Imperial College London

Computational Genomics of Developmental Gene Regulation

Author:
Malcolm Perry
London Institute for Medical Science
Imperial College London

Supervisors:
Professor Boris Lenhard
Professor Matthias Merkenschlager

Thesis Submitted in part fulfilment of the requirements for the degree of Doctor of Philosophy of Imperial College London

November 3, 2017

Declaration

I certify that the work presented in this thesis is my own, and that any work carried out by others has been properly acknowledged and appropriately referenced in the text.

Malcolm Perry

Abstract

The development of multicellular organisms requires the precise execution of complex transcriptional programs. The demands posed by development, coupled with the relatively late evolution of multicellularity, could have led to a separate mode of gene regulation for gene involved in, and regulated throughout development. I investigated the regulation of genes by enhancers using histone modifications coupled to gene expression, based on the observation that developmental genes are surrounded by dense clusters of conserved enhancers which act in concert. Genes regulated by enhancers are much more likely to be developmentally regulated genes, and many enhancers at each loci co-ordinate to direct transcription across multiple tissues.

CAGE-seq is a powerful tool for determining the structure of promoters. I analysed promoters in Amphioxus using CAGE-seq to determine if the diverse promoter architectures observed in vertebrates had ancestral origins. Promoters in amphioxus can be divided