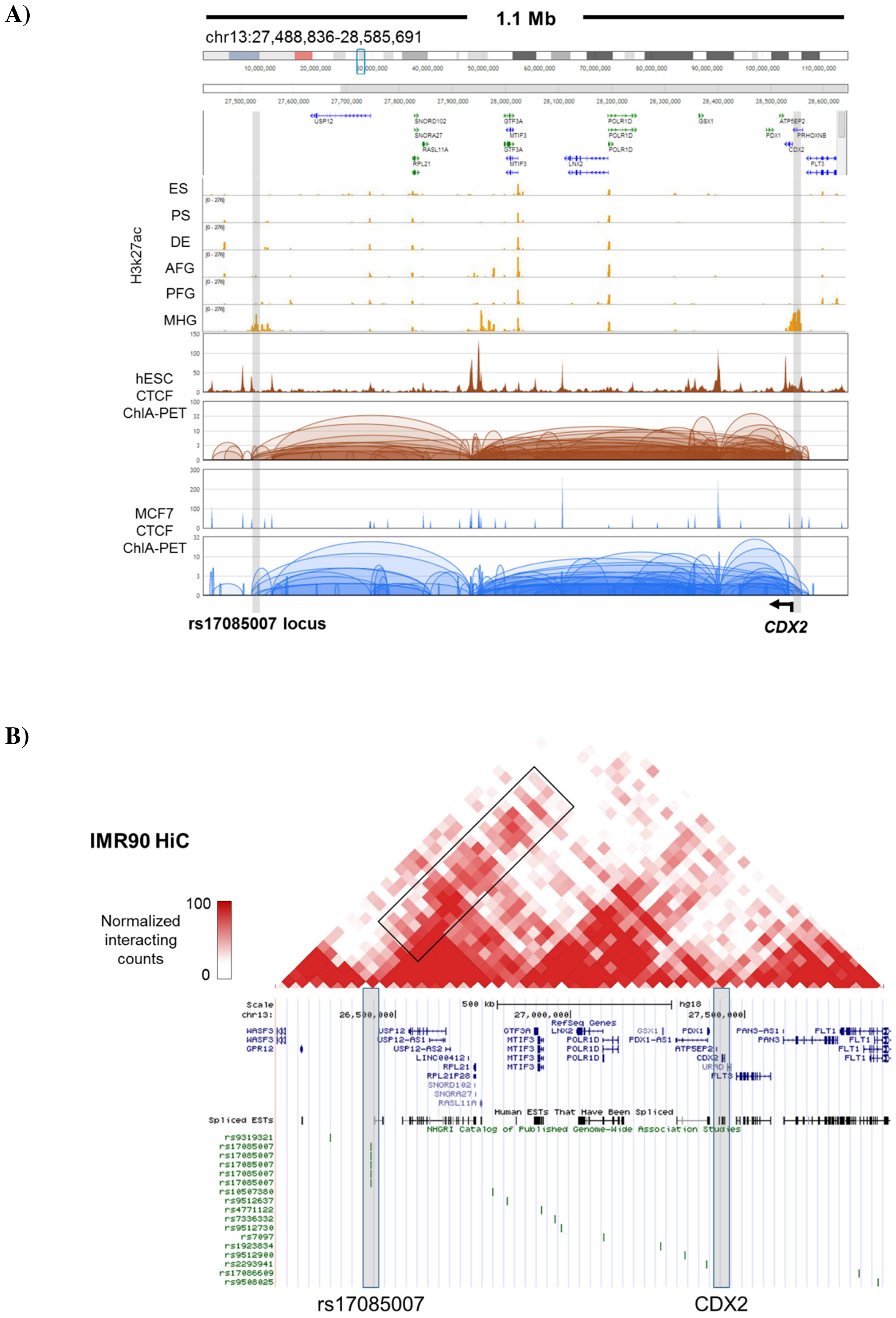
(A) Distribution of 1000 Genomes SNPs (1kgSNPs) by $R^2$ values to ulcerative colitis tagSNPs. (B) Correlated risk SNPS which overlapped MHG enhancers, refered to as putative causal variants. These SNPs are sorted according to strength of linkage and distance from tagSNPs, and red dots represent putative causal SNPs in relatively strong linkage ($R^2$ >0.3) with tagSNPs. (C) Fine mapping of a 290kb region around the ulcerative colitis-associated SNP rs17085007, adapted from (Asano et al., 2009). A 74kb linkage block encompass the SNP and several MHG enhancer elements. (D) Histone modification profiles of MHG cells and five ENCODE cell types at the rs17085007 risk SNP, together with basewise phyloP conservation score of the locus.

**Table 5 List of tagSNPs and their corresponding SNPs overlapping MHG enhancers**

| TagSNP | TagSNP coordinates | Corresponding SNP | Corresponding SNP coordinates | R.squared | p.value |
|---|---|---|---|---|---|
| rs10975003 | chr17:5213687 | rs16922908 | chr17:5270239 | 0.477855 | 9.10E-51 |
| rs17085007 | chr17:27531267 | rs1556039 | chr17:27532164 | 1 | 1.34E-131 |
| | | rs1556040 | chr17:27532117 | 1 | 1.34E-131 |
| | | rs9512464 | chr17:27530715 | 1 | 1.34E-131 |
| | | rs9553939 | chr17:27533642 | 0.990226 | 1.34E-129 |
| | | rs9512464 | chr17:27530715 | 0.593892 | 1.63E-94 |
| | | rs11618775 | chr17:27526265 | 0.562111 | 3.58E-88 |
| | | rs17587460 | chr17:27525790 | 0.562111 | 3.58E-88 |
| | | rs12856714 | chr17:27525461 | 0.554135 | 2.27E-87 |
| | | rs74529802 | chr17:27529096 | 0.449735 | 2.50E-15 |
| | | rs12585310 | chr17:27528347 | 0.442424 | 6.02E-76 |
| | | rs78362746 | chr17:27555181 | 0.412351 | 2.26E-57 |
| | | rs35669887 | chr17:27553691 | 0.367284 | 3.41E-52 |
| rs1728785 | chr17:68591230 | rs7184242 | chr17:68757572 | 0.452546 | 1.07E-69 |
| rs1992950 | chr17:200290359 | rs1348813 | chr17:200245245 | 0.863428 | 1.36E-150 |
| | | rs2305262 | chr17:200246318 | 0.863428 | 1.10E-47 |
| rs561722 | chr17:114386830 | rs581015 | chr17:114349840 | 0.492249 | 5.39E-95 |
| | | rs12797627 | chr17:114349787 | 0.418051 | 7.80E-69 |
| rs6499188 | chr17:68674788 | rs7184242 | chr17:68757572 | 0.456654 | 1.18E-70 |
| rs7210086 | chr17:70641698 | rs16977300 | chr17:70619547 | 0.322582 | 3.03E-39 |
| rs798502 | chr17:2789880 | rs798563 | chr17:2757867 | 0.785152 | 1.01E-138 |
| | | rs4719648 | chr17:2756832 | 0.497309 | 6.78E-93 |
| | | rs4719647 | chr17:2756721 | 0.494028 | 1.36E-92 |
| | | rs6965989 | chr17:2757420 | 0.427976 | 2.89E-77 |

## 4.5 Risk locus target gene identification

Local chromosome structure organized by CTCF binding has been recently shown to be important for the expression super-enhancer driven genes (Dowen et al., 2014). To investigate chromatin organization at 13q12.13, we made use of CTCF chromatin interactions detected by ChIA-PET in hESC and MCF7 cells (unpublished data from Ruan Lab). CTCF binding is associated with looping interactions to both active and inactive promoters, and has been proposed as a global organizer of long range chromatin interactions (Jin et al., 2013; Phillips and Corces, 2009). Given that CTCF binding is largely invariant between different cell types (Kim et al., 2007), we reasoned that CTCF-associated chromatin interactions can provide general insights into higher-order chromatin organization across multiple cell-types. Extensive CTCF-associated chromatin loops were observed extending over 1 Mb, linking the rs17085007 locus to *CDX2* (Fig 4.6A) in both hESC and MCF7 cells. In parallel, we examined Hi-C chromatin interactions in IMR90 fibroblast cells (Jin et al., 2013) and observed similar looping interactions with the *CDX2* locus (Fig 4.6B), suggesting that chromatin topology at this locus is broadly similar between different cell types. We validated this interaction using DNA FISH in the colon cell line HCT116 and observed significant colocalization of the two loci for both alleles (Fig 4.6C). *CDX2* is crucial for intestine specification from hindgut endoderm (Gao et al., 2009; Grainger et al., 2010) and is important for maintaining the balance of epithelium proliferation-differentiation (Lorentz et al., 1997; Suh and Traber, 1996). Importantly, CDX2 expression was markedly reduced in ulcerative colitis patients (Coskun et al., 2012). These data suggest that CDX2 is the target gene of the 13q12.13 ulcerative colitis risk locus.


A)
1.1 Mb
chr13:27,488,836-28,585,691
USP12
SNORD102
SNORA27
RASL11A
RPL21
GTF3A
MTIF3
LNX2
POLR1D
GSX1
PDX1
ATP5EP2
PRHOXNB
CDX2
FLT3
H3k27ac
ES
PS
DE
AFG
PFG
MHG
hESC
CTCF
ChIA-PET
MCF7
CTCF
ChIA-PET
rs17085007 locus
CDX2
B)
IMR90 HiC
Normalized interacting counts
100
0
Scale
chr13:
500 kb
hg18
26,500,000
27,000,000
27,500,000
RefSeq Genes
Spliced ESTs
Human ESTs That Have Been Spliced
NHGRI Catalog of Published Genome-Wide Association Studies
rs9319321
rs17085007
rs17085007
rs17085007
rs17085007
rs17085007
rs10507380
rs9512637
rs4771122
rs7336332
rs9512730
rs7097
rs1923834
rs9512900
rs2293941
rs17086609
rs9508025
rs17085007
CDX2

C)

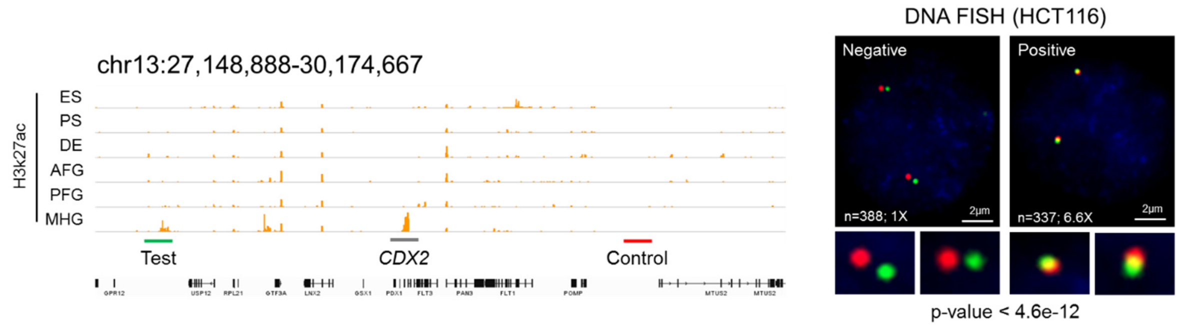


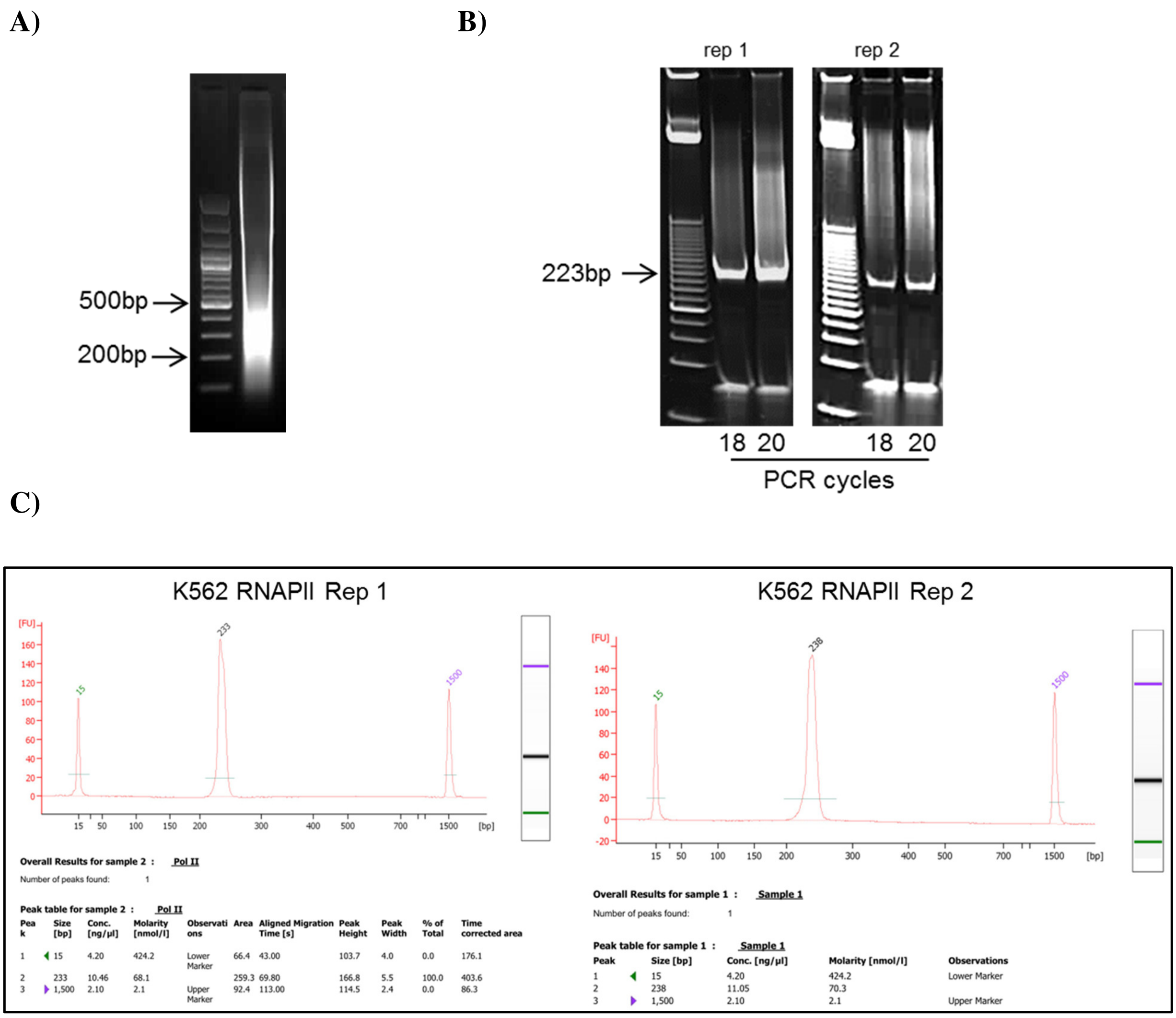
**Figure 4.6 Target gene identification for risk locus 13q12.13**

(A) H3K27ac signals at 13q12.13 in endoderm derivatives, and long range interactions with *CDX2* revealed through CTCF ChIA-PET. (B) Genome topology of the same locus based on Hi-C data. (C) DNA-FISH validation of interaction between 13q12.13 risk locus (test/positive) and CDX2 locus (bait) in HCT116 colon cell line. A control probe (negative) was selected equidistant from CDX2 as the test probe. n refers to the number of nuclei counted; 1x and 6.6x refers to the control-normalized fold changes in interacting nuclei count in the control and test experiment respectively. P values were calculated using the Fisher’s Exact Test.

# CHAPTER 5 ANALYSIS OF GLOBAL CHROMATIN INTERACTIONS IN TRANSCRIPTION CONTROL

## 5.1 RNAPII ChIA-PET library construction in K562 cells

In this study, we hypothesized that application of ChIA-PET to components of the general transcription machinery, such as RNAPII, allows investigation of global transcription-associated interactions between enhancers and promoters. To this end, we have constructed ChIA-PET libraries for RNAPII in the myelogenous leukemia cell line K562 (Zhang et al., 2012a), as part of a more comprehensive study involving five human cell lines (Li et al., 2012). Formaldehyde cross-linked chromatin was sonicated for RNAPII ChIP to a size range of 200 – 500 bp (Fig 5.1A), and 100 ng of ChIP DNA was used as input for ChIA-PET library construction. The 8WG16 RNAPII antibody, which recognizes the initiation form of RNAPII, was used for all ChIP experiments. This antibody binds the unphosphorylated consensus repeat, YSPTSPS, and allows for enrichment of both pre-initiatiation and initiating forms of RNAPII (Phatnani and Greenleaf, 2006). Hence, the 8WG16 epitope specifically enriches for promoter-bound RNAPII much more efficiently than RNAPII on coding regions or 3' ends (O'Brien et al., 1994). This allows the ChIP protocol to enrich for total promoter-associated RNAPII, which will ensure a comprehensive overview of all RNAPII- and promoter-associated chromatin interactions. Subsequent to proximity ligation and PET purification, PCR was performed for library amplification and resolved by gel electrophoresis, yielding an approximately 230 bp product (Fig 5.1B). Amplified libraries were resolved using on-chip electrophoresis to validate fragment size and concentration (Fig 5.1C) and subsequently sequenced on an Illumina GAIIx machine.



**Overall Results for sample 2 :** Pol II

Number of peaks found: 1

**Peak table for sample 2 :** Pol II

| Peak | | Size [bp] | Conc. [ng/µl] | Molarity [nmol/l] | Observations | Area | Aligned Migration Time [s] | Peak Height | Peak Width | % of Total | Time corrected area |
|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | ◀ | 15 | 4.20 | 424.2 | Lower Marker | 66.4 | 43.00 | 103.7 | 4.0 | 0.0 | 176.1 |
| 2 | | 233 | 10.46 | 68.1 | | 259.3 | 69.80 | 166.8 | 5.5 | 100.0 | 403.6 |
| 3 | ▶ | 1,500 | 2.10 | 2.1 | Upper Marker | 92.4 | 113.00 | 114.5 | 2.4 | 0.0 | 86.3 |

**Overall Results for sample 1 :** Sample 1

Number of peaks found: 1

**Peak table for sample 1 :** Sample 1

| Peak | | Size [bp] | Conc. [ng/µl] | Molarity [nmol/l] | Observations |
|---|---|---|---|---|---|
| 1 | ◀ | 15 | 4.20 | 424.2 | Lower Marker |
| 2 | | 238 | 11.05 | 70.3 | |
| 3 | ▶ | 1,500 | 2.10 | 2.1 | Upper Marker |

**Figure 5.1 RNAPII ChIA-PET library construction (Zhang et al., 2012a)**

A) Gel electrophoresis of formaldehyde-fixed, sonicated chromatin DNA from K562 cells. B) Purified ChIA-PETs were amplified using 18 and 20 PCR cycles. Amplified PETs appear as a sharp, distinct 223 bp band. The 40 bp bands represent excess unligated adapters. C) Bioanalyzer analysis of the amplified library.

## 5.2 Distinct models of RNAPII-associated chromatin interactions

K562 RNAPII ChIA-PET libraries were constructed and sequenced in replicates, generating about 14 million tags per library. Sequenced tags were mapped to the human genome for binding peak and interaction cluster calling by Dr Li Guoliang using the ChIA-PET tool (Li et al., 2010) (Table 6). Close to 27,000 RNAPII peaks and 65,000 interaction clusters were identified per library. Among the identified interactions, more than 34,000 interactions were observed in both libraries, representing approximately 53% of interactions identified in each library (Fig 5.2). We subsequently pooled all interactions from both libraries for subsequent analyses. Among the genome-wide RNAPII interactions identified, 16,597 (17.3%) represented promoter-promoter (P-P) interactions, while 27,438 (28.6%) and 39,046 (40.7%) were promoter-enhancer (P-E) and enhancer-enhancer (E-E) interactions respectively. The large number of P-P and E-E interactions suggests that a significant number of genetic loci may be coordinately transcribed. Such P-P and E-E interactions were also widely observed in MCF7 cells (Table 6). Based on the connectivity between pairs of interactions, these interactions could be combined to form complex interactions, from which three distinct interaction types were identified (Fig 5.3A). “Basal” promoters (BP) are characterized by isolated RNAPII binding; “single-gene” interactions (SG) include a single promoter interacting with one/multiple enhancers; “multi-gene” interactions (MG) consist of multiple interacting promoters and enhancers. A total of 6,223 interaction models were identified. Although there were only 1,328 (21%) MG interaction clusters, these clusters contained 11,723 genes, constituting 61% of all genes involved in the interaction models (Fig 5.3B). These data suggest that in addition to the classical ‘enhancer-promoter’ transcription model, promoters and enhancers are engaged in extensive interaction clusters for transcription.

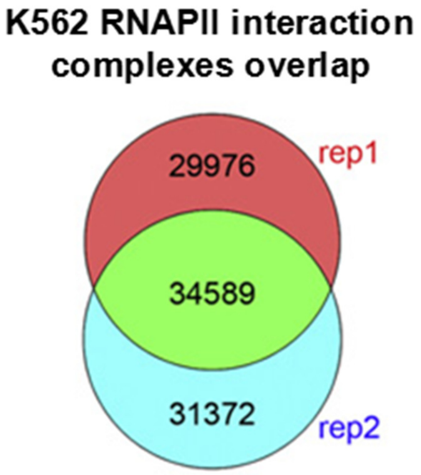


**Figure 5.2 Interaction overlap analysis between two biological replicates of RNAPII ChIA-PET libraries**

**Table 6 Statistics of sequenced tag counts, RNAPII peaks, interaction clusters and interaction types in K562 cells$^{\%}$**

| Library | Unique tags | RNAPII peaks | Interaction clusters$^{\#}$ | Interaction type* | | | |
|---|---|---|---|---|---|---|---|
| | | | | Promoter-promoter | Promoter-enhancer | Enhancer-enhancer | Promoter-terminator |
| K562-1 | 14,177,547 | 26,922 | 64,565 | 16,597<br>(17.3%) | 27,438<br>(28.6%) | 39,046<br>(40.7%) | 12,856<br>(13.4%) |
| K562-2 | 14,365,592 | 27,046 | 65,961 | | | | |
| MCF-1$^{\wedge}$ | 15,283,270 | 27,198 | 23,440 | 8,599<br>(24.8%) | 11,720<br>(33.8%) | 8,495<br>(24.5%) | 5,860<br>(16.9%)) |
| MCF-2$^{\wedge}$ | 15,622,720 | 27,683 | 24,126 | | | | |

$^{\#}$ Interaction clusters with PET counts greater than 3

* Interaction type computed using total number of interaction clusters from both replicates

$^{\wedge}$ MCF7 interaction data were generated by other members of the lab

$^{\%}$ Data analysis using ChIA-PET tool were performed by Dr Li Guoliang (Genome Institute of Singapore)

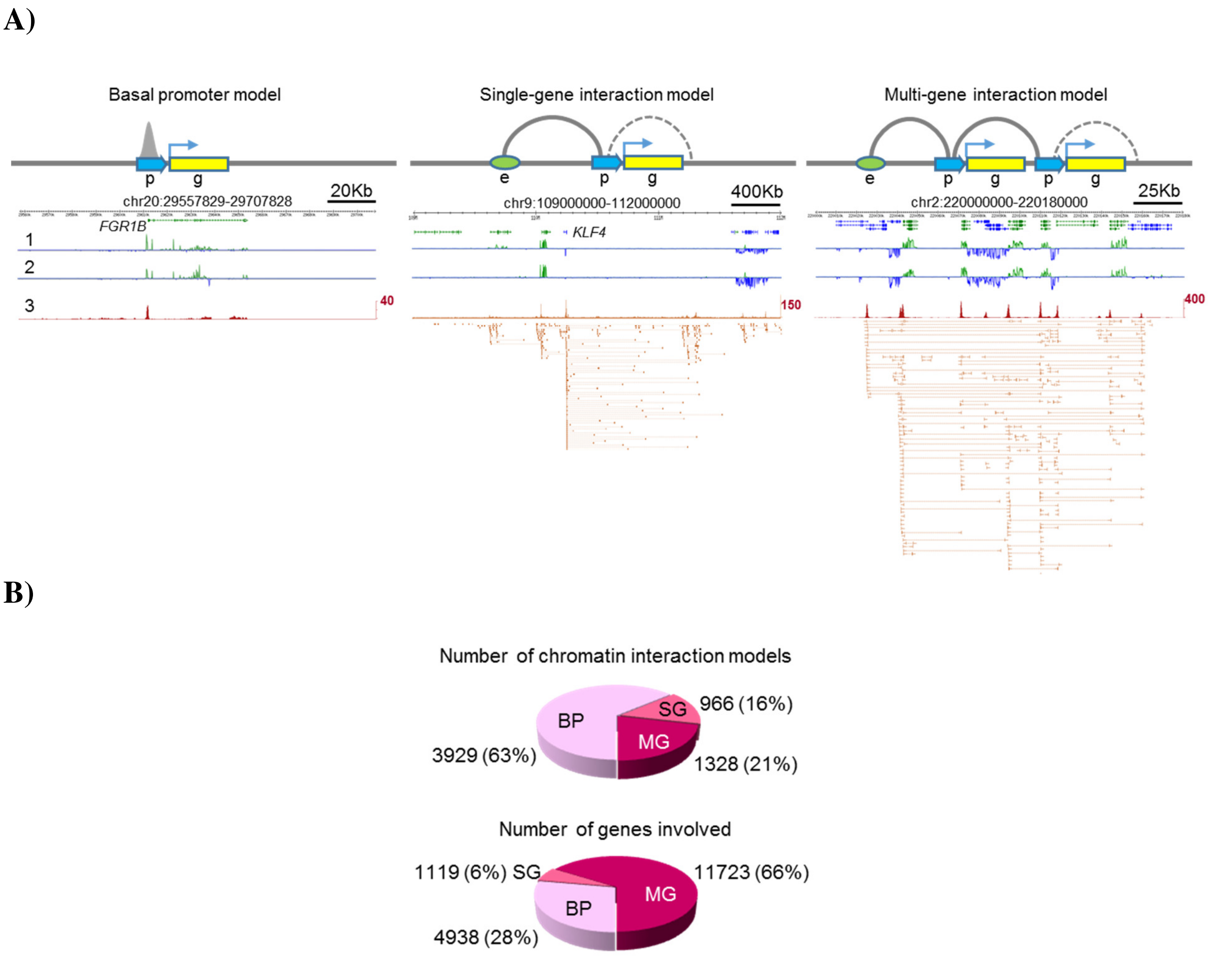


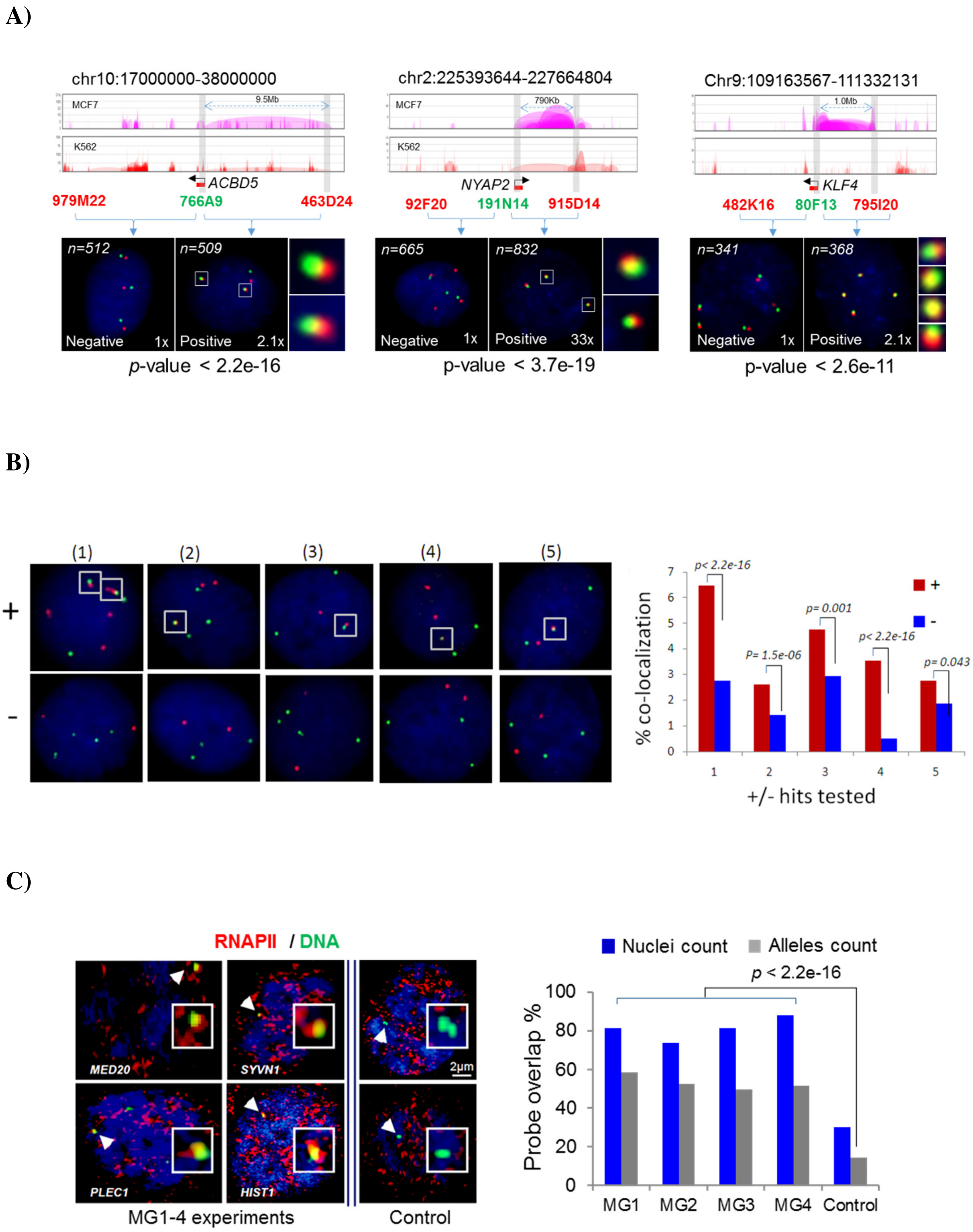
**Figure 5.3 Three models for RNAPII-associated chromatin interactions (Li et al., 2012)**

A) Basal – RNAPII peak at promoter. Single-gene – single enhancer to promoter. Multigene – several enhancers and promoters converging. B) Proportion of single-gene, basal-promoter and multi-gene interactions models, as well as the number of genes associated with each model.

### 5.3 Validation of chromatin interactions

To validate the RNAPII-associated interactions identified by ChIA-PET, DNA FISH analysis was performed on three randomly-selected intrachromosomal interactions. At each test site, a control probe was selected equidistant from the predicted interacting probe pair, and probe interactions were counted in at least 300 nuclei for each interaction. All test-probe pairs showed significantly higher levels of co-localization than the control probe (Fig 5.4A). In addition, 5 randomly selected interchromosomal interactions were similarly validated (Fig 5.4B). Compared to the control probe (-), test probes (+) targeting different chromosomes showed significantly higher co-localization levels. These data indicate that most chromatin interactions captured by RNAPII ChIA-PET were genuine.

The multi-gene interactions identified suggest that interacting genes may be coordinately transcribed in a postulated “transcription factory” (Cook, 1999). To study the link between multi-gene interacting loci and transcription factories, I performed 3D-DNA fish combined with RNAPII IF staining of MCF7 nuclei for 4 multi-gene interacting loci. All 4 tested multi-gene interaction loci (green) showed significant association with RNAPII IF staining (red) (Fig 5.4C). These data demonstrated a convergence of multi-gene interacting loci with regions of active RNAPII transcription, supporting the hypothesis that chromatin interactions provide a structural framework for coordinated gene transcription.



A)
chr10:17000000-38000000
MCF7
9.5Mb
K562
ACBD5
979M22
766A9
463D24
n=512
n=509
Negative 1x
Positive 2.1x
p-value < 2.2e-16
chr2:225393644-227664804
MCF7
790Kb
K562
NYAP2
92F20
191N14
915D14
n=665
n=832
Negative 1x
Positive 33x
p-value < 3.7e-19
Chr9:109163567-111332131
1.0Mb
KLF4
482K16
80F13
795I20
n=341
n=368
Negative 1x
Positive 2.1x
p-value < 2.6e-11
B)
(1)
(2)
(3)
(4)
(5)
+
-
% co-localization
0
1
2
3
4
5
6
7
p< 2.2e-16
P= 1.5e-06
p= 0.001
p< 2.2e-16
p= 0.043
+
-
1
2
3
4
5
+/- hits tested
C)
RNAPII / DNA
MED20
SYVN1
PLEC1
HIST1
2µm
MG1-4 experiments
Control
Nuclei count
Alleles count
p < 2.2e-16
Probe overlap %
0
20
40
60
80
100
MG1
MG2
MG3
MG4
Control

**Figure 5.4 Validation of RNAPII-associated chromatin interactions**

A) DNA-FISH validation for three single-gene interactions in MCF7. Negative denotes a control BAC probe approximately equidistant from the bait (green), while positive denotes a test BA probe. *n* denotes the total number of nuclei analyzed. Control-normalized fold changes in interacting nuclei count in the control and test experiments were denoted by "x" at the bottom right corner. P values were calculated using the Fisher's Exact Test. B) DNA FISH validation of five randomly chosen interchromosomal interactions. "+" denotes positive test probe pairs, while "-" denotes non-interacting control probe pairs. C) Colocalization analysis of multi-gene loci with RNAPII foci. RNAPII-IF staining (red) with four multi-gene loci (green, MED20, PLEC1, SYVN1, HIST1). Overlaps between RNAPII foci, multi-gene loci and control locus were visually determined for nuclei (n = 436) and alleles and shown as a barchart.

## 5.4 Enhancer-promoter interactions and disease-associated loci

The identification of thousands of enhancer-promoter interactions using model cell lines in this study suggests that the target genes of non-coding disease-causing mutations or regulatory variants may be revealed. Consistent with this hypothesis, RNAPII ChIA-PET in MCF cells revealed abundant interactions at the *SHH* locus with an enhancer ~1Mb upstream (Fig 5.5A). Importantly, mutations within this enhancer has been reported to cause the congenital disorder *preaxial polydactylyl* through disruption of *SHH* transcription (Lettice et al., 2002).

In another example, we examined the *IRS1* locus, which was demonstrated to harbor a type II diabetes (T2D) susceptibility locus (Kilpeläinen et al., 2011; Rung et al., 2009). The disease locus resides in a gene desert ~600Mb downstream of the *IRS1* promoter, yet no direct molecular link has been demonstrated between these two genetic loci. Here, we identified chromatin interactions between the IRS1 promoter and two downstream enhancers ~600 kb and 1 Mb away. Importantly, the more proximal enhancer coincides with the reported T2D risk locus (Fig 5.5B), suggesting that genetic variants at this locus may impair enhancer function and disrupt *IRS1* expression, leading to increased disease risk. We further validated the 1 Mb chromatin interaction with *IRS1* using DNA-FISH (Fig 5.5B) and showed that interactions identified within this region are genuine. Hence, chromatin interaction data may reveal the genetic targets of disease risk loci and facilitate subsequent functionalization studies.

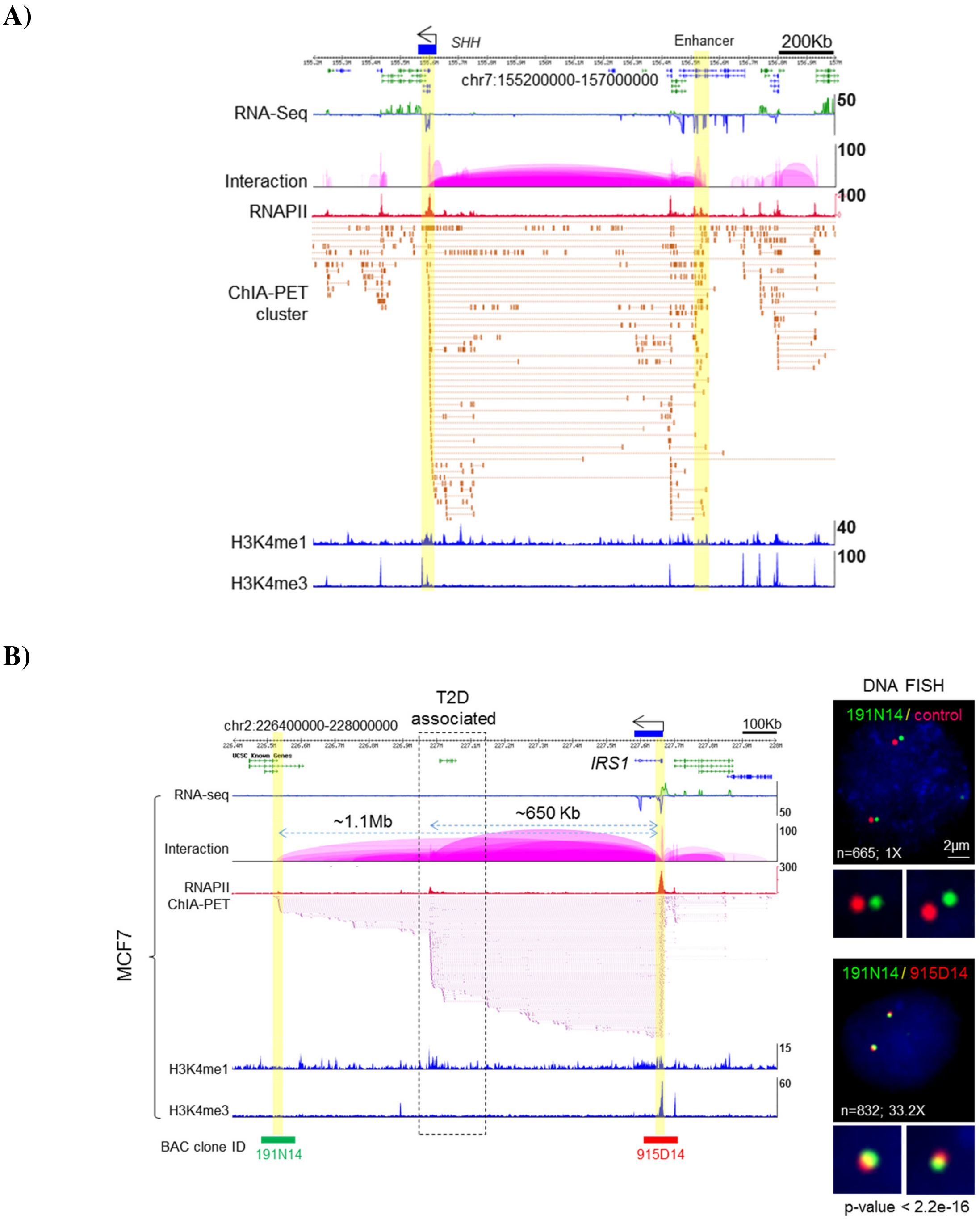


**Figure 5.5 Identification and validation of enhancer-promoter interactions involved in disease**

A) Interaction between the SHH gene and its upstream enhancer located approximately 1Mb away. B) Interaction cluster involving two putative enhancers of *IRS1*, one of which harbors an insulin resistance susceptibility locus. The 1.1Mb interaction was validated by DNA-FISH. Screenshots of genetic loci and chromatin interactions were adopted from (Li et al., 2012). Control-normalized fold changes in interacting nuclei count in the control and test experiments were denoted by “x” at the bottom right corner. P values were calculated using the Fisher’s Exact Test.

# CHAPTER 6 CONCLUDING REMARKS

## 6.1 Endoderm enhancer discovery

The precise, yet variable and complex gene expression patterns characteristic of development are largely mediated by enhancers, which function as an integrating platform mediating crosstalk between extracellular signals and encoded genetic information. The development of ChIP and high-throughput sequencing has made possible the unbiased identification of thousands of enhancers in multiple cell and tissue types (Bernstein et al., 2012; Visel et al., 2009b). In this study we have coupled an *in vitro* endoderm differentiation model with high-throughput sequencing for transcriptome and epigenome mapping, allowing comprehensive enhancer prediction and discovery in these developmentally transient cell states. Despite the physiological importance of endoderm-derived organs, including the lung, liver, pancreas and intestines, our understanding of the regulatory mechanisms driving endoderm differentiation remains rudimentary. Enhancer discovery through epigenome profiling may not only contribute to a better understanding of genome-environment crosstalk underlying distinct cellular states and behavior, but also have the potential to aid in annotation of genetic determinants underlying disease predisposition, as illustrated in this study.

Efforts to identify transcriptional enhancers in a high-throughput manner have focused on the trans-acting factor EP300 (ref), as well as the chromatin marks H3K4me1 H3K4me2 and H3K27ac. Among these modifications, H3K27ac has been identified to be strongly enriched with active promoters and enhancers in tissue culture models (Creyghton et al., 2010; Hawkins et al., 2011; Rada-Iglesias et al., 2011). Subsequent studies successfully used H3K27ac to identify enhancers during zebrafish embryogenesis (Bogdanović et al., 2012), as well as enhancers for limb specification in mouse, rhesus macaque and human (Cotney et al., 2013). Cell-type-specific enhancers can be further detected through cross-comparisons of H3K27ac in multiple cell types for increased specificity (Cotney et al., 2012). In this study, we focused on H3K27ac as cells transit from hESC to endoderm derivatives to obtain a comprehensive catalogue of putative enhancers driving dynamic and developmental-stage-specific transcriptional programs during endoderm formation. Our stringent criteria for defining putative enhancers resulted in a catalogue of strongly enriched (> 4 fold tag count over input), gene-distal H3K27ac+ elements, which are also enriched in H3K4me2 while depleted in H3K4me3 and H3K27me3 (Fig 4.1B). We note that H3K27ac alone is insufficient for conclusive enhancer discovery (Cotney et al., 2012), and that supplementation of our current datasets with endoderm-specifying TF binding profiles, such as SOX17, CXCR4 and MIXL1, will allow for more comprehensive and accurate enhancer identification. The feasibility of such approaches is illustrated by a recent study examining the binding dynamics of 38 TFs in relation to epigenome and transcriptome changes during hESC differentiation (Tsankov et al., 2015). Such enhancer catalogues not only serve as a valuable resource for understanding embryonic transcription regulation, but may also facilitate the discovery of regulatory variants underlying common diseases afflicting endoderm-derived organs, as illustrated here for rs17085007 associated with ulcerative colitis risk. The approach used in this study should be broadly applicable to other *in vitro* cell differentiation models, such as cardiomyocytes (Lian et al., 2013) and renal progenitors (Takasato et al., 2014), and is expected to reveal further insights in development and disease relevant to these cell lineages.

Our analysis of DE enhancers revealed that sequence information encoded in DE enhancers are interpreted by both the endoderm-specifying TF, EOMES, as well as the NODAL signaling effectors SMAD2/3, SMAD4 and FOXH1. In addition to endoderm (Teo et al., 2011), EOMES is also required for specification of cardiac mesoderm (Costello et al., 2011), and it is unclear how a single TF can specify progenitor cell fates in these two non-overlapping germ layers. Our integrative analysis of the binding profiles of these TFs in DE, coupled with H3K27ac as readout of functional binding, revealed a unique set of co-bound regulatory elements associated with an endoderm transcriptional program. Cell-fate decisions within the primitive streak have been suggested to be regulated by graded levels of NODAL signaling; a high level of *Nodal/Smad2/3* signaling is required for DE specification, while low levels of *Nodal* are sufficient to induce mesoderm formation (Costello et al., 2011; Dunn et al., 2004; Vincent et al., 2003). Epigenome profiling in DE revealed that strong H3K27ac enrichment is associated with high SMAD2/3/4 occupancy, whereas elements bound by EOMES with weak SMAD2/3/4 are associated with mesoderm and cardiac development (Fig 3.8). Hence, a subset of DNA-bound EOMES cooperates with NODAL signaling to induce endoderm differentiation, whereas at other loci EOMES may associate with other DNA binding partners for mesoderm specification through a SMAD2/3- and FOXH1-independent mechanism. It will be interesting to examine the relationship between NODAL effectors, EOMES and other regulatory TFs predicted to be enriched on DE enhancers, e.g. GATA4 and NANOG (Fig 3.6). This work further identified an overrepresentation of AP-1 motifs on DE enhancers, suggesting that transcriptional effectors of the JNK signaling pathway assemble on DE enhancers. Impaired JNK activity in mouse ES cells and embryoid bodies led to reduced expression of endoderm lineage markers such as *Sox17* and *Hnf1* (Loebel et al., 2011; Xu and Davis, 2010). This is consistent with the current finding that JNK signaling effectors assemble on DE enhancers and highlight the potential of JNK pathway modulation as a strategy to improve current directed-DE differentiation protocols. In sum these observations suggest that chromatin signatures in the relevant cell types can reveal the role of specific signaling pathways in developmental gene expression.

Enhancer discovery through chromatin profiling is complement by other next-generation sequencing approaches, such as DNase-seq and TF ChIP-seq. Collectively, these assays have greatly expanded our knowledge of enhancer localizations, properties and functions (Bulger and Groudine, 2010; Shlyueva et al., 2014), which in turn have spurred other novel methods for enhancer discovery. For example, publicly available enhancer catalogues have facilitated computational dissection of enhancer sequence features and highlighted the potential of *in silico* approaches for accurate enhancer prediction (Yáñez-Cuna et al., 2014). Sequencing of enhancer-associated bidirectional capped RNAs has been reported as a more sensitive and reliable predictor of enhancer activity (Andersson et al., 2014). Despite the high-throughput nature of these approaches, the predicted elements require additional validation assays as readout for enhancer function, which are laborious and low-throughput in nature. To overcome this limitation, several innovative assays were recently developed, combining enhancer discovery and validation into a single high-throughput experiment. These methods include STARR-seq (Arnold et al., 2013), SIF-seq (Dickel et al., 2014) and FIREWAch (Murtha et al., 2014). The application of these novel enhancer discovery techniques, together with targeted genome-editing techniques using site-specific nucleases (Gaj et al., 2013) are foreseen to translate genomic information into functional knowledge and ultimately, clinically-relevant endpoints.

## 6.2 ChIA-PET analysis of global chromatin interactions

Through ChIA-PET analyses of RNAPII, we have comprehensively mapped transcription-associated chromatin interactions between promoters and distal regulatory elements. Not only do promoters contact their respective enhancers, we also observed extensive enhancer-enhancer and promoter-promoter interactions, suggesting extensive looping of RNAPII binding sites into higher-order interaction complexes. The clustering of a large number of genes into higher-order structures may provide the structural basis for coordinated transcription regulation of different genes, consistent with the hypothesis that highly active transcriptional units gather at certain concentrated foci of RNAPII and offers mechanistic insight into transcription factories. This idea is further confirmed through DNA-FISH coupled with immunostaining, where we demonstrated the physical association of these interaction complexes with RNAPII foci. The clustering of multiple genes into transcription units may offer a structural explanation to the recent phenomena of pervasive transcription in intergenic regions (Jensen et al., 2013). In addition, this model is in principle akin to the bacterial operon as a mechanism for coordinated transcription regulation of related genes, further advancing the notion that similar biological concepts apply in eukaryotic and bacterial systems despite differences in complexity and structure. We note that the RNAPII interactions observed in this study are associated with pre-initiation forms of the protein, and not all interactions captured represents active transcription or elongation events. The interactions identified can be complemented with transcriptomic and epigenetic data to allow dissection of different transcriptional states. Alternatively, ChIA-PET can be performed on other phosphorylated formed of RNAPII, such as Ser5P (initiating form) or Ser2,5P (elongating form) (Phatnani and Greenleaf, 2006). These data, together with increasing sequencing depth and coverage, will allow even more comprehensive analysis of transcription-associated chromatin interactions and potentially reveal more diverse types of regulatory interaction classes.

Current 3C-based methods investigating genome organization relies on capturing crosslinked and ligated chromatin from a large population of cells for contact frequency quantification. The result is an averaged set of chromatin interactions which is assumed to represent the dominant conformation within the entire cell population. However, population-averaged assays disregard the presence of biologically relevant subpopulations and may poorly reflect the actual state in either the majority or any subpopulation of cells (Altschuler and Wu, 2010). To address this issue, Kalhor *et al*. developed a computational strategy which employed a scoring function to define the genome as a set of models each representing a spatial variant in different cells (Kalhor et al., 2011). By iteratively deriving each model from primary input data, cell-to-cell variants are independently and reproducibly reflected. The combination of this strategy with existing genome-wide methods, such as Hi-C and ChIA-PET, may overcome the limitations of chromatin interaction averaging.

Experimentally, the requirement for large cell numbers for current high-throughput chromatin interaction assays precludes their application to many interesting biological samples, such as patient DNA. This requirement may be augmented by increasing the length of current ChIA-PET tags from the current 20 bp by *Mme*I restriction digestion, to 27 bp by *EcoP*15I digestion. Longer tags map to the genome with greater accuracy, reducing the number of false positive chromatin interactions and improving the coverage of chromatin interactions at a given sequencing depth. A non-enzymatic approach for obtaining longer PETs is to sonicate the ChIA-PET ligation products to obtain DNA fragments of 400 – 600 bp. The biotinylated linker-containing DNA can then be purified using streptavidin-magnetic beads and subjected to Illumina paired end sequencing analysis at 75/105 bp from both ends. Another factor influencing cell number requirement involves the over-amplification of DNA templates derived from small number of cells, resulting in high redundancy and low interaction complexity. A potential solution involves the use a single molecule sequencing platform, such as the SMRT system by Pacific Biosciences, to overcome the requirement for library PCR amplification.

All 3C-derived methods rely on formaldehyde fixation of cells to crosslink proteins and DNA which are in physical proximity, providing a snapshot representation of chromatin structure. As formaldehyde has a relatively short crosslinking arm of 2Å, it may not, in principle, efficiently crosslink certain proteins which are more distantly bound to DNA. For example, both the histone deacetylase Rpd3 [95] and MTA3 [96], a component of the Mi-2/NuRD histone deacetylase complex, are refractory to formaldehyde crosslinking, necessitating the use of a second crosslinker with a longer spacer arm. These studies suggest that additional crosslinking using reagents of varying spacer arms are more effective in preserving large multiprotein complexes and possibly their associated chromatin interactions. Indeed, ChIA-PET experiments using ethylene glycol bis(succinimidyl succinate) (EGS) in combination with formaldehyde increases chromatin interaction complexity (unpublished observations, Ruan Lab).

The current improvements in sensitivities and scales at which interactions between genomic loci can be detected and analyzed calls for assays for functional testing of the identified interactions. The synthetic zinc-finger nucleases (ZFNs) and transcription activator-like effector nucleases (TALENs) contains independent DNA-binding and DNA-cleavage protein domains, and can produce simultaneous targeted breaks on the same chromosome at distances up to 15Mbs, producing large deletions at frequencies of 0.1-10% in a native chromatin environment in human cells (Lee et al., 2010; Urnov et al., 2010). CRISPR represents a more efficient alternative to ZFNs and TALENs, where target specificity of the Cas9 endonuclease is obtained through a guide RNA which can be designed complementary to any genomic locus (Wiedenheft et al., 2012). The versatility of the CRISPR systems is likely to make it the method of choice for targeted perturbations of chromatin interactions.

Genome-wide association studies (GWAS) have identified a large number of genetic loci associated with disease susceptibility. Most of these loci reside in noncoding regions, suggesting that mutations in enhancers may account for a large fraction of disease-associated risk (Visel et al., 2009a). While GWAS-associated loci can be overlapped with ENCODE data-defined genetic/epigenetic descriptors obtained through DNase-seq and ChIP-seq (Hazelett et al., 2014; Rhie et al., 2013), the question of how noncoding disease risk elements are connected to gene functions remains unresolved. Using RNAPII ChIA-PET analyses, we have comprehensively identified enhancers and their target genes in several human cell types. As illustrated by the *SHH* and *IRS1* loci, long range interaction data can provide the connectivity of GWAS risk loci to their target genes and facilitate functionalization of these disease risk-associated elements. As transcriptional enhancers are highly cell-type-specific, we expect that the application of ChIA-PET to disease-relevant cell types is expected to provide further mechanistic explanations for increased risk and transcription dysregulation underlying human complex diseases.

# REFERENCES

Ahlgren, U., Jonsson, J., and Edlund, H. (1996). The morphogenesis of the pancreatic mesenchyme is uncoupled from that of the pancreatic epithelium in IPF1/PDX1-deficient mice. Development *122*, 1409–1416.

Akhtar-Zaidi, B., Cowper-Sal·lari, R., Corradin, O., Saiakhova, A., Bartels, C.F., Balasubramanian, D., Myeroff, L., Lutterbaugh, J., Jarrar, A., Kalady, M.F., et al. (2012). Epigenomic Enhancer Profiling Defines a Signature of Colon Cancer. Science *336*, 736–739.

Alber, F., Dokudovskaya, S., Veenhoff, L.M., Zhang, W., Kipper, J., Devos, D., Suprapto, A., Karni-Schmidt, O., Williams, R., Chait, B.T., et al. (2007). Determining the architectures of macromolecular assemblies. Nature *450*, 683–694.

Alder, O., Cullum, R., Lee, S., Kan, A.C., Wei, W., Yi, Y., Garside, V.C., Bilenky, M., Griffith, M., Morrissy, A.S., et al. (2014). Hippo Signaling Influences HNF4A and FOXA2 Enhancer Switching during Hepatocyte Differentiation. Cell Rep. *9*, 261–271.

Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990). Basic local alignment search tool. J. Mol. Biol. *215*, 403–410.

Altschuler, S.J., and Wu, L.F. (2010). Cellular Heterogeneity: Do Differences Make a Difference? Cell *141*, 559–563.

Anders, S., and Huber, W. (2010). Differential expression analysis for sequence count data. Genome Biol. *11*, R106.

Anderson, C.A., Boucher, G., Lees, C.W., Franke, A., D'Amato, M., Taylor, K.D., Lee, J.C., Goyette, P., Imielinski, M., Latiano, A., et al. (2011). Meta-analysis identifies 29 additional ulcerative colitis risk loci, increasing the number of confirmed associations to 47. Nat Genet *43*, 246–252.

Andersson, R., Gebhard, C., Miguel-Escalada, I., Hoof, I., Bornholdt, J., Boyd, M., Chen, Y., Zhao, X., Schmidl, C., Suzuki, T., et al. (2014). An atlas of active enhancers across human cell types and tissues. Nature *507*, 455–461.

Ang, S.L., Conlon, R.A., Jin, O., and Rossant, J. (1994). Positive and negative signals from mesoderm regulate the expression of mouse Otx2 in ectoderm explants. Development *120*, 2979–2989.

Antonarakis, S.E., Irkin, S.H., Cheng, T.C., Scott, a F., Sexton, J.P., Trusko, S.P., Charache, S., and Kazazian, H.H. (1984). beta-Thalassemia in American Blacks: novel mutations in the "TATA" box and an acceptor splice site. Proc. Natl. Acad. Sci. U. S. A. *81*, 1154–1158.

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

Arnold, S.J., and Robertson, E.J. (2009). Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. Nat. Rev. Mol. Cell Biol. *10*, 91–103.

Arnold, C.D., Gerlach, D., Stelzer, C., Boryń, Ł.M., Rath, M., and Stark, A. (2013). Genome-wide quantitative enhancer activity maps identified by STARR-seq. Science *339*, 1074–1077.

Arnold, S.J., Huang, G.-J., Cheung, A.F.P., Era, T., Nishikawa, S.-I., Bikoff, E.K., Molnár, Z., Robertson, E.J., and Groszer, M. (2008a). The T-box transcription factor Eomes/Tbr2 regulates neurogenesis in the cortical subventricular zone. Genes Dev. *22*, 2479–2484.

Arnold, S.J., Hofmann, U.K., Bikoff, E.K., and Robertson, E.J. (2008b). Pivotal roles for eomesodermin during axis formation, epithelium-to-mesenchyme transition and endoderm specification in the mouse. Development *135*, 501–511.

Arnone, M.I., and Davidson, E.H. (1997). The hardwiring of development: organization and function of genomic regulatory systems. Development *124*, 1851–1864.

Arnosti, D.N., and Kulkarni, M.M. (2005). Transcriptional enhancers: Intelligent enhanceosomes or flexible billboards? J. Cell. Biochem. *94*, 890–898.

Arnosti, D.N., Barolo, S., Levine, M., and Small, S. (1996). The eve stripe 2 enhancer employs multiple modes of transcriptional synergy. Development *122*, 205–214.

Asano, K., Matsushita, T., Umeno, J., Hosono, N., Takahashi, A., Kawaguchi, T., Matsumoto, T., Matsui, T., Kakuta, Y., Kinouchi, Y., et al. (2009). A genome-wide association study identifies three new susceptibility loci for ulcerative colitis in the Japanese population. Nat. Genet. *41*, 1325–1329.

Banerji, J., Rusconi, S., and Schaffner, W. (1981). Expression of a beta-globin gene is enhanced by remote SV40 DNA sequences. Cell *27*, 299–308.

Banerji, J., Olson, L., and Schaffner, W. (1983). A lymphocyte-specific cellular enhancer is located downstream of the joining region in immunoglobulin heavy chain genes. Cell *33*, 729–740.

Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Schones, D.E., Wang, Z., Wei, G., Chepelev, I., and Zhao, K. (2007). High-resolution profiling of histone methylations in the human genome. Cell *129*, 823–837.

Baù, D., Sanyal, A., Lajoie, B.R., Capriotti, E., Byron, M., Lawrence, J.B., Dekker, J., and Marti-Renom, M.A. (2011). The three-dimensional folding of the α-globin gene domain reveals formation of chromatin globules. Nat. Struct. Mol. Biol. *18*, 107–114.

Bauer, D.E., Kamran, S.C., Lessard, S., Xu, J., Fujiwara, Y., Lin, C., Shao, Z., Canver, M.C., Smith, E.C., Pinello, L., et al. (2013). An erythroid enhancer of BCL11A subject to genetic variation determines fetal hemoglobin level. Science *342*, 253–257.

Bell, a C., and Felsenfeld, G. (2000). Methylation of a CTCF-dependent boundary controls imprinted expression of the Igf2 gene. Nature *405*, 482–485.

Bell, a C., West, a G., and Felsenfeld, G. (1999). The protein CTCF is required for the enhancer blocking activity of vertebrate insulators. Cell *98*, 387–396.

Ben-Haim, N., Lu, C., Guzman-Ayala, M., Pescatore, L., Mesnard, D., Bischofberger, M., Naef, F., Robertson, E.J., and Constam, D.B. (2006). The Nodal Precursor Acting via Activin Receptors Induces Mesoderm by Maintaining a Source of Its Convertases and BMP4. Dev. Cell *11*, 313–323.

Vanden Berghe, W., De Bosscher, K., Boone, E., Plaisance, S., and Haegeman, G. (1999). The nuclear factor-kappaB engages CBP/p300 and histone acetyltransferase activity for transcriptional activation of the interleukin-6 gene promoter. J Biol Chem *274*, 32091–32098.

Berman, B.P., Nibu, Y., Pfeiffer, B.D., Tomancak, P., Celniker, S.E., Levine, M., Rubin, G.M., and Eisen, M.B. (2002). Exploiting transcription factor binding site clustering to identify cis-regulatory modules involved in pattern formation in the Drosophila genome. Proc. Natl. Acad. Sci. U. S. A. *99*, 757–762.

Bernardo, A.S., Faial, T., Gardner, L., Niakan, K.K., Ortmann, D., Senner, C.E., Callery, E.M., Trotter, M.W., Hemberger, M., Smith, J.C., et al. (2011). Article BRACHYURY and CDX2 Mediate BMP-Induced Differentiation of Human and Mouse Pluripotent Stem Cells into Embryonic and Extraembryonic Lineages. Stem Cell *9*, 144–155.

Bernstein, B.E., Mikkelsen, T.S., Xie, X., Kamal, M., Huebert, D.J., Cuff, J., Fry, B., Meissner, A., Wernig, M., Plath, K., et al. (2006). A bivalent chromatin structure marks key developmental genes in embryonic stem cells. Cell *125*, 315–326.

Bernstein, B.E., Birney, E., Dunham, I., Green, E.D., Gunter, C., and Snyder, M. (2012). An integrated encyclopedia of DNA elements in the human genome. Nature *489*, 57–74.

Bessa, J., Tena, J.J., de la Calle-Mustienes, E., Fernández-Miñán, A., Naranjo, S., Fernández, A., Montoliu, L., Akalin, A., Lenhard, B., Casares, F., et al. (2009). Zebrafish enhancer detection (ZED) vector: a new tool to facilitate transgenesis and the functional analysis of cis-regulatory regions in zebrafish. Dev. Dyn. *238*, 2409–2417.

Bhattacharya, D., Talwar, S., Mazumder, A., and Shivashankar, G. V (2009). Spatio-Temporal Plasticity in Chromatin Organization in Mouse Cell Differentiation and during Drosophila Embryogenesis. Biophysj *96*, 3832–3839.

Bird, A. (2002). DNA methylation patterns and epigenetic memory. Genes Dev. *16*, 6–21.

Bird, A.P. (1987). CpG islands as gene markers in the vertebrate nucleus. Trends Genet. *3*, 342–347.

Bird, A.P., and Southern, E.M. (1978). Use of restriction enzymes to study eukaryotic DNA methylation: I. The methylation pattern in ribosomal DNA from Xenopus laevis. J. Mol. Biol. *118*, 27–47.

Blow, M.J., McCulley, D.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2010). ChIP-Seq identification of weakly conserved heart enhancers. Nat. Genet. *42*, 806–810.

Bogdanović, O., Fernandez-Miñán, A., Tena, J.J., De La Calle-Mustienes, E., Hidalgo, C., Van Kruysbergen, I., Van Heeringen, S.J., Veenstra, G.J.C., and Gómez-Skarmeta, J.L. (2012). Dynamics of enhancer chromatin signatures mark the transition from pluripotency to cell specification during embryogenesis. Genome Res. *22*, 2043–2053.

Bonn, S., Zinzen, R.P., Girardot, C., Gustafson, E.H., Perez-Gonzalez, A., Delhomme, N., Ghavi-Helm, Y., Wilczyński, B., Riddell, A., and Furlong, E.E.M. (2012). Tissue-specific analysis of chromatin state identifies temporal signatures of enhancer activity during embryonic development. Nat. Genet. *44*, 148–156.

Bossard, P., and Zaret, K.S. (1998). GATA transcription factors as potentiators of gut endoderm differentiation. Development *125*, 4909–4917.

Boyle, A.P., Davis, S., Shulha, H.P., Meltzer, P., Margulies, E.H., Weng, Z., Furey, T.S., and Crawford, G.E. (2008). High-Resolution Mapping and Characterization of Open Chromatin across the Genome. Cell *132*, 311–322.

Bricaud, O., and Collazo, A. (2006). The transcription factor six1 inhibits neuronal and promotes hair cell fate in the developing zebrafish (Danio rerio) inner ear. J. Neurosci. *26*, 10438–10451.

Brown, J.L., Snir, M., Noushmehr, H., Kirby, M., Hong, S.-K., Elkahloun, A.G., and Feldman, B. (2008). Transcriptional profiling of endogenous germ layer precursor cells identifies dusp4 as an essential gene in zebrafish endoderm specification. Proc. Natl. Acad. Sci. U. S. A. *105*, 12337–12342.

Buecker, C., and Wysocka, J. (2012). Enhancers as information integration hubs in development: lessons from genomics. Trends Genet. *28*, 276–284.

Buenrostro, J.D., Giresi, P.G., Zaba, L.C., Chang, H.Y., and Greenleaf, W.J. (2013). Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. Nat. Methods *10*, 1213–1218.

Bulger, M., and Groudine, M. (2010). Enhancers: the abundance and function of regulatory sequences beyond promoters. Dev. Biol. *339*, 250–257.

Bulger, M., and Groudine, M. (2011). Functional and mechanistic diversity of distal transcription enhancers. Cell *144*, 327–339.

Cai, S., Han, H.-J., and Kohwi-Shigematsu, T. (2003). Tissue-specific nuclear architecture and gene expression regulated by SATB1. Nat. Genet. *34*, 42–51.

Calo, E., and Wysocka, J. (2013). Modification of enhancer chromatin: what, how, and why? Mol. Cell *49*, 825–837.

Carter, D., Chakalova, L., Osborne, C.S., Dai, Y., and Fraser, P. (2002). Long-range chromatin regulatory interactions in vivo. Nat. Genet. *32*, 623–626.

Chapuy, B., McKeown, M.R., Lin, C.Y., Monti, S., Roemer, M.G., Qi, J., Rahl, P.B., Sun, H.H., Yeda, K.T., Doench, J.G., et al. (2013). Discovery and characterization of super-enhancer-associated dependencies in diffuse large B cell lymphoma. Cancer Cell *24*, 777–790.

Chen, X. (2008). Integration of external signaling pathways with the core transcriptional network in embryonic stem cells. Cell *133*, 1106–1117.

Chen, X., Xu, H., Yuan, P., Fang, F., Huss, M., Vega, V.B., Wong, E., Orlov, Y.L., Zhang, W., Jiang, J., et al. (2008). Resource Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. 1106–1117.

Cheng, X., Ying, L., Lu, L., Galvão, A.M., Mills, J.A., Lin, H.C., Kotton, D.N., Shen, S.S., Nostro, M.C., Choi, J.K., et al. (2012). Self-renewing endodermal progenitor lines generated from human pluripotent stem cells. Cell Stem Cell *10*, 371–384.

Chepelev, I., Wei, G., Wangsa, D., Tang, Q., and Zhao, K. (2012). Characterization of genome-wide enhancer-promoter interactions reveals co-expression of interacting genes and modes of higher order chromatin organization. Cell Res. *22*, 490–503.

Chiarle, R., Zhang, Y., Frock, R.L., Lewis, S.M., Molinie, B., Ho, Y.J., Myers, D.R., Choi, V.W., Compagno, M., Malkin, D.J., et al. (2011). Genome-wide translocation sequencing reveals mechanisms of chromosome breaks and rearrangements in B cells. Cell *147*, 107–119.

Chung, J.H., Whiteley, M., and Felsenfeld, G. (1993). A 5’ element of the chicken beta-globin domain serves as an insulator in human erythroid cells and protects against position effect in Drosophila. Cell *74*, 505–514.

Cirillo, L.A., Lin, F.R., Cuesta, I., Friedman, D., Jarnik, M., and Zaret, K.S. (2002). Opening of compacted chromatin by early developmental transcription factors HNF3 (FoxA) and GATA-4. Mol. Cell *9*, 279–289.

Clark, S.J., Harrison, J., Paul, C.L., and Frommer, M. (1994). High sensitivity mapping of methylated cytosines. Nucleic Acids Res. *22*, 2990–2997.

Coetzee, S.G., Rhie, S.K., Berman, B.P., Coetzee, G.A., and Noushmehr, H. (2012). FunciSNP: an R/bioconductor tool integrating functional non-coding data sets with genetic association studies to identify candidate regulatory SNPs. Nucleic Acids Res. *40* , e139–e139.

Conaway, R.C., and Conaway, J.W. (1993). General Initiation Factors for RNA Polymerase II. Annu. Rev. Biochem. *62*, 161–190.

Consortium, E.P. (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799–816.

Consortium, I.H.G.. (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931–945.

Cook, P.R. (1999). The Organization of Replication and Transcription. Science *284*, 1790–1795.

Coskun, M., Olsen, A.K., Holm, T.L., Kvist, P.H., Nielsen, O.H., Riis, L.B., Olsen, J., and Troelsen, J.T. (2012). TNF-α-induced down-regulation of CDX2 suppresses MEP1A expression in colitis. Biochim. Biophys. Acta - Mol. Basis Dis. *1822*, 843–851.

Costello, I., Pimeisl, I.-M., Dräger, S., Bikoff, E.K., Robertson, E.J., and Arnold, S.J. (2011). The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation. Nat. Cell Biol. *13*, 1084–1091.

Cotney, J., Leng, J., Oh, S., DeMare, L.E., Reilly, S.K., Gerstein, M.B., and Noonan, J.P. (2012). Chromatin state signatures associated with tissue-specific gene expression and enhancer activity in the embryonic limb. Genome Res. *22*, 1069–1080.

Cotney, J., Leng, J., Yin, J., Reilly, S.K., DeMare, L.E., Emera, D., Ayoub, A.E., Rakic, P., and Noonan, J.P. (2013). The Evolution of Lineage-Specific Regulatory Activities in the Human Embryonic Limb. Cell *154*, 185–196.

Creyghton, M.P., Cheng, A.W., Welstead, G.G., Kooistra, T., Carey, B.W., Steine, E.J., Hanna, J., Lodato, M. a, Frampton, G.M., Sharp, P. a, et al. (2010). Histone H3K27ac separates active from poised enhancers and predicts developmental state. Proc. Natl. Acad. Sci. U. S. A. *107*, 21931–21936.

Cross, S.H., Charlton, J.A., Nan, X., and Bird, A.P. (1994). Purification of CpG islands using a methylated DNA binding column. Nat. Genet. *6*, 236–244.

Cuddapah, S., Jothi, R., Schones, D.E., Roh, T.Y., Cui, K., and Zhao, K. (2009). Global analysis of the insulator binding protein CTCF in chromatin barrier regions reveals demarcation of active and repressive domains. Genome Res. *19*, 24–32.

Cullen, K.E., Kladde, M.P., and Seyfred, M.A. (1993). Interaction between transcription regulatory regions of prolactin chromatin. Science *261*, 203–206.

D'Amour, K.A., Agulnick, A.D., Eliazer, S., Kelly, O.G., Kroon, E., and Baetge, E.E. (2005). Efficient differentiation of human embryonic stem cells to definitive endoderm. Nat. Biotechnol. *23*, 1534–1541.

Das, P.M., and Singal, R. (2004). DNA methylation and cancer. J. Clin. Oncol. *22*, 4632–4642.

Dekker, J., Rippe, K., Dekker, M., and Kleckner, N. (2002). Capturing chromosome conformation. Science *295*, 1306–1311.

Dessimoz, J., Opoka, R., Kordich, J.J., Grapin-Botton, A., and Wells, J.M. (2006). FGF signaling is necessary for establishing gut tube domains along the anterior-posterior axis in vivo. Mech. Dev. *123*, 42–55.

Dhar, S.S., Ongwijitwat, S., and Wong-Riley, M.T.T. (2009). Chromosome conformation capture of all 13 genomic loci in the transcriptional regulation of the multisubunit bigenomic cytochrome c oxidase in neurons. J. Biol. Chem. *284*, 18644–18650.

Dickel, D.E., Zhu, Y., Nord, A.S., Wylie, J.N., Akiyama, J. a, Afzal, V., Plajzer-Frick, I., Kirkpatrick, A., Göttgens, B., Bruneau, B.G., et al. (2014). Function-based identification of mammalian enhancers using site-specific integration. Nat. Methods *11*, 566–571.

Dixon, J.R., Selvaraj, S., Yue, F., Kim, A., Li, Y., Shen, Y., Hu, M., Liu, J.S., and Ren, B. (2012). Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature *485*, 376–380.

Dong, J.M., and Lim, L. (1996). The human neuronal alpha 1-chimaerin gene contains a position-dependent negative regulatory element in the first exon. Neurochem. Res. *21*, 1023–1030.

Dostie, J., Richmond, T.A., Arnaout, R.A., Selzer, R.R., Lee, W.L., Honan, T.A., Rubio, E.D., Krumm, A., Lamb, J., Nusbaum, C., et al. (2006). Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements. Genome Res. *16*, 1299–1309.

Dovey, O.M., Foster, C.T., and Cowley, S.M. (2010). Emphasizing the positive: A role for histone deacetylases in transcriptional activation. Cell Cycle *9*, 2700–2701.

Dowen, J.M., Fan, Z.P., Hnisz, D., Ren, G., Abraham, B.J., Zhang, L.N., Weintraub, A.S., Schuijers, J., Lee, T.I., Zhao, K., et al. (2014). Control of Cell Identity Genes Occurs in Insulated Neighborhoods in Mammalian Chromosomes. Cell *159*, 374–387.

Driscoll, M.C., Dobkin, C.S., and Alter, B.P. (1989). Gamma delta beta-thalassemia due to a de novo mutation deleting the 5’ beta-globin gene activation-region hypersensitive sites. Proc. Natl. Acad. Sci. U. S. A. *86*, 7470–7474.

Duan, Z., Andronescu, M., Schutz, K., McIlwain, S., Kim, Y.J., Lee, C., Shendure, J., Fields, S., Blau, C.A., and Noble, W.S. (2010). A three-dimensional model of the yeast genome. Nature *465*, 363–367.

Dufort, D., Schwartz, L., Harpal, K., and Rossant, J. (1998). The transcription factor HNF3beta is required in visceral endoderm for normal primitive streak morphogenesis. Development *125*, 3015–3025.

Dunn, N.R., Vincent, S.D., Oxburgh, L., Robertson, E.J., and Bikoff, E.K. (2004). Combinatorial activities of Smad2 and Smad3 regulate mesoderm formation and patterning in the mouse embryo. Development *131*, 1717–1728.

Egelhofer, T. a, Minoda, A., Klugman, S., Lee, K., Kolasinska-Zwierz, P., Alekseyenko, A. a, Cheung, M.-S., Day, D.S., Gadel, S., Gorchakov, A. a, et al. (2011). An assessment of histone-modification antibody quality. Nat. Struct. Mol. Biol. *18*, 91–93.

Ehrlich, M., A-gama-sosa, M., Huang, L., Midgett, R.M., Kenneth, C., Mccune, R.A., and Gehrke, C. (1982). Amount and distribution of 5-methylcytosine in human DNA from different types of tissues or cells. Nucleic Acids Res. *10*, 2709–2722.

Eldholm, V., Haugen, A., and Zienolddiny, S. (2014). CTCF mediates the TERT enhancer-promoter interactions in lung cancer cells: Identification of a novel enhancer region involved in the regulation of TERT gene. Int. J. Cancer *134*, 2305–2313.

El-hashash, A.H.K., Al, D., Turcatel, G., Rogers, O., Li, X., Bellusci, S., and Warburton, D. (2011). Six1 transcription factor is critical for coordination of epithelial , mesenchymal and vascular morphogenesis in the mammalian lung. Dev. Biol. *353*, 242–258.

Engel, J.D., and Tanimoto, K. (2000). Looping, linking, and chromatin activity: new insights into beta-globin locus regulation. Cell *100*, 499–502.

Ernst, J., Kheradpour, P., Mikkelsen, T.S., Shoresh, N., Ward, L.D., Epstein, C.B., Zhang, X., Wang, L., Issner, R., Coyne, M., et al. (2011). Mapping and analysis of chromatin state dynamics in nine human cell types. Nature *473*, 43–49.

Van Es, J.H., Haegebarth, A., Kujala, P., Itzkovitz, S., Koo, B.-K., Boj, S.F., Korving, J., van den Born, M., van Oudenaarden, A., Robine, S., et al. (2012). A Critical Role for the Wnt Effector Tcf4 in Adult Intestinal Homeostatic Self-Renewal. Mol. Cell. Biol. *32*, 1918–1927.

Fan, X., Hagos, E.G., Xu, B., Sias, C., Kawakami, K., Burdine, R.D., and Dougan, S.T. (2007). Nodal signals mediate interactions between the extra-embryonic and embryonic tissues in zebrafish. Dev. Biol. *310*, 363–378.

Farh, K.K.-H., Marson, A., Zhu, J., Kleinewietfeld, M., Housley, W.J., Beik, S., Shoresh, N., Whitton, H., Ryan, R.J.H., Shishkin, A.A., et al. (2015). Genetic and epigenetic fine mapping of causal autoimmune disease variants. Nature *518*, 337–343.

Feldmann, A., Ivanek, R., Murr, R., Gaidatzis, D., Burger, L., and Schübeler, D. (2013). Transcription factor occupancy can mediate active turnover of DNA methylation at regulatory regions. PLoS Genet. *9*, e1003994.

Felsenfeld, G., and Groudine, M. (2003). Controlling the double helix. Nature *421*, 448–453.

Fisher, S., Grice, E. a, Vinton, R.M., Bessling, S.L., and McCallion, A.S. (2006). Conservation of RET regulatory function from human to zebrafish without sequence similarity. Science *312*, 276–279.

Fisher, W.W., Li, J.J., Hammonds, A.S., Brown, J.B., Pfeiffer, B.D., Weiszmann, R., MacArthur, S., Thomas, S., Stamatoyannopoulos, J. a, Eisen, M.B., et al. (2012). DNA regions bound at low occupancy by transcription factors do not drive patterned reporter gene expression in Drosophila. Proc. Natl. Acad. Sci. U. S. A. *109*, 21330–21335.

Fraser, J., Rousseau, M., Shenker, S., Ferraiuolo, M.A., Hayashizaki, Y., Blanchette, M., and Dostie, J. (2009). Chromatin conformation signatures of cellular differentiation. Genome Biol. *10*, R37.

Freisinger, C.M., Fisher, R. a, and Slusarski, D.C. (2010). Regulator of g protein signaling 3 modulates wnt5b calcium dynamics and somite patterning. PLoS Genet. *6*, e1001020.

French, J.D., Ghoussaini, M., Edwards, S.L., Meyer, K.B., Michailidou, K., Ahmed, S., Khan, S., Maranian, M.J., O’Reilly, M., Hillman, K.M., et al. (2013). Functional variants at the 11q13 risk locus for breast cancer regulate cyclin D1 expression through long-range enhancers. Am. J. Hum. Genet. *92*, 489–503.

Fudenberg, G., Getz, G., Meyerson, M., and Mirny, L.A. (2011). High order chromatin architecture shapes the landscape of chromosomal alterations in cancer. Nat. Biotechnol. *29*, 1109–1113.

Fukuda-Taira, S. (1981). Hepatic induction in the avian embryo: specificity of reactive endoderm and inductive mesoderm. J. Embryol. Exp. Morphol. *63*, 111–125.

Fullwood, M., Huang, P.Y.H., Han, Y., Handoko, L., Velkov, S., Wong, E., Cheung, E., Ruan, X., Wei, C.-L., Fullwood, M.J., et al. (2010). Protocol: Sonication-based Circular Chromosome Conformation Capture with next-generation sequencing analysis for the detection of chromatin interactions. Protoc. Exch.

Fullwood, M.J., Liu, M.H., Pan, Y.F., Liu, J., Xu, H., Mohamed, Y. Bin, Orlov, Y.L., Velkov, S., Ho, A., Mei, P.H., et al. (2009). An oestrogen-receptor-alpha-bound human chromatin interactome. Nature *462*, 58–64.

Gabellini, D., Green, M.R., and Tupler, R. (2002). Inappropriate gene activation in FSHD: a repressor complex binds a chromosomal repeat deleted in dystrophic muscle. Cell *110*, 339–348.

Gaj, T., Gersbach, C. a, and Barbas, C.F. (2013). ZFN, TALEN, and CRISPR/Cas-based methods for genome engineering. Trends Biotechnol. *31*, 397–405.

Gallagher, P.G., Steiner, L. a, Liem, R.I., Owen, A.N., Cline, A.P., Seidel, N.E., Garrett, L.J., and Bodine, D.M. (2010). Mutation of a barrier insulator in the human ankyrin-1 gene is associated with hereditary spherocytosis. J. Clin. Invest. *120*, 4453–4465.

Gao, N., White, P., and Kaestner, K.H. (2009). Article Establishment of Intestinal Identity and Epithelial-Mesenchymal Signaling by Cdx2. Dev. Cell *16*, 588–599.

Gartenberg, M.R., and Merkenschlager, M. (2008). Condensin goes with the family but not with the flow. Genome Biol. *9*, 236.

Gaszner, M., Vazquez, J., and Schedl, P. (1999). The Zw5 protein, a component of the scs chromatin domain boundary, is able to block enhancer-promoter interaction. Genes Dev. *13*, 2098–2107.

Gaulton, K.J., Nammo, T., Pasquali, L., Simon, J.M., Giresi, P.G., Fogarty, M.P., Panhuis, T.M., Mieczkowski, P., Secchi, A., Bosco, D., et al. (2010). A map of open chromatin in human pancreatic islets. Nat. Genet. *42*, 255–259.

Geiger, a, Salazar, G., Le Cam, a, and Kervran, a (2001). Characterization of an enhancer element in the proximal promoter of the mouse glucagon receptor gene. Biochim. Biophys. Acta *1517*, 236–242.

Gershenzon, N.I., and Ioshikhes, I.P. (2005). Synergy of human Pol II core promoter elements revealed by statistical sequence analysis. Bioinformatics *21*, 1295–1300.

Ghavi-helm, Y., Klein, F. a, Pakozdi, T., Ciglar, L., Noordermeer, D., Huber, W., and Furlong, E.E.M. (2014). Enhancer loops appear stable during development and are associated with paused polymerase. Nature *512*, 96–100.

Gifford, C., Ziller, M., Gu, H., and Trapnell, C. (2013). Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. Cell *153*, 1149–1163.

Gillies, S.D., Morrison, S.L., Oi, V.T., and Tonegawa, S. (1983). A tissue-specific transcription enhancer element is located in the major intron of a rearranged immunoglobulin heavy chain gene. Cell *33*, 717–728.

Giresi, P.G., Kim, J., McDaniell, R.M., Iyer, V.R., and Lieb, J.D. (2007). FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. Genome Res. *17*, 877–885.

Goessling, W., North, T.E., Lord, A.M., Ceol, C., Lee, S., Weidinger, G., Bourque, C., Strijbosch, R., Haramis, A.P., Puder, M., et al. (2008). APC mutant zebrafish uncover a changing temporal requirement for wnt signaling in liver development. Dev. Biol. *320*, 161–174.

Golan-Mashiach, M., Grunspan, M., Emmanuel, R., Gibbs-Bar, L., Dikstein, R., and Shapiro, E. (2012). Identification of CTCF as a master regulator of the clustered protocadherin genes. Nucleic Acids Res. *40*, 3378–3391.

Goldman, J. a, Garlick, J.D., and Kingston, R.E. (2010). Chromatin remodeling by imitation switch (ISWI) class ATP-dependent remodelers is stimulated by histone variant H2A.Z. J. Biol. Chem. *285*, 4645–4651.

Gosalia, N., Neems, D., Kerschner, J.L., Kosak, S.T., and Harris, A. (2014). Architectural proteins CTCF and cohesin have distinct roles in modulating the higher order structure and expression of the CFTR locus. Nucleic Acids Res. 1–11.

Gotea, V., Visel, A., Westlund, J.M., Nobrega, M.A., Pennacchio, L.A., and Ovcharenko, I. (2010). Homotypic clusters of transcription factor binding sites are a key component of human promoters and enhancers. Genome Res. *20*, 565–577.

Grainger, S., Savory, J.G.A., and Lohnes, D. (2010). Cdx2 regulates patterning of the intestinal epithelium. Dev. Biol. *339*, 155–165.

Grant, S.F.A., Reid, D.M., Blake, G., Herd, R., Fogelman, I., and Ralston, S.H. (1996). Reduced bone density and osteoporosis associated with a polymorphic Sp1 binding site in the collagen type I [alpha] 1 gene. Nat Genet *14*, 203–205.

Grapin-Botton, A., and Melton, D.A. (2000). Endoderm development: From patterning to organogenesis. Trends Genet. *16*, 124–130.

Green, M.D., Chen, A., Nostro, M., D'Souza, S.L., Schaniel, C., Lemischka, I.R., Gouon-Evans, V., Keller, G., and Snoeck, H. (2011). Generation of anterior foregut endoderm from human embryonic and induced pluripotent stem cells. Nat. Biotechnol. *29*, 267–272.

Gregorieff, A., Grosschedl, R., and Clevers, H. (2004). Hindgut defects and transformation of the gastro-intestinal tract in Tcf4(-/-)/Tcf1(-/-) embryos. EMBO J. *23*, 1825–1833.

Gross, D.S., and Garrard, W.T. (1988). Nuclease Hypersensitive Sites in Chromatin. Annu. Rev. Biochem. *57*, 159–197.

Grosveld, F., Assendelft, G. van, Greaves, D., and Kollias, G. (1987). Position-independent, high-level expression of the human β-globin gene in transgenic mice. Cell *51*, 975–985.

Grzeskowiak, R., Amin, J., Oetjen, E., and Knepel, W. (2000). Insulin responsiveness of the glucagon gene conferred by interactions between proximal promoter and more distal enhancer-like elements involving the paired-domain transcription factor Pax6. J. Biol. Chem. *275*, 30037–30045.

Gualdi, R., Bossard, P., Zheng, M., Hamada, Y., Coleman, J.R., and Zaret, K.S. (1996). Hepatic specification of the gut endoderm in vitro: Cell signaling and transcriptional control. Genes Dev. *10*, 1670–1682.

Handoko, L., Xu, H., Li, G., Ngan, C.Y., Chew, E., Schnapp, M., Lee, C.W.H., Ye, C., Ping, J.L.H., Mulawadi, F., et al. (2011). CTCF-mediated functional chromatin interactome in pluripotent cells. Nat. Genet. *43*, 630–638.

Hare, E.E., Peterson, B.K., Iyer, V.N., Meier, R., and Eisen, M.B. (2008). Sepsid even-skipped enhancers are functionally conserved in Drosophila despite lack of sequence conservation. PLoS Genet. *4*.

Harismendy, O., Notani, D., Song, X., Rahim, N.G., Tanasa, B., Heintzman, N., Ren, B., Fu, X.-D., Topol, E.J., Rosenfeld, M.G., et al. (2011). 9p21 DNA variants associated with coronary artery disease impair interferon-γ signalling response. Nature *470*, 264–268.

Harris, M.B., Mostecki, J., and Rothman, P.B. (2005). Repression of an interleukin-4-responsive promoter requires cooperative BCL-6 function. J. Biol. Chem. *280*, 13114–13121.

Hart, A.H., Hartley, L., Sourris, K., Stadler, E.S., Li, R., Stanley, E.G., Tam, P.P.L., Elefanty, A.G., and Robb, L. (2002). Mixl1 is required for axial mesendoderm morphogenesis and patterning in the murine embryo. Development *129*, 3597–3608.

Hartmann, H., Guthöhrlein, E.W., Siebert, M., Luehr, S., and Söding, J. (2013). P-value-based regulatory motif discovery using positional weight matrices. Genome Res. *23*, 181–194.

Hashimshony, T., Zhang, J., Keshet, I., Bustin, M., and Cedar, H. (2003). The role of DNA methylation in setting up chromatin structure during development. Nat. Genet. *34*, 187–192.

Hatton, C., Wilkie, A., Drysdale, H., Wood, W., Vickers, M., Sharpe, J., Ayyub, H., Pretorius, I., Buckle, V., and Higgs, D. (1990). Alpha-thalassemia caused by a large (62 kb) deletion upstream of the human alpha globin gene cluster. Blood *76*, 221–227.

Hawkins, R.D., Hon, G.C., Yang, C., Antosiewicz-Bourget, J.E., Lee, L.K., Ngo, Q.-M., Klugman, S., Ching, K.A., Edsall, L.E., Ye, Z., et al. (2011). Dynamic chromatin states in human ES cells reveal potential regulatory sequences and genes involved in pluripotency. Cell Res. *21*, 1393–1409.

Hazelett, D.J., Rhie, S.K., Gaddis, M., Yan, C., Lakeland, D.L., Coetzee, S.G., Henderson, B.E., Noushmehr, H., Cozen, W., Kote-Jarai, Z., et al. (2014). Comprehensive Functional Annotation of 77 Prostate Cancer Risk Loci. PLoS Genet. *10*.

He, A., Kong, S.W., Ma, Q., and Pu, W.T. (2011). Co-occupancy by multiple cardiac transcription factors identifies transcriptional enhancers active in heart. Proc. Natl. Acad. Sci. U. S. A. *108*, 5632–5637.

Heintzman, N.D., Hon, G.C., Hawkins, R.D., Kheradpour, P., Stark, A., Harp, L.F., Ye, Z., Lee, L.K., Stuart, R.K., Ching, C.W., et al. (2009). Histone modifications at human enhancers reflect global cell-type-specific gene expression. Nature *459*, 108–112.

Heinz, S., Benner, C., Spann, N., Bertolino, E., Lin, Y.C., Laslo, P., Cheng, J.X., Murre, C., Singh, H., and Glass, C.K. (2010). Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. Mol. Cell *38*, 576–589.

Hengartner, C.J., Myer, V.E., Liao, S.M., Wilson, C.J., Koh, S.S., and Young, R. a (1998). Temporal regulation of RNA polymerase II by Srb10 and Kin28 cyclin-dependent kinases. Mol. Cell *2*, 43–53.

Hering, N. a, Fromm, M., and Schulzke, J.-D. (2012). Determinants of colonic barrier function in inflammatory bowel disease and potential therapeutics. J. Physiol. *590*, 1035–1044.

Hnisz, D., Abraham, B.J., Lee, T.I., Lau, A., Saint-André, V., Sigova, A. a, Hoke, H. a, and Young, R. a (2013). Super-enhancers in the control of cell identity and disease. Cell *155*, 934–947.

Hobbs, K., Negri, J., Klinnert, M., Rosenwasser, L.J., and Borish, L. (1998). Interleukin-10 and transforming growth factor-beta promoter polymorphisms in allergies and asthma. Am. J. Respir. Crit. Care Med. *158*, 1958–1962.

Hock, T.D., Nick, H.S., and Agarwal, A. (2004). Upstream stimulatory factors, USF1 and USF2, bind to the human haem oxygenase-1 proximal promoter in vivo and regulate its transcription. Biochem. J. *383*, 209–218.

Hodge, R.D., Nelson, B.R., Kahoud, R.J., Yang, R., Mussar, K.E., Reiner, S.L., and Hevner, R.F. (2012). Tbr2 is essential for hippocampal lineage progression from neural stem cells to intermediate progenitors and neurons. J. Neurosci. *32*, 6275–6287.

Hoffman, M.M., Ernst, J., Wilder, S.P., Kundaje, A., Harris, R.S., Libbrecht, M., Giardine, B., Ellenbogen, P.M., Bilmes, J. a, Birney, E., et al. (2013). Integrative annotation of chromatin elements from ENCODE data. Nucleic Acids Res. *41*, 827–841.

Hong, S.K., Tsang, M., and Dawid, I.B. (2008). The Mych gene is required for neural crest survival during zebrafish development. PLoS One *3*.

De Hoon, M.J.L., Imoto, S., Nolan, J., and Miyano, S. (2004). Open source clustering software. Bioinformatics *20*, 1453–1454.

Hu, G., Cui, K., Northrup, D., Liu, C., Wang, C., Tang, Q., Ge, K., Levens, D., Crane-robinson, C., and Zhao, K. (2013). Article H2A . Z Facilitates Access of Active and Repressive Complexes to Chromatin in Embryonic Stem Cell Self-Renewal and Differentiation. Stem Cell *12*, 180–192.

Hudson, C., Clements, D., Friday, R. V., Stott, D., and Woodland, H.R. (1997). Xsox17-alpha and -beta mediate endoderm formation in xenopus. Cell *91*, 397–405.

Huelsken, J., Vogel, R., Brinkmann, V., Erdmann, B., Birchmeier, C., and Birchmeier, W. (2000). Requirement for β-catenin in anterior-posterior axis formation in mice. J. Cell Biol. *148*, 567–578.

Ishii, H., Sen, R., and Pazin, M.J. (2004). Combinatorial Control of DNase I-hypersensitive Site Formation and Erasure by Immunoglobulin Heavy Chain Enhancer-binding Proteins. J. Biol. Chem. *279*, 7331–7338.

Jensen, T., Jacquier, A., and Libri, D. (2013). Dealing with pervasive transcription. Mol. Cell *52*, 473–484.

Jin, C., Zang, C., Wei, G., Cui, K., Peng, W., Zhao, K., and Felsenfeld, G. (2009). H3.3/H2A.Z double variant-containing nucleosomes mark "nucleosome-free regions" of active promoters and other regulatory regions. Nat. Genet. *41*, 941–945.

Jin, F., Li, Y., Dixon, J.R., Selvaraj, S., Ye, Z., Lee, A.Y., Yen, C.-A., Schmitt, A.D., Espinoza, C. a, and Ren, B. (2013). A high-resolution map of the three-dimensional chromatin interactome in human cells. Nature *503*, 290–294.

Jin, O., Harpal, K., Ang, S.L., and Rossant, J. (2001). Otx2 and HNF3beta genetically interact in anterior patterning. Int. J. Dev. Biol. *45*, 357–365.

Jin, Q.W., Trelles-Sticken, E., Scherthan, H., and Loidl, J. (1998). Yeast nuclei display prominent centromere clustering that is reduced in nondividing cells and in meiotic prophase. J. Cell Biol. *141*, 21–29.

John, S., Reeves, R.B., Lin, J.X., Child, R., Leiden, J.M., Thompson, C.B., and Leonard, W.J. (1995). Regulation of cell-type-specific interleukin-2 receptor alpha-chain gene expression: potential role of physical interactions between Elf-1, HMG-I(Y), and NF-kappa B family proteins. Mol. Cell. Biol. *15*, 1786–1796.

Johnson, D.S., Mortazavi, A., Myers, R.M., and Wold, B. (2007). Genome-wide mapping of in vivo protein-DNA interactions. Science *316*, 1497–1502.

Jostins, L., Ripke, S., Weersma, R.K., Duerr, R.H., McGovern, D.P., Hui, K.Y., Lee, J.C., Schumm, L.P., Sharma, Y., Anderson, C. a, et al. (2012). Host-microbe interactions have shaped the genetic architecture of inflammatory bowel disease. Nature *491*, 119–124.

Junion, G., Spivakov, M., Girardot, C., Braun, M., Gustafson, E.H., Birney, E., and Furlong, E.E.M. (2012). A transcription factor collective defines cardiac cell fate and reflects lineage history. Cell *148*, 473–486.

Kadonaga, J.T. (2012). Perspectives on the RNA polymerase II core promoter. Wiley Interdiscip. Rev. Dev. Biol. *1*, 40–51.

Kagey, M.H., Newman, J.J., Bilodeau, S., Zhan, Y., Orlando, D. a, van Berkum, N.L., Ebmeier, C.C., Goossens, J., Rahl, P.B., Levine, S.S., et al. (2010). Mediator and cohesin connect gene expression and chromatin architecture. Nature *467*, 430–435.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2011). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kalhor, R., Tjong, H., Jayathilaka, N., Alber, F., and Chen, L. (2012). Genome architectures revealed by tethered chromosome conformation capture and population-based modeling. Nat. Biotechnol. *30*, 90–98.

Kapoor-Vazirani, P., and Vertino, P.M. (2014). A Dual Role for the Histone Methyltransferase PR-SET7/SETD8 and Histone H4 Lysine 20 Monomethylation in the Local Regulation of RNA Polymerase II Pausing. J. Biol. Chem. *289*, 7425–7437.

Katsumoto, K., and Kume, S. (2013). The role of CXCL12-CXCR4 signaling pathway in pancreatic development. Theranostics *3*, 11–17.

Kawakami, K. (2007). Tol2: a versatile gene transfer vector in vertebrates. Genome Biol. *8 Suppl 1*, S7.

Kellum, R., and Schedl, P. (1991). A position-effect assay for boundaries of higher order chromosomal domains. Cell *64*, 941–950.

Kilpeläinen, T.O., Zillikens, M.C., Stančákova, A., Finucane, F.M., Ried, J.S., Langenberg, C., Zhang, W., Beckmann, J.S., Luan, J., Vandenput, L., et al. (2011). Genetic variation near IRS1 associates with reduced adiposity and an impaired metabolic profile. Nat. Genet. *43*, 753–760.

Kim, S., Kim, Y.W., Shim, S.H., Kim, C.G., and Kim, A. (2012). Chromatin structure of the LCR in the human β-globin locus transcribing the adult δ- and β-globin genes. Int. J. Biochem. Cell Biol. *44*, 505–513.

Kim, S.W., Yoon, S.-J., Chuong, E., Oyolu, C., Wills, A.E., Gupta, R., and Baker, J. (2011). Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. Dev. Biol. *357*, 492–504.

Kim, T.H., Abdullaev, Z.K., Smith, A.D., Ching, K. a, Loukinov, D.I., Green, R.D., Zhang, M.Q., Lobanenkov, V. V, and Ren, B. (2007). Analysis of the vertebrate insulator protein CTCF-binding sites in the human genome. Cell *128*, 1231–1245.

Kim, T.-K., Hemberg, M., Gray, J.M., Costa, A.M., Bear, D.M., Wu, J., Harmin, D. a, Laptewicz, M., Barbara-Haley, K., Kuersten, S., et al. (2010). Widespread transcription at neuronal activity-regulated enhancers. Nature *465*, 182–187.

King, D.C., Taylor, J., Elnitski, L., Chiaromonte, F., Miller, W., and Hardison, R.C. (2005). Evaluation of regulatory potential and conservation scores for detecting cis-regulatory modules in aligned mammalian genome sequences. Genome Res. *15*, 1051–1060.

Kinkel, M.D., Eames, S.C., Alonzo, M.R., and Prince, V.E. (2008). Cdx4 is required in the endoderm to localize the pancreas and limit beta-cell number. Development *135*, 919–929.

Koch, F., Fenouil, R., Gut, M., Cauchy, P., Albert, T.K., Zacarias-Cabeza, J., Spicuglia, S., de la Chapelle, A.L., Heidemann, M., Hintermair, C., et al. (2011). Transcription initiation platforms and GTF recruitment at tissue-specific enhancers and promoters. Nat. Struct. Mol. Biol. *18*, 956–963.

Kochilas, L.K., Potluri, V., Gitler, A., Balasubramanian, K., and Chin, A.J. (2003). Cloning and characterization of zebrafish tbx1. Gene Expr. Patterns *3*, 645–651.

Kornberg, R.D. (1974). Chromatin Structure: A Repeating Unit of Histones and DNA. Science *184*, 868–871.

Kornberg, R.D., and Lorch, Y. (1999). Twenty-five years of the nucleosome, fundamental particle of the eukaryote chromosome. Cell *98*, 285–294.

Korzh, V. (2007). Transposons as tools for enhancer trap screens in vertebrates. Genome Biol. *8 Suppl 1*, S8.

Kouzarides, T. (2007). Chromatin modifications and their function. Cell *128*, 693–705.

Kudoh, T., Tsang, M., Hukriede, N. a, Chen, X., Dedekian, M., Clarke, C.J., Kiang, a, Schultz, S., Epstein, J. a, Toyama, R., et al. (2001). A gene expression screen in zebrafish embryogenesis. Genome Res. *11*, 1979–1987.

Kulkarni, M.M., and Arnosti, D.N. (2003). Information display by transcriptional enhancers. Development *130*, 6569–6575.

Kurukuti, S., Tiwari, V.K., Tavoosidana, G., Pugacheva, E., Murrell, A., Zhao, Z., Lobanenkov, V., Reik, W., and Ohlsson, R. (2006). CTCF binding at the H19 imprinting control region mediates maternally inherited higher-order chromatin conformation to restrict enhancer access to Igf2. Proc. Natl. Acad. Sci. U. S. A. *103*, 10684–10689.

Van der Laan, M.J., and Pollard, K.S. (2003). A new algorithm for hybrid hierarchical clustering with visualization and the booatstrap. J. Stat. Plan. Inference *117*, 275–303.

Laclef, C., Souil, E., Demignon, J., and Maire, P. (2003). Thymus, kidney and craniofacial abnormalities in Six1 deficient mice. Mech. Dev. *120*, 669–679.

Landegren, U., Kaiser, R., Sanders, J., and Hood, L. (1988). A ligase-mediated gene detection technique. Science *241*, 1077–1080.

Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., et al. (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860–921.

Landry, J.R., Kinston, S., Knezevic, K., Donaldson, I.J., Green, A.R., and Göttgens, B. (2005). Fli1, Elf1, and Ets1 regulate the proximal promoter of the LMO2 gene in endothelial cells. Blood *106*, 2680–2687.

Landy, A. (1989). Dynamic, Structural, and Regulatory Aspects of lambda Site-Specific Recombination. Annu. Rev. Biochem. *58*, 913–941.

Lauderdale, J.D., Wilensky, J.S., Oliver, E.R., Walton, D.S., and Glaser, T. (2000). 3' deletions cause aniridia by preventing PAX6 gene expression. Proc. Natl. Acad. Sci. U. S. A. *97*, 13755–13759.

Laukoetter, M.G., Nava, P., and Nusrat, A. (2008). Role of the intestinal barrier in inflammatory bowel disease. World J. Gastroenterol. *14*, 401–407.

Lee, C.S., Friedman, J.R., Fulmer, J.T., and Kaestner, K.H. (2005). The initiation of liver development is dependent on Foxa transcription factors. Nature *435*, 944–947.

Lee, H.J., Kim, E., and Kim, J.S. (2010). Targeted chromosomal deletions in human cells using zinc finger nucleases. Genome Res. *20*, 81–89.

Lettice, L.A., Horikoshi, T., Heaney, S.J.H., van Baren, M.J., van der Linde, H.C., Breedveld, G.J., Joosse, M., Akarsu, N., Oostra, B.A., Endo, N., et al. (2002). Disruption of a long-range cis-acting regulator for Shh causes preaxial polydactyly. Proc. Natl. Acad. Sci. U. S. A. *99*, 7548–7553.

Lettice, L.A., Heaney, S.J.H., Purdie, L.A., Li, L., de Beer, P., Oostra, B.A., Goode, D., Elgar, G., Hill, R.E., and de Graaff, E. (2003). A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. Hum. Mol. Genet. *12*, 1725–1735.

Levo, M., and Segal, E. (2014). In pursuit of design principles of regulatory sequences. Nat. Rev. Genet. *15*, 453–468.

Levy, S., Hannenhalli, S., and Workman, C. (2001). Enrichment of regulatory signals in conserved non-coding genomic sequence. Bioinformatics *17*, 871–877.

Li, C., and Manley, J.L. (1998). Even-skipped represses transcription by binding TATA binding protein and blocking the TFIID-TATA box interaction. Mol. Cell. Biol. *18*, 3771–3781.

Li, G., Fullwood, M.J., Xu, H., Mulawadi, F.H., Velkov, S., Vega, V., Ariyaratne, P.N., Mohamed, Y. Bin, Ooi, H.-S., Tennakoon, C., et al. (2010). ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. Genome Biol. *11*, R22.

Li, G., Ruan, X., Auerbach, R.K., Sandhu, K.S., Zheng, M., Wang, P., Poh, H.M., Goh, Y., Lim, J., Zhang, J., et al. (2012). Extensive Promoter-Centered Chromatin Interactions Provide a Topological Basis for Transcription Regulation. Cell *148*, 84–98.

Li, W., Notani, D., Ma, Q., Tanasa, B., Nunez, E., Chen, A.Y., Merkurjev, D., Zhang, J., Ohgi, K., Song, X., et al. (2013). Functional roles of enhancer RNAs for oestrogen-dependent transcriptional activation. Nature *498*, 516–520.

Li, X., MacArthur, S., Bourgon, R., Nix, D., Pollard, D. a, Iyer, V.N., Hechmer, A., Simirenko, L., Stapleton, M., Luengo Hendriks, C.L., et al. (2008). Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. PLoS Biol. *6*, e27.

Li, Kenneth R., X. Fang, G.S. (2002). Locus control regions. Blood *100*, 3077–3086.

Lian, X., Zhang, J., Azarin, S.M., Zhu, K., Hazeltine, L.B., Bao, X., Hsiao, C., Kamp, T.J., and Palecek, S.P. (2013). Directed cardiomyocyte differentiation from human pluripotent stem cells by modulating Wnt/β-catenin signaling under fully defined conditions. Nat. Protoc. *8*, 162–175.

Lieberman-Aiden, E., van Berkum, N.L., Williams, L., Imakaev, M., Ragoczy, T., Telling, A., Amit, I., Lajoie, B.R., Sabo, P.J., Dorschner, M.O., et al. (2009). Comprehensive mapping of long-range interactions reveals folding principles of the human genome. Science *326*, 289–293.

Ling, J.Q., Li, T., Hu, J.F., Vu, T.H., Chen, H.L., Qiu, X.W., Cherry, A.M., and Hoffman, A.R. (2006). CTCF mediates interchromosomal colocalization between Igf2/H19 and Wsb1/Nf1. Science *312*, 269–272.

Lister, R., Mukamel, E. a, Nery, J.R., Urich, M., Puddifoot, C. a, Johnson, N.D., Lucero, J., Huang, Y., Dwork, A.J., Schultz, M.D., et al. (2013). Global epigenomic reconfiguration during mammalian brain development. Science *341*, 1237905.

Liu, F., and Posakony, J.W. (2012). Role of architecture in the function and specificity of two notch-regulated transcriptional enhancer modules. PLoS Genet. *8*.

Liu, Z., and Garrard, W.T. (2005). Long-range interactions between three transcriptional enhancers, active Vkappa gene promoters, and a 3' boundary sequence spanning 46 kilobases. Mol. Cell. Biol. *25*, 3220–3231.

Liu, J.K., DiPersio, C.M., and Zaret, K.S. (1991). Extracellular signals that regulate liver transcription factors during hepatic differentiation in vitro. Mol. Cell. Biol. *11*, 773–784.

Liu, W., Ma, Q., Wong, K., Li, W., Ohgi, K., Zhang, J., Aggarwal, A.K., and Rosenfeld, M.G. (2013). Brd4 and JMJD6-associated anti-pause enhancers in regulation of transcriptional pause release. Cell *155*, 1581–1595.

Liu, Y., Festing, M., Thompson, J.C., Hester, M., Rankin, S., El-Hodiri, H.M., Zorn, A.M., and Weinstein, M. (2004). Smad2 and Smad3 coordinately regulate craniofacial and endodermal development. Dev. Biol. *270*, 411–426.

Liu, Z., Scannell, D.R., Eisen, M.B., and Tjian, R. (2011). Control of embryonic stem cell lineage commitment by core promoter factor, TAF3. Cell *146*, 720–731.

Loebel, D. a F., Studdert, J.B., Power, M., Radziewic, T., Jones, V., Coultas, L., Jackson, Y., Rao, R.S., Steiner, K., Fossat, N., et al. (2011). Rhou maintains the epithelial architecture and facilitates differentiation of the foregut endoderm. Development *138*, 4511–4522.

Loh, K.M., Ang, L.T., Zhang, J., Kumar, V., Ang, J., Auyeong, J.Q., Lee, K.L., Choo, S.H., Lim, C.Y.Y., Nichane, M., et al. (2014). Resource Efficient Endoderm Induction from Human Pluripotent Stem Cells by Logically Directing Signals Controlling Lineage Bifurcations. Stem Cell *14*, 237–252.

Lokmane, L., Haumaitre, C., Garcia-Villalba, P., Anselme, I., Schneider-Maunoury, S., and Cereghini, S. (2008). Crucial role of vHNF1 in vertebrate hepatic specification. Development *135*, 2777–2786.

Lomvardas, S., Barnea, G., Pisapia, D.J., Mendelsohn, M., Kirkland, J., and Axel, R. (2006). Interchromosomal Interactions and Olfactory Receptor Choice. Cell *126*, 403–413.

Loots, G.G. (2008). Genomic identification of regulatory elements by evolutionary sequence comparison and functional analysis. Adv. Genet. *61*, 269–293.

Loots, G.G., Locksley, R.M., Blankespoor, C.M., Wang, Z.E., Miller, W., Rubin, E.M., and Frazer, K.A. (2000). Identification of a coordinate regulator of interleukins 4, 13, and 5 by cross-species sequence comparisons. Science *288*, 136–140.

Loots, G.G., Kneissel, M., Keller, H., Baptist, M., Chang, J., Collette, N.M., Ovcharenko, D., Plajzer-Frick, I., and Rubin, E.M. (2005). Genomic deletion of a long-range bone enhancer misregulates sclerostin in Van Buchem disease. Genome Res. *15*, 928–935.

Lorentz, O., Duluc, I., De Arcangelis, A., Simon-Assmann, P., Kedinger, M., and Freund, J.N. (1997). Key role of the Cdx2 homeobox gene in extracellular matrix-mediated intestinal cell differentiation. J. Cell Biol. *139*, 1553–1565.

Lovén, J., Hoke, H. a, Lin, C.Y., Lau, A., Orlando, D. a, Vakoc, C.R., Bradner, J.E., Lee, T.I., and Young, R. a (2013). Selective inhibition of tumor oncogenes by disruption of super-enhancers. Cell *153*, 320–334.

Lu, C.C., Brennan, J., and Robertson, E.J. (2001). From fertilization to gastrulation: axis formation in the mouse embryo. Curr. Opin. Genet. Dev. *11*, 384–392.

Majumder, P., and Boss, J.M. (2010). CTCF controls expression and chromatin architecture of the human major histocompatibility complex class II locus. Mol. Cell. Biol. *30*, 4211–4223.

Malik, S., and Roeder, R. (2000). Transcriptional regulation through Mediator-like coactivators in yeast and metazoan cells. Trends Biochem. Sci. *25*, 277–283.

Manco, L., Ribeiro, M.L., Máximo, V., Almeida, H., Costa, a, Freitas, O., Barbot, J., Abade, a, and Tamagnini, G. (2000). A new PKLR gene mutation in the R-type promoter region affects the gene transcription causing pyruvate kinase deficiency. Br. J. Haematol. *110*, 993–997.

Maston, G.A., Evans, S.K., and Green, M.R. (2006). Transcriptional regulatory elements in the human genome. Annu. Rev. Genomics Hum. Genet. *7*, 29–59.

Maurano, M.T., Humbert, R., Rynes, E., Thurman, R.E., Haugen, E., Wang, H., Reynolds, A.P., Sandstrom, R., Qu, H., Brody, J., et al. (2012). Systematic Localization of Common Disease-Associated Variation in Regulatory DNA. Science *337*, 1190–1195.

May, D., Blow, M.J., Kaplan, T., Mcculley, D.J., Jensen, B.C., Akiyama, J.A., Holt, A., Plajzer-frick, I., Shoukry, M., Wright, C., et al. (2011). Large-scale discovery of enhancers from human heart tissue. Nat. Genet. *44*, 89–93.

McGuckin, M. a, Eri, R., Simms, L. a, Florin, T.H.J., and Radford-Smith, G. (2009). Intestinal barrier dysfunction in inflammatory bowel diseases. Inflamm. Bowel Dis. *15*, 100–113.

McKay, D.J., and Lieb, J.D. (2013). A common set of DNA regulatory elements shapes Drosophila appendages. Dev. Cell *27*, 306–318.

McKnight, K.D., Wang, P., and Kim, S.K. (2010). Deconstructing Pancreas Development to Reconstruct Human Islets from Pluripotent Stem Cells. Cell Stem Cell *6*, 300–308.

McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nat. Biotechnol. *28*, 495–501.

McLin, V. a, Rankin, S. a, and Zorn, A.M. (2007). Repression of Wnt/beta-catenin signaling in the anterior endoderm is essential for liver and pancreas development. Development *134*, 2207–2217.

McPherson, J.D., Marra, M., Hillier, L., Waterston, R.H., Chinwalla, a, Wallis, J., Sekhon, M., Wylie, K., Mardis, E.R., Wilson, R.K., et al. (2001). A physical map of the human genome. Nature *409*, 934–941.

Meno, C., Gritsman, K., Ohishi, S., Ohfuji, Y., Heckscher, E., Mochida, K., Shimono, A., Kondoh, H., Talbot, W.S., Robertson, E.J., et al. (1999). Mouse lefty2 and zebrafish antivin are feedback inhibitors of nodal signaling during vertebrate gastrulation. Mol. Cell *4*, 287–298.

Mercer, T.R., and Mattick, J.S. (2013). Structure and function of long noncoding RNAs in epigenetic regulation. Nat. Struct. Mol. Biol. *20*, 300–307.

Merika, M., and Thanos, D. (2001). Enhanceosomes. Curr Opin Genet Dev *11*, 205–208.

Meyer, K.B., O'Reilly, M., Michailidou, K., Carlebur, S., Edwards, S.L., French, J.D., Prathalingham, R., Dennis, J., Bolla, M.K., Wang, Q., et al. (2013). Fine-scale mapping of the FGFR2 breast cancer risk locus: putative functional variants differentially bind FOXA1 and E2F1. Am. J. Hum. Genet. *93*, 1046–1060.

Mikawa, T., Poh, A.M., Kelly, K. a, Ishii, Y., and Reese, D.E. (2004). Induction and patterning of the primitive streak, an organizing center of gastrulation in the amniote. Dev. Dyn. *229*, 422–432.

Mikkelsen, T., Ku, M., Jaffe, D., and Issac, B. (2007). Genome-wide maps of chromatin state in pluripotent and lineage-committed cells. Nature *448*.

Misteli, T. (2010). Higher-order genome organization in human disease. Cold Spring Harb Perspect Biol *2*, a000794.

Moffat, G.J., McLaren, a. W., and Wolf, C.R. (1996). Functional Characterization of the Transcription Silencer Element Located within the Human Pi Class Glutathione S-Transferase Promoter. J. Biol. Chem. *271*, 20740–20747.

Mohandas, T., Sparkes, R.S., and Shapiro, L.J. (1981). Reactivation of an inactive human X chromosome: evidence for X inactivation by DNA methylation. Science *211*, 393–396.

Moorman, C., Sun, L. V, Wang, J., de Wit, E., Talhout, W., Ward, L.D., Greil, F., Lu, X., White, K.P., Bussemaker, H.J., et al. (2006). Hotspots of transcription factor colocalization in the genome of Drosophila melanogaster. Proc. Natl. Acad. Sci. U. S. A. *103*, 12027–12032.

Mozas, P., Galetto, R., Albajar, M., Ros, E., Pocoví, M., and Rodríguez-Rey, J.C. (2002). A mutation (-49C>T) in the promoter of the low density lipoprotein receptor gene associated with familial hypercholesterolemia. J. Lipid Res. *43*, 13–18.

Mullen, A.C., Orlando, D. a, Newman, J.J., Lovén, J., Kumar, R.M., Bilodeau, S., Reddy, J., Guenther, M.G., DeKoter, R.P., and Young, R. a (2011). Master transcription factors determine cell-type-specific responses to TGF-β signaling. Cell *147*, 565–576.

Murakami, K., Elmlund, H., and Kalisman, N. (2013). Architecture of an RNA polymerase II transcription pre-initiation complex. Science (80-. ). *342*, 1238724.

Murrell, A., Heeson, S., and Reik, W. (2004). Interaction between differentially methylated regions partitions the imprinted genes Igf2 and H19 into parent-specific chromatin loops. Nat. Genet. *36*, 889–893.

Murtha, M., Tokcaer-Keskin, Z., Tang, Z., Strino, F., Chen, X., Wang, Y., Xi, X., Basilico, C., Brown, S., Bonneau, R., et al. (2014). FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. Nat. Methods *11*, 559–565.

Naranjo, S., Voesenek, K., de la Calle-Mustienes, E., Robert-Moreno, A., Kokotas, H., Grigoriadou, M., Economides, J., Van Camp, G., Hilgert, N., Moreno, F., et al. (2010). Multiple enhancers located in a 1-Mb region upstream of POU3F4 promote expression during inner ear development and may be required for hearing. Hum. Genet. *128*, 411–419.

Narlikar, L., Sakabe, N.J., Blanski, A.A., Arimura, F.E., Westlund, J.M., Nobrega, M.A., and Ovcharenko, I. (2010). Genome-wide discovery of human heart enhancers. Genome Res. *20*, 381–392.

Neely, K., Hassan, A., and Wallberg, A. (1999). Activation Domain–Mediated Targeting of the SWI/SNF Complex to Promoters Stimulates Transcription from Nucleosome Arrays. Mol. Cell *4*, 649–655.

Nóbrega, M. a, Zhu, Y., Plajzer-Frick, I., Afzal, V., and Rubin, E.M. (2004). Megabase deletions of gene deserts result in viable mice. Nature *431*, 988–993.

O'Brien, T., Hardin, S., Greenleaf, A., and Lis, J.T. (1994). Phosphorylation of RNA polymerase II C-terminal domain and transcriptional elongation. Nature *370*, 75–77.

O'Kane, C.J., and Gehring, W.J. (1987). Detection in situ of genomic regulatory elements in Drosophila. Proc. Natl. Acad. Sci. U. S. A. *84*, 9123–9127.

Ochi, H., Tamai, T., Nagano, H., Kawaguchi, A., Sudou, N., and Ogino, H. (2012). Evolution of a tissue-specific silencer underlies divergence in the expression of pax2 and pax8 paralogues. Nat. Commun. *3*, 848.

Ohlsson, R., and Göndör, A. (2007). The 4C technique: the "Rosetta stone" for genome biology in 3D? Curr. Opin. Cell Biol. *19*, 321–325.

Oliveira, A.M.M., Hemstedt, T.J., and Bading, H. (2012). Rescue of aging-associated decline in Dnmt3a2 expression restores cognitive abilities. Nat. Neurosci. *15*, 1111–1113.

Ong, C.-T., and Corces, V.G. (2014). CTCF: an architectural protein bridging genome topology and function. Nat. Rev. Genet. *15*, 234–246.

Orford, K., Kharchenko, P., Lai, W., Dao, M.C., Worhunsky, D.J., Ferro, A., Janzen, V., Park, P.J., and Scadden, D.T. (2008). Differential H3K4 Methylation Identifies Developmentally Poised Hematopoietic Genes. Dev. Cell *14*, 798–809.

Orphanides, G., Lagrange, T., and Reinberg, D. (1996). The general transcription factors of RNA polymerase II. Genes Dev. *10*, 2657–2683.

Pabo, C., and Sauer, R. (1992). Transcription factors: structural families and principles of DNA recognition. Annu. Rev. Biochem. *61*, 1053–1095.

Panne, D., Maniatis, T., and Harrison, S.C. (2004). Crystal structure of ATF-2/c-Jun and IRF-3 bound to the interferon-beta enhancer. EMBO J. *23*, 4384–4393.

Panne, D., Maniatis, T., and Harrison, S.C. (2007). An atomic model of the interferon-beta enhanceosome. Cell *129*, 1111–1123.

Park, P. (2009). ChIP–seq: advantages and challenges of a maturing technology. Nat. Rev. Genet. *10*, 669–680.

Park, D., Lee, Y., Bhupindersingh, G., and Iyer, V.R. (2013). Widespread misinterpretable ChIP-seq bias in yeast. PLoS One *8*.

Parker, S.C.J., Stitzel, M.L., Taylor, D.L., Orozco, J.M., Erdos, M.R., Akiyama, J. a, van Bueren, K.L., Chines, P.S., Narisu, N., Black, B.L., et al. (2013). Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. Proc. Natl. Acad. Sci. U. S. A. *110*, 17921–17926.

Pennacchio, L.A., Ahituv, N., Moses, A.M., Prabhakar, S., Nobrega, M.A., Shoukry, M., Minovitsky, S., Dubchak, I., Holt, A., Lewis, K.D., et al. (2006). In vivo enhancer analysis of human conserved non-coding sequences. Nature *444*, 499–502.

Pera, M.F., and Tam, P.P.L. (2010). Extrinsic regulation of pluripotent stem cells. Nature *465*, 713–720.

Peric-Hupkes, D., Meuleman, W., Pagie, L., Bruggeman, S.W.M., Solovei, I., Brugman, W., Gräf, S., Flicek, P., Kerkhoven, R.M., van Lohuizen, M., et al. (2010). Molecular Maps of the Reorganization of Genome-Nuclear Lamina Interactions during Differentiation. Mol. Cell *38*, 603–613.

Phatnani, H.P., and Greenleaf, A.L. (2006). Phosphorylation and functions of the RNA polymerase II CTD. Genes Dev. *20*, 2922–2936.

Phillips, J.E., and Corces, V.G. (2009). CTCF: Master Weaver of the Genome. Cell *137*, 1194–1211.

Pickersgill, H., Kalverda, B., de Wit, E., Talhout, W., Fornerod, M., and van Steensel, B. (2006). Characterization of the Drosophila melanogaster genome at the nuclear lamina. Nat. Genet. *38*, 1005–1014.

Privalsky, M.L. (2004). The role of corepressors in transcriptional regulation by nuclear hormone receptors. Annu. Rev. Physiol. *66*, 315–360.

Qin, Y., Kong, L., Poirier, C., Truong, C., Overbeek, P. a, and Bishop, C.E. (2004). Long-range activation of Sox9 in Odd Sex (Ods) mice. Hum. Mol. Genet. *13*, 1213–1218.

Rada-Iglesias, A., Bajpai, R., Swigut, T., Brugmann, S.A., Flynn, R.A., and Wysocka, J. (2011). A unique chromatin signature uncovers early developmental enhancers in humans. Nature *470*, 279–283.

Rada-Iglesias, A., Bajpai, R., Prescott, S., Brugmann, S. a, Swigut, T., and Wysocka, J. (2012). Epigenomic annotation of enhancers predicts transcriptional regulators of human neural crest. Cell Stem Cell *11*, 633–648.

Raghu Ram, E.V.S., and Meshorer, E. (2009). Transcriptional competence in pluripotency. Genes Dev. *23*, 2793–2798.

Rastegar, S., Hess, I., Dickmeis, T., Nicod, J.C., Ertzer, R., Hadzhiev, Y., Thies, W.G., Scherer, G., and Strähle, U. (2008). The words of the regulatory code are arranged in a variable manner in highly conserved enhancers. Dev. Biol. *318*, 366–377.

Reitsma, P., Bertina, R., Amstel, J.P. van, Riemens, A., and Briet, E. (1988). The putative factor IX gene promoter in hemophilia B Leyden. Blood *72*, 1074–1076.

Ren, X., Siegel, R., Kim, U., and Roeder, R.G. (2011). Direct Interactions of OCA-B and TFII-I Regulate Immunoglobulin Heavy-Chain Gene Transcription by Facilitating Enhancer-Promoter Communication. Mol. Cell *42*, 342–355.

Reynolds, N., O'Shaughnessy, A., and Hendrich, B. (2013). Transcriptional repressors: multifaceted regulators of gene expression. Development *140*, 505–512.

Rhee, H.S., and Pugh, B.F. (2011). Resource Comprehensive Genome-wide Protein-DNA Interactions Detected at Single-Nucleotide Resolution. Cell *147*, 1408–1419.

Rhie, S.K., Coetzee, S.G., Noushmehr, H., Yan, C., Kim, J.M., Haiman, C.A., and Coetzee, G.A. (2013). Comprehensive functional annotation of seventy-one breast cancer risk Loci. PLoS One *8*, e63925.

Rivera, C.M., and Ren, B. (2013). Mapping human epigenomes. Cell *155*, 39–55.

Rojas, A., Schachterle, W., Xu, S., Martín, F., and Black, B.L. (2010). Direct transcriptional regulation of Gata4 during early endoderm specification is controlled by FoxA2 binding to an intronic enhancer. Dev. Biol. *346*, 346–355.

Roy, S., Tan, Y.Y., and Hart, C.M. (2007). A genetic screen supports a broad role for the Drosophila insulator proteins BEAF-32A and BEAF-32B in maintaining patterns of gene expression. Mol. Genet. Genomics *277*, 273–286.

Rung, J., Cauchi, S., Albrechtsen, A., Shen, L., Rocheleau, G., Cavalcanti-Proença, C., Bacot, F., Balkau, B., Belisle, A., Borch-Johnsen, K., et al. (2009). Genetic variant near IRS1 is associated with type 2 diabetes, insulin resistance and hyperinsulinemia. Nat. Genet. *41*, 1110–1115.

Russ, a P., Wattler, S., Colledge, W.H., Aparicio, S. a, Carlton, M.B., Pearce, J.J., Barton, S.C., Surani, M. a, Ryan, K., Nehls, M.C., et al. (2000). Eomesodermin is required for mouse trophoblast development and mesoderm formation. Nature *404*, 95–99.

Ryba, T., Hiratani, I., Lu, J., Itoh, M., Kulik, M., Zhang, J., Schulz, T.C., Robins, A.J., Dalton, S., and Gilbert, D.M. (2010). Evolutionarily conserved replication timing profiles predict long-range chromatin interactions and distinguish closely related cell types. Genome Res. *20*, 761–770.

Saldanha, A.J. (2004). Java Treeview--extensible visualization of microarray data. Bioinformatics *20*, 3246–3248.

Salim, S.Y., and Söderholm, J.D. (2011). Importance of disrupted intestinal barrier in inflammatory bowel diseases. Inflamm. Bowel Dis. *17*, 362–381.

Sandmann, T., Girardot, C., Brehme, M., Tongprasit, W., Stolc, V., and Furlong, E.E.M. (2007). A core transcriptional network for early mesoderm development in Drosophila melanogaster. Genes Dev. *21*, 436–449.

Sanyal, A., Lajoie, B.R., Jain, G., and Dekker, J. (2012). The long-range interaction landscape of gene promoters. Nature *489*, 109–113.

Schier, A.F. (2003). Nodal signaling in vertebrate development. Annu. Rev. Cell Dev. Biol. *19*, 589–621.

Schoenfelder, S., Sexton, T., Chakalova, L., Cope, N.F., Horton, A., Andrews, S., Kurukuti, S., Mitchell, J.A., Umlauf, D., Dimitrova, D.S., et al. (2009). Preferential associations between co-regulated genes reveal a transcriptional interactome in erythroid cells. Nat. Genet. *42*, 53–61.

Schones, D.E., Schones, D.E., Cui, K., Cui, K., Cuddapah, S., Cuddapah, S., Roh, T., Roh, T., Barski, A., Barski, A., et al. (2008). Dynamic regulation of nucleosome positioning in the human genome. Cell *132*, 887–898.

Seipel, K., Georgiev, O., and Schaffner, W. (1992). Different activation domains stimulate transcription from remote (’enhancer') and proximal ('promoter') positions. EMBO J. *11*, 4961–4968.

Senger, K., Armstrong, G.W., Rowell, W.J., Kwan, J.M., Markstein, M., and Levine, M. (2004). Immunity Regulatory DNAs Share Common Organizational Features in Drosophila. Mol. Cell *13*, 19–32.

Sexton, T., Yaffe, E., Kenigsberg, E., Bantignies, F., Leblanc, B., Hoichman, M., Parrinello, H., Tanay, A., and Cavalli, G. (2012). Three-dimensional folding and functional organization principles of the Drosophila genome. Cell *148*, 458–472.

Shen, M.M. (2007). Nodal signaling: developmental roles and regulation. Development *134*, 1023–1034.

Shen, L., Su, L., and Turner, J.R. (2009). Mechanisms and functional implications of intestinal barrier defects. Dig. Dis. *27*, 443–449.

Shen, Y., Yue, F., McCleary, D.F., Ye, Z., Edsall, L., Kuan, S., Wagner, U., Dixon, J., Lee, L., Lobanenkov, V. V., et al. (2012). A map of the cis-regulatory sequences in the mouse genome. Nature *488*, 116–120.

Shlyueva, D., Stampfel, G., and Stark, A. (2014). Transcriptional enhancers: from properties to genome-wide predictions. Nat. Rev. Genet. *15*, 272–286.

Shook, D., and Keller, R. (2003). Mechanisms, mechanics and function of epithelial–mesenchymal transitions in early development. Mech. Dev. *120*, 1351–1383.

Siggers, T., and Gordân, R. (2014). Protein-DNA binding: Complexities and multi-protein codes. Nucleic Acids Res. *42*, 2099–2111.

Simonis, M., Klous, P., Splinter, E., Moshkin, Y., Willemsen, R., de Wit, E., van Steensel, B., and de Laat, W. (2006). Nuclear organization of active and inactive chromatin domains uncovered by chromosome conformation capture-on-chip (4C). Nat. Genet. *38*, 1348–1354.

Smale, S.T., and Kadonaga, J.T. (2003). The RNA polymerase II core promoter. Annu. Rev. Biochem. *72*, 449–479.

Smemo, S., Tena, J.J., Kim, K.-H., Gamazon, E.R., Sakabe, N.J., Gómez-Marín, C., Aneas, I., Credidio, F.L., Sobreira, D.R., Wasserman, N.F., et al. (2014). Obesity-associated variants within FTO form long-range functional connections with IRX3. Nature *507*, 371–375.

Smith, R.P., Taher, L., Patwardhan, R.P., Kim, M.J., Inoue, F., Shendure, J., Ovcharenko, I., and Ahituv, N. (2013). Massively parallel decoding of mammalian regulatory sequences supports a flexible organizational model. Nat. Genet. *45*, 1021–1028.

Song, L., and Crawford, G.E. (2010). DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. Cold Spring Harb. Protoc. *5*.

Song, J., Kim, H.J., Gong, Z., Liu, N.-A., and Lin, S. (2007). Vhnf1 acts downstream of Bmp, Fgf, and RA signals to regulate endocrine beta cell development in zebrafish. Dev. Biol. *303*, 561–575.

Spilianakis, C.G., Lalioti, M.D., Town, T., Lee, G.R., and Flavell, R. a (2005). Interchromosomal associations between alternatively expressed loci. Nature *435*, 637–645.

Splinter, E., Heath, H., Kooren, J., Palstra, R.J., Klous, P., Grosveld, F., Galjart, N., and De Laat, W. (2006). CTCF mediates long-range chromatin looping and local histone modification in the ??-globin locus. Genes Dev. *20*, 2349–2354.

Srinivasan, L., and Atchison, M.L. (2004). YY1 DNA binding and PcG recruitment requires CtBP. Genes Dev. *18*, 2596–2601.

Stadler, M.B., Murr, R., Burger, L., Ivanek, R., Lienert, F., Schöler, A., van Nimwegen, E., Wirbelauer, C., Oakeley, E.J., Gaidatzis, D., et al. (2011). DNA-binding factors shape the mouse methylome at distal regulatory regions. Nature *480*, 490–495.

Stanley, E.G., Biben, C., Allison, J., Hartley, L., Wicks, I.P., Campbell, I.K., McKinley, M., Barnett, L., Koentgen, F., Robb, L., et al. (2000). Targeted insertion of a lacZ reporter gene into the mouse Cer1 locus reveals complex and dynamic expression during embryogenesis. Genesis *26*, 259–264.

Van Steensel, B., and Henikoff, S. (2000). Identification of in vivo DNA targets of chromatin proteins using tethered dam methyltransferase. Nat. Biotechnol. *18*, 424–428.

Stumm, R.K., Zhou, C., Ara, T., Lazarini, F., Dubois-Dalcq, M., Nagasawa, T., Höllt, V., and Schulz, S. (2003). CXCR4 regulates interneuron migration in the developing neocortex. J. Neurosci. *23*, 5123–5130.

Subramanian, A., Subramanian, A., Tamayo, P., Tamayo, P., Mootha, V.K., Mootha, V.K., Mukherjee, S., Mukherjee, S., Ebert, B.L., Ebert, B.L., et al. (2005). Gene set enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proc. Natl. Acad. Sci. U. S. A. *102*, 15545–15550.

Suh, E., and Traber, P.G. (1996). An intestine-specific homeobox gene regulates proliferation and differentiation. Mol. Cell. Biol. *16*, 619–625.

Tada, S., Era, T., Furusawa, C., Sakurai, H., Nishikawa, S., Kinoshita, M., Nakao, K., Chiba, T., and Nishikawa, S.-I. (2005). Characterization of mesendoderm: a diverging point of the definitive endoderm and mesoderm in embryonic stem cell differentiation culture. Development *132*, 4363–4374.

Taher, L., Smith, R.P., Kim, M.J., Ahituv, N., and Ovcharenko, I. (2013). Sequence signatures extracted from proximal promoters can be used to predict distal enhancers. Genome Biol. *14*, R117.

Takasato, M., Er, P.X., Becroft, M., Vanslambrouck, J.M., Stanley, E.G., Elefanty, A.G., and Little, M.H. (2014). Directing human embryonic stem cell differentiation towards a renal lineage generates a self-organizing kidney. Nat. Cell Biol. *16*, 118–126.

Tam, P.P.L., and Loebel, D. a F. (2007). Gene function in mouse embryogenesis: get set for gastrulation. Nat. Rev. Genet. *8*, 368–381.

Tam, P.P.L., Khoo, P.-L., Lewis, S.L., Bildsoe, H., Wong, N., Tsang, T.E., Gad, J.M., and Robb, L. (2007). Sequential allocation and global pattern of movement of the definitive endoderm in the mouse embryo during gastrulation. Development *134*, 251–260.

Tamplin, O.J., Kinzel, D., Cox, B.J., Bell, C.E., Rossant, J., and Lickert, H. (2008). Microarray analysis of Foxa2 mutant mouse embryos reveals novel gene expression and inductive roles for the gastrula organizer and its derivatives. BMC Genomics *9*, 511.

Teo, A.K.K., Arnold, S.J., Trotter, M.W.B., Brown, S., Ang, L.T., Chng, Z., Robertson, E.J., Dunn, N.R., and Vallier, L. (2011). Pluripotency factors regulate definitive endoderm specification through eomesodermin. Genes Dev. *25*, 238–250.

Teytelman, L., Thurtle, D.M., Rine, J., and van Oudenaarden, A. (2013). Highly expressed loci are vulnerable to misleading ChIP localization of multiple unrelated proteins. Proc. Natl. Acad. Sci. U. S. A. *110*, 18602–18607.

Thambirajah, A. a, Dryhurst, D., Ishibashi, T., Li, A., Maffey, A.H., and Ausió, J. (2006). H2A.Z stabilizes chromatin in a way that is dependent on core histone acetylation. J. Biol. Chem. *281*, 20036–20044.

Thanos, D., and Maniatis, T. (1995). Virus induction of human IFNβ gene expression requires the assembly of an enhanceosome. Cell *83*, 1091–1100.

Thurman, R.E., Rynes, E., Humbert, R., Vierstra, J., Maurano, M.T., Haugen, E., Sheffield, N.C., Stergachis, A.B., Wang, H., Vernot, B., et al. (2012). The accessible chromatin landscape of the human genome. Nature *489*, 75–82.

Tolhuis, B., Palstra, R., Splinter, E., Grosveld, F., and Laat, W. De (2002). Looping and Interaction between Hypersensitive Sites in the Active ␤ -globin Locus. *10*, 1453–1465.

Tolhuis, B., Wit, E. De, Muijrers, I., Teunissen, H., Talhout, W., and Steensel, B. Van (2006). Genome-wide profiling of PRC1 and PRC2 Polycomb chromatin binding in Drosophila melanogaster. *38*, 694–700.

Touboul, T., Hannan, N.R.F., Corbineau, S., Martinez, A., Martinet, C., Branchereau, S., Mainot, S., Strick-Marchand, H., Pedersen, R., Santo, J. Di, et al. (2010). Generation of functional hepatocytes from human embryonic stem cells under chemically defined conditions that recapitulate liver development. Hepatology *51*, 1754–1765.

Trompouki, E., Bowman, T. V, Lawton, L.N., Fan, Z.P., Wu, D.-C., DiBiase, A., Martin, C.S., Cech, J.N., Sessa, A.K., Leblanc, J.L., et al. (2011). Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. Cell *147*, 577–589.

Tsankov, A.M., Gu, H., Akopian, V., Ziller, M.J., Donaghey, J., Amit, I., Gnirke, A., and Meissner, A. (2015). Transcription factor binding dynamics during human ES cell differentiation. Nature *518*, 344–349.

Umbarger, M.A., Toro, E., Wright, M.A., Porreca, G.J., Baù, D., Hong, S.H., Fero, M.J., Zhu, L.J., Marti-Renom, M.A., McAdams, H.H., et al. (2011). The three-dimensional architecture of a bacterial genome and its alteration by genetic perturbation. Mol. Cell *44*, 252–264.

Urnov, F.D., Rebar, E.J., Holmes, M.C., Zhang, H.S., and Gregory, P.D. (2010). Genome editing with engineered zinc finger nucleases. Nat. Rev. Genet. *11*, 636–646.

Valenzuela, L., and Kamakaka, R.T. (2006). Chromatin insulators. Annu. Rev. Genet. *40*, 107–138.

De Vas, M.G., Kopp, J.L., Heliot, C., Sander, M., Cereghini, S., and Haumaitre, C. (2015). Hnf1b controls pancreas morphogenesis and the generation of Ngn3+ endocrine progenitors. Dev. *142* , 871–882.

Vernimmen, D., De Gobbi, M., Sloane-Stanley, J.A., Wood, W.G., and Higgs, D.R. (2007). Long-range chromosomal interactions regulate the timing of the transition between poised and active gene expression. EMBO J. *26*, 2041–2051.

Verzi, M.P., Hatzis, P., Sulahian, R., Philips, J., Schuijers, J., Shin, H., Freed, E., Lynch, J.P., Dang, D.T., Brown, M., et al. (2010). TCF4 and CDX2, major transcription factors for intestinal function, converge on the same cis-regulatory regions. Proc. Natl. Acad. Sci. U. S. A. *107*, 15157–15162.

Vincent, S.D., Dunn, N.R., Hayashi, S., Norris, D.P., and Robertson, E.J. (2003). Cell fate decisions within the mouse organizer are governed by graded Nodal signals. Genes Dev. *17*, 1646–1662.

Visel, A., Prabhakar, S., Akiyama, J. a, Shoukry, M., Lewis, K.D., Holt, A., Plajzer-Frick, I., Afzal, V., Rubin, E.M., and Pennacchio, L. a (2008). Ultraconservation identifies a small subset of extremely constrained developmental enhancers. Nat. Genet. *40*, 158–160.

Visel, A., Rubin, E.M., and Pennacchio, L. a (2009a). Genomic views of distant-acting enhancers. Nature *461*, 199–205.

Visel, A., Blow, M.J., Li, Z., Zhang, T., Akiyama, J.A., Holt, A., Plajzer-Frick, I., Shoukry, M., Wright, C., Chen, F., et al. (2009b). ChIP-seq accurately predicts tissue-specific activity of enhancers. Nature *457*, 854–858.

Vogel, M.J., Peric-Hupkes, D., and van Steensel, B. (2007). Detection of in vivo protein-DNA interactions using DamID in mammalian cells. Nat. Protoc. *2*, 1467–1478.

Vorachek, W.R., Steppan, C.M., Lima, M., Black, H., Bhattacharya, R., Wen, P., Kajiyama, Y., and Locker, J. (2000). Distant enhancers stimulate the albumin promoter through complex proximal binding sites. J. Biol. Chem. *275*, 29031–29041.

Wang, D., Garcia-bassets, I., Benner, C., Li, W., Su, X., Zhou, Y., Qiu, J., and Liu, W. (2011). Reprogramming transcription by distinct classes of enhancers functionally defined by eRNA. Nature *474*, 390–394.

Wang, H.L., Wu, T., Chang, W.T., Li, a H., Chen, M.S., Wu, C.Y., and Fang, W. (2000). Point mutation associated with X-linked dominant Charcot-Marie-Tooth disease impairs the P2 promoter activity of human connexin-32 gene. Brain Res. Mol. Brain Res. *78*, 146–153.

Wang, Z., Zang, C., Rosenfeld, J.A., Schones, D.E., Barski, A., Cuddapah, S., Cui, K., Roh, T.-Y., Peng, W., Zhang, M.Q., et al. (2008). Combinatorial patterns of histone acetylations and methylations in the human genome. Nat. Genet. *40*, 897–903.

Wang, Z., Zang, C., Cui, K., Schones, D.E., Barski, A., Peng, W., and Zhao, K. (2009). Resource Genome-wide Mapping of HATs and HDACs Reveals Distinct Functions in Active and Inactive Genes. Cell *138*, 1019–1031.

Welter, D., MacArthur, J., Morales, J., Burdett, T., Hall, P., Junkins, H., Klemm, A., Flicek, P., Manolio, T., Hindorff, L., et al. (2014). The NHGRI GWAS Catalog, a curated resource of SNP-trait associations. Nucleic Acids Res. *42*, D1001–D1006.

Whyte, W. a, Orlando, D. a, Hnisz, D., Abraham, B.J., Lin, C.Y., Kagey, M.H., Rahl, P.B., Lee, T.I., and Young, R. a (2013). Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell *153*, 307–319.

Wiedenheft, B., Sternberg, S.H., and Doudna, J.A. (2012). RNA-guided genetic silencing systems in bacteria and archaea. Nature *482*, 331–338.

Wijchers, P.J., and de Laat, W. (2011). Genome organization influences partner selection for chromosomal rearrangements. Trends Genet. *27*, 63–71.

Williamson, I., Hill, R.E., and Bickmore, W.A. (2011). Forum Enhancers : From Developmental Genetics to the Genetics of Common Human Disease. Dev. Cell *21*, 17–19.

De Wit, E., and de Laat, W. (2012). A decade of 3C technologies: Insights into nuclear organization. Genes Dev. *26*, 11–24.

Won, K.-J., Zhang, X., Wang, T., Ding, B., Raha, D., Snyder, M., Ren, B., and Wang, W. (2013). Comparative annotation of functional regions in the human genome using epigenomic data. Nucleic Acids Res. *41*, 4423–4432.

Wood, W.M., Dowding, J.M., Sarapura, V.D., McDermott, M.T., Gordon, D.F., and Ridgway, E.C. (1998). Functional interactions of an upstream enhancer of the mouse glycoprotein hormone alpha-subunit gene with proximal promoter sequences. Mol. Cell. Endocrinol. *142*, 141–152.

Würtele, H., and Chartrand, P. (2006). Genome-wide scanning of HoxB1-associated loci in mouse ES cells using an open-ended Chromosome Conformation Capture methodology. Chromosom. Res. *14*, 477–495.

Xi, H., Shulha, H.P., Lin, J.M., Vales, T.R., Fu, Y., Bodine, D.M., McKay, R.D.G., Chenoweth, J.G., Tesar, P.J., Furey, T.S., et al. (2007). Identification and characterization of cell type-specific and ubiquitous chromatin regulatory structures in the human genome. PLoS Genet. *3*, e136.

Xi, Q., Wang, Z., Zaromytidou, A.-I., Zhang, X.H.-F., Chow-Tsang, L.-F., Liu, J.X., Kim, H., Barlas, A., Manova-Todorova, K., Kaartinen, V., et al. (2011). A poised chromatin platform for TGF-β access to master regulators. Cell *147*, 1511–1524.

Xiang, J.-F., Yin, Q.-F., Chen, T., Zhang, Y., Zhang, X.-O., Wu, Z., Zhang, S., Wang, H.-B., Ge, J., Lu, X., et al. (2014). Human colorectal cancer-specific CCAT1-L lncRNA regulates long-range chromatin interactions at the MYC locus. Cell Res. *24*, 513–531.

Xie, R., Everett, L.J., Lim, H.W., Patel, N.A., Schug, J., Kroon, E., Kelly, O.G., Wang, A., D'Amour, K.A., Robins, A.J., et al. (2013). Dynamic chromatin remodeling mediated by polycomb proteins orchestrates pancreatic differentiation of human embryonic stem cells. Cell Stem Cell *12*, 224–237.

Xu, P., and Davis, R.J. (2010). c-Jun NH2-terminal kinase is required for lineage-specific differentiation but not stem cell self-renewal. Mol. Cell. Biol. *30*, 1329–1340.

Xu, Z., Wei, G., Chepelev, I., Zhao, K., and Felsenfeld, G. (2011). Mapping of INS promoter interactions reveals its role in long-range regulation of SYT8 transcription. Nat. Struct. Mol. Biol. *18*, 372–378.

Yáñez-Cuna, J.O., Dinh, H.Q., Kvon, E.Z., Shlyueva, D., and Stark, A. (2012). Uncovering cis-regulatory sequence requirements for context-specific transcription factor binding. Genome Res. *22*, 2018–2030.

Yáñez-Cuna, J.O., Arnold, C.D., Stampfel, G., Boryń, L.M., Gerlach, D., Rath, M., and Stark, A. (2014). Dissection of thousands of cell type-specific enhancers identifies dinucleotide repeat motifs as general enhancer features. Genome Res. *24*, 1147–1156.

Yang, S.-K., Hong, M., Zhao, W., Jung, Y., Tayebi, N., Ye, B.D., Kim, K.-J., Park, S.H., Lee, I., Shin, H.D., et al. (2013). Genome-Wide Association Study of Ulcerative Colitis in Koreans Suggests Extensive Overlapping of Genetic Susceptibility With Caucasians. Inflamm. Bowel Dis. *19*.

Yun, W.J., Kim, Y.W., Kang, Y., Lee, J., Dean, A., and Kim, A. (2014). The hematopoietic regulator TAL1 is required for chromatin looping between the β-globin LCR and human γ-globin genes to activate transcription. Nucleic Acids Res. *42*, 4283–4293.

Zaret, K.S., and Carroll, J.S. (2011). Pioneer transcription factors: establishing competence for gene expression. Genes Dev. *25*, 2227–2241.

Zhang, F., Tanasa, B., Merkurjev, D., Lin, C., Song, X., Li, W., Tan, Y., Liu, Z., Zhang, J., Ohgi, K. a., et al. (2015). Enhancer-bound LDB1 regulates a corticotrope promoter-pausing repression program. Proc. Natl. Acad. Sci. *112*, 1380–1385.

Zhang, H., Jiao, W., Sun, L., Fan, J., Chen, M., Wang, H., Xu, X., Shen, A., Li, T., Niu, B., et al. (2013a). Intrachromosomal looping is required for activation of endogenous pluripotency genes during reprogramming. Cell Stem Cell *13*, 30–35.

Zhang, J., Poh, H.M., Peh, S.Q., Sia, Y.Y., Li, G., Mulawadi, F.H., Goh, Y., Fullwood, M.J., Sung, W., Ruan, X., et al. (2012a). ChIA-PET analysis of transcriptional chromatin interactions. Methods *58*, 289–299.

Zhang, X., Cowper-Sal-lari, R., Bailey, S.D., Moore, J.H., and Lupien, M. (2012b). Integrative functional genomics identifies an enhancer looping to the SOX9 gene disrupted by the 17q24.3 prostate cancer risk locus. Genome Res. *22*, 1437–1446.

Zhang, Y., McCord, R.P., Ho, Y.-J., Lajoie, B.R., Hildebrand, D.G., Simon, A.C., Becker, M.S., Alt, F.W., and Dekker, J. (2012c). Spatial organization of the mouse genome and its role in recurrent chromosomal translocations. Cell *148*, 908–921.

Zhang, Y., Wong, C.-H., Birnbaum, R.Y., Li, G., Favaro, R., Ngan, C.Y., Lim, J., Tai, E., Poh, H.M., Wong, E., et al. (2013b). Chromatin connectivity maps reveal dynamic promoter-enhancer long-range associations. Nature *504*, 306–310.

Zhao, K., Hart, C.M., and Laemmli, U.K. (1995). Visualization of chromosomal domains with boundary element-associated factor BEAF-32. Cell *81*, 879–889.

Zhao, R., Watt, A.J., Li, J., Luebke-Wheeler, J., Morrisey, E.E., and Duncan, S.A. (2005). GATA6 Is Essential for Embryonic Development of the Liver but Dispensable for Early Heart Formation. Mol. Cell. Biol. *25* , 2622–2631.

Zhao, Z., Tavoosidana, G., Sjölinder, M., Göndör, A., Mariano, P., Wang, S., Kanduri, C., Lezcano, M., Sandhu, K.S., Singh, U., et al. (2006). Circular chromosome conformation capture (4C) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. Nat. Genet. *38*, 1341–1347.

Zhou, V.W., Goren, A., and Bernstein, B.E. (2011). Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. *12*, 7–18.

Zhou, Y., Kurukuti, S., Saffrey, P., Vukovic, M., Michie, A.M., Strogantsev, R., West, A.G., and Vetrie, D. (2013). Chromatin looping defines expression of TAL1, its flanking genes, and regulation in T-ALL. Blood *122*, 4199–4209.

Zhu, J., Adli, M., Zou, J.Y., Verstappen, G., Coyne, M., Zhang, X., Durham, T., Miri, M., Deshpande, V., Jager, P.L. De, et al. (2013). Resource Genome-wide Chromatin State Transitions Associated with Developmental and Environmental Cues. Cell *152*, 642–654.

Zimmerman, N.P., Vongsa, R. a, Faherty, S.L., Salzman, N.H., and Dwinell, M.B. (2011). Targeted intestinal epithelial deletion of the chemokine receptor CXCR4 reveals important roles for extracellular-regulated kinase-1/2 in restitution. Lab. Invest. *91*, 1040–1055.

Zorn, A.M., and Wells, J.M. (2007). Molecular basis of vertebrate endoderm development. Int. Rev. Cytol. *259*, 49–111.

Zorn, A.M., and Wells, J.M. (2009). Vertebrate endoderm development and organ formation. Annu. Rev. Cell Dev. Biol. *25*, 221–251.

Zorn, A.M., Butler, K., and Gurdon, J.B. (1999). Anterior endomesoderm specification in Xenopus by Wnt/beta-catenin and TGF-beta signalling pathways. Dev. Biol. *209*, 282–297.

## PUBLICATIONS DURING THE COURSE OF THIS STUDY

1. **Zhang J**, Lim B. Disruption of long-range chromatin interactions by the 13q12.13 ulcerative colitis risk locus. (Manuscript in preparation. Poster presented at Cell Symposia: Transcriptional Regulation in Development. July 2014, Chicago, IL)

2. **Zhang J**, Lim B. Genomic approaches for the identification of transcriptional regulatory elements. (Review manuscript in preparation)

3. Loh KM#, Ang LT#, **Zhang J***, Kumar V*, Ang J, Auyeong JQ, Lee KL, Choo SH, Lim CY, Nichane M, Tan J, Noghabi MS, Azzola L, Ng ES, Durruthy-Durruthy J, Sebastiano V, Poellinger L, Elefanty AG, Stanley EG, Chen Q, Prabhakar S, Weissman IL, Lim B. Efficient endoderm induction from human pluripotent stem cells by logically directing signals controlling lineage bifurcations. Cell Stem Cell. 2014 Feb 6;14(2):237-52.

4. **Zhang J**, Poh HM, Peh SQ, Sia YY, Li G, Mulawadi FH, Goh Y, Fullwood MJ, Sung WK, Ruan X, Wold B, Ruan Y. ChIA-PET Analysis of Transcriptional Chromatin Interactions. Methods. 2012;Nov;58(3):289-99

5. Goh Y, Fullwood MJ, Poh HM, Peh SQ, Ong CT, **Zhang J**, Ruan X, Ruan Y. Chromatin Interaction Analysis with Paired-End Tag Sequencing (ChIA-PET) for mapping chromatin interactions and understanding transcription regulation. J Vis Exp. 2012;Apr;30;(62)

6. Li G, Ruan X, Auerbach RK, Sandhu KS, Zheng M, Wang P, Poh HM, Goh Y, Lim J, **Zhang J**, Sim HS, Peh SQ, Mulawadi FH, Ong CT, Orlov YL, Hong S, Zhang Z, Landt S, Raha D, Euskirchen G, Wei CL, Ge W, Wang H, Davis C, Fisher-Aylor KI, Mortazavi A, Gerstein M, Gingeras T, Wold B, Sun Y, Fullwood MJ, Cheung E, Liu E, Sung WK, Snyder M, Ruan Y. Extensive promoter-centered chromatin interactions provide a topological basis for transcription regulation. Cell. 2012 Jan 20;148(1-2):84-98

## APPENDIX I - OLIGONUCLEOTIDE SEQUENCES USED IN THIS STUDY

qPCR primers for RNA-seq validation

| **Reverse transcription qPCR primers** | **Sequence** |
|---|---|
| PBGD-F (housekeeping) | GGAGCCATGTCTGGTAACGG |
| PBGD-R (housekeeping) | CCACGCGAATCACTCTCATCT |
| ISL1-F | AGATTATATCAGGTTGTACGGGATCA |
| ISL1-R | ACACAGCGGAAACACTCGAT |
| IRX3-F | CTCCGCACCTGCTGGGACTTC |
| IRX3-R | CTCCACTTCCAAGGCACTACAG |
| PAX9-F | TGGTTATGTTGCTGGACATGGGTG |
| PAX9-R | GGAAGCCGTGACAGAATGACTACCT |
| TBX1-F | CGGCTCCTACGACTATTGCCC |
| TBX1-R | GGAACGTATTCCTTGCTTGCCCT |
| HNF1B-F | AGGCCACAATCTCCTCTCAC |
| HNF1B-R | TTGCTGGGGATTATGGTGGGA |
| HNF4A-F | CATGGCCAAGATTGACAACCT |
| HNF4A-R | TTCCCATATGTTCCTGCATCAG |
| HOXA1-F | CGTGAGAAGGAGGGTCTCTTG |
| HOXA1-R | GTGGGAGGTAGTCAGAGTGTC |
| HOXB4-F | GTTCCCTCCATGCGAGGAATA |
| HOXB4-R | GCTGGGTAGGTAATCGCTCTG |
| HOXC5-F | GCAGAGCCCCAATATCCCTG |
| HOXC5-R | CCGATCCATAGTTCCCACAAGTT |
| HOXC6-F | F:ACCCCTGGATGCAGCGAATGAATTCG |
| HOXC6-R | GTTCCAGGGTCTGGTACCGCGAGTA |
| CDX2-F | F: GGGCTCTCTGAGAGGCAGGT |
| CDX2-R | CCTTTGCTCTGCGGTTCTG |
| PDX1-F | F: GCGTTGTTTGTGGCTGTTGCGCA |
| PDX1-R | AGCTTCCCCGCTGTGTGTGTTAGG |

qPCR primers for ChIP enrichment validation

| **ChIP-qPCR primers** | **Sequence** |
|---|---|
| NC1-F | TTCAAGTGACTCCCCTGTCTC |
| NC1-R | TTCAGCAGAATTACAAGAAACAAAAT |
| NC2-F | TTTTGCAAGATTAGTTGATAAGAGAGA |
| NC2-R | GCGTTGGTTGTGGGACTATT |
| WNT8B-F | CCTTCCACCTCTTGTGTGCT |
| WNT8B-R | CAGGGACAGGAAGGAGACAG |
| FOXP4-F | AAGATGAAATCAGCGCATCC |
| FOXP4-R | CCTCCCTTCATTTCCTCAGA |
| NODAL-F | GGCAGGGATCCCAGGTGAGGT |
| NODAL-R | GCAGGCTTGTCCCGGCAGAT |

| | |
|---|---|
| DPPA4-F | CATTCTCAGCACCCTCGGTT |
| DPPA4-R | TGGGGGCTAGAGGGAAATGG |
| GAPDH-F | GCCTCTGCGCCCTTGAGCTA |
| GAPDH-R | GATGCGGCCGTCTCTGGAAC |
| ACTB-F | GGGTGGGTCACTAGGGAGAGA |
| ACTB-R | GACTCCCCCAACACCACACT |
| EIF3B-F | GAAGCCACATGCACCCAATG |
| EIF3B-R | ACTCAACAGGCGATTGCTCA |
| BTF3-F | TATTCGCTCCGACAAGGTACAA |
| BTF3-R | CCGCTCCCGTCCTCCTA |
| PRDM14-F | ACCCCGTACAGAACGAAGTG |
| PRDM14-R | AAACCCTCCAACCAAGAAGG |
| SOX2-F | GCCCTGCAGTACAACTCCAT |
| SOX2-R | GACTTGACCACCGAACCCAT |
| RNAPII-F | AACGGCGAATTCCACAAC |
| RNAPII-R | CGCGTCTGCTAACGTAGTCC |
| RNAPII-neg-F | AGTCTGAGCTTTGTGGACAGC |
| RNAPII-neg-R | CCCTCCCAGTATACAGTCTTGC |

ChIA-PET linker and adapter sequences

| **Linkers** | **Sequence** |
|---|---|
| Linker A | GGCCGCGATATCTTATCCAAC |
| Linker B | GGCCGCGATATACATTCCAAC |
| **Adapters** | **Sequence** |
| Adapter A | CCATCTCATCCCTGCGTGTCCCATCTGTTCCCTCCCTGTCTCAGNN |
| Adapter B | CTGAGACACGCAACAGGGGATAGGCAAGGCACACAGGGGATAGG |
| **PCR primers** | **Sequence** |
| Primer A | AATGATACGGCGACCACCGAGATCTACACCCTATCCCCTGTGTGCCTTG |
| Primer B | CAAGCAGAAGACGGCATACGAGATCGGTCCATCTCATCCCTGCGTGTC |
| **Sequencing primers** | **Sequence** |
| Forward | AATGATACGGCGACCACCGAGAT |
| Reverse | CAAGCAGAAGACGGCATACGA |

PCR primers for enhancer validation in zebrafish

| **Zebrafish PCR primers** | **Sequence** |
|---|---|
| TBX1-F | CACCCCTCCGGGTGACCAAAATCA |
| TBX1-R | GGATTGTCCCTCCTAGGCCA |
| HNF1B-F | CACCACTTAGCAGATGCTGTCAACAC |
| HNF1B-R | CGGCAGGCCCATAGAGATTA |
| WNT5B-F | CACCTGGCATCTCGCATGTCCTTT |
| WNT5B-R | AGACGAGTGCAGTTCCTTGG |

## APPENDIX II – BAC PROBES USED FOR DNA-FISH

Intrachromosome (related to Fig 4.6C, Fig 5.4A and Fig 5.5B)

| Test mix | Control mix |
|---|---|
| RP11-766A9 + RP11-463D24 | RP11-766A9 + RP11-979M22 |
| RP11-191N14 + RP11-915D14 | RP11-191N14 + RP11-92F20 |
| RP11-80F13 + RP11-795I20 | RP11-80F13 + RP11-482K16 |
| RP11-136G6 + RP11-780D7 | RP11-136G6 + RP11-903H1 |

Interchromosome (related to Fig 5.4B)

| Test mix | Control mix |
|---|---|
| RP11-727F15 + RP11-143M10 | RP11-727F15 + RP11-563H6 |
| RP11-626F12 + RP11-556I13 | RP11-626F12 + RP11-563H6 |
| RP11-727F15 + RP11-419E4 | RP11-727F15 + RP11-563H6 |
| RP11-286L5 + RP11-107L14 | RP11-286L5 + RP11-563H6 |
| RP11-286L5 + RP11-419E4 | RP11-286L5 + RP11-563H6 |

Immunofluorescence-DNA FISH

| MG locus | Control locus |
|---|---|
| RP11-973N23 | RP11-699B7 |
| RP11-399J13 | RP11-699B7 |
| RP11-34B20 | RP11-699B7 |
| RP11-143M10 | RP11-699B7 |

## APPENDIX III – ULCERATIVE COLITIS SNPS FROM NHGRI GWAS CATALOG

108 TagSNPs for ulcerative colitis from the NHGRI GWAS catalog

| No. | Coordinates (chr:position) | SNP ID | Population |
|---|---|---|---|
| 1 | 13:27531267 | rs17085007 | ASN |
| 2 | 16:86009740 | rs16940186 | ASN |
| 3 | 1:20200990 | rs4654903 | ASN |
| 4 | 16:86014241 | rs16940202 | ASN |
| 5 | 1:2501338 | rs10797432 | EUR |
| 6 | 1:20171860 | rs6426833 | EUR |
| 7 | 1:200101920 | rs2816958 | EUR |
| 8 | 2:198881668 | rs1016883 | EUR |
| 9 | 2:199523122 | rs17229285 | EUR |
| 10 | 3:53062661 | rs9847710 | EUR |
| 11 | 4:103511114 | rs3774959 | EUR |
| 12 | 5:594083 | rs11739663 | EUR |
| 13 | 5:134443606 | rs254560 | EUR |
| 14 | 6:32612397 | rs6927022 | EUR |
| 15 | 7:2789880 | rs798502 | EUR |
| 16 | 7:27231762 | rs4722672 | EUR |
| 17 | 7:107480315 | rs4380874 | EUR |
| 18 | 7:128573967 | rs4728142 | EUR |
| 19 | 11:96023427 | rs483905 | EUR |
| 20 | 11:114386830 | rs561722 | EUR |
| 21 | 15:41563950 | rs28374715 | EUR |
| 22 | 16:30482494 | rs11150589 | EUR |
| 23 | 16:68591230 | rs1728785 | EUR |
| 24 | 17:70641698 | rs7210086 | EUR |
| 25 | 19:47123783 | rs1126510 | EUR |
| 26 | 20:33799280 | rs6088765 | EUR |
| 27 | 20:43065028 | rs6017342 | EUR |
| 28 | 1:67705958 | rs11209026 | EUR |
| 29 | 1:161479745 | rs1801274 | EUR |
| 30 | 1:206939904 | rs3024505 | EUR |
| 31 | 2:61204856 | rs7608910 | EUR |
| 32 | 2:241579108 | rs4676406 | EUR |
| 33 | 3:49719729 | rs9822268 | EUR |

| | | | |
|---|---|---|---|
| 34 | 4:123329362 | rs17388568 | EUR |
| 35 | 7:107492789 | rs4510766 | EUR |
| 36 | 10:101290301 | rs6584283 | EUR |
| 37 | 12:68500075 | rs7134599 | EUR |
| 38 | 16:68674788 | rs6499188 | EUR |
| 39 | 17:38040763 | rs2872507 | EUR |
| 40 | 21:40465534 | rs2836878 | EUR |
| 41 | 22:50435480 | rs5771069 | EUR |
| 42 | 1:2513216 | rs734999 | EUR |
| 43 | 1:8021973 | rs35675666 | EUR |
| 44 | 1:22698447 | rs7524102 | EUR |
| 45 | 2:102663628 | rs2310173 | EUR |
| 46 | 2:219010146 | rs11676348 | EUR |
| 47 | 5:10752315 | rs267939 | EUR |
| 48 | 5:35876274 | rs3194051 | EUR |
| 49 | 5:40410935 | rs6451493 | EUR |
| 50 | 5:158826792 | rs6871626 | EUR |
| 51 | 6:43795968 | rs943072 | EUR |
| 52 | 6:106522027 | rs6911490 | EUR |
| 53 | 6:138006504 | rs6920220 | EUR |
| 54 | 9:4981602 | rs10758669 | EUR |
| 55 | 9:117553249 | rs4246905 | EUR |
| 56 | 9:139266405 | rs10781499 | EUR |
| 57 | 10:35554054 | rs12261843 | EUR |
| 58 | 11:1874072 | rs907611 | EUR |
| 59 | 11:76299194 | rs2155219 | EUR |
| 60 | 13:27531267 | rs17085007 | EUR |
| 61 | 13:41013977 | rs941823 | EUR |
| 62 | 16:86014241 | rs16940202 | EUR |
| 63 | 20:62327582 | rs2297441 | EUR |
| 64 | 21:16817051 | rs1297265 | EUR |
| 65 | 21:45615023 | rs2838519 | EUR |
| 66 | 1:200877562 | rs7554511 | EUR |
| 67 | 4:180284814 | rs6811556 | EUR |
| 68 | 5:107834247 | rs4571457 | EUR |
| 69 | 7:18800413 | rs11764116 | EUR |
| 70 | 7:81857893 | rs929351 | EUR |

| | | | |
|---|---|---|---|
| 71 | 13:79550934 | rs7319358 | EUR |
| 72 | 14:29132877 | rs1956388 | EUR |
| 73 | 20:31718653 | rs6059101 | EUR |
| 74 | 6:32079567 | rs17207986 | EUR |
| 75 | 9:117605070 | rs11554257 | EUR |
| 76 | 7:98760504 | rs7809799 | EUR |
| 77 | 1:20227723 | rs4654925 | EUR |
| 78 | 1:206943968 | rs3024493 | EUR |
| 79 | 1:20140036 | rs1317209 | EUR |
| 80 | 1:67694202 | rs2201841 | EUR |
| 81 | 1:161472158 | rs10800309 | EUR |
| 82 | 2:61186829 | rs13003464 | EUR |
| 83 | 3:49721532 | rs3197999 | EUR |
| 84 | 5:583442 | rs4957048 | EUR |
| 85 | 7:107503441 | rs4598195 | EUR |
| 86 | 9:139266496 | rs4077515 | EUR |
| 87 | 10:101291593 | rs11190140 | EUR |
| 88 | 12:68504592 | rs1558744 | EUR |
| 89 | 17:38062196 | rs2305480 | EUR |
| 90 | 1:200935866 | rs11584383 | EUR |
| 91 | 2:200290359 | rs1992950 | EUR |
| 92 | 17:38051348 | rs8067378 | EUR |
| 93 | 21:16805220 | rs1736135 | EUR |
| 94 | 16:10975311 | rs4781011 | EUR |
| 95 | 1:161479745 | rs1801274 | ASN |
| 96 | 9:5213687 | rs10975003 | ASN |
| 97 | 7:107453103 | rs2108225 | ASN |
| 98 | 7:107495434 | rs886774 | EUR |
| 99 | 13:40505510 | rs9548988 | EUR |
| 100 | 3:49701983 | rs9858542 | EUR |
| 101 | 9:139269338 | rs10781500 | EUR |
| 102 | 1:20142866 | rs3806308 | EUR |
| 103 | 1:67725120 | rs10889677 | EUR |
| 104 | 12:68596661 | rs2870946 | EUR |
| 105 | 9:85311147 | rs668853 | EUR |
| 106 | 7:107479519 | rs4730273 | EUR |
| 107 | 7:107484437 | rs4730276 | EUR |

| | | | |
|---|---|---|---|
| 108 | 7:107580839 | rs2158836 | EUR |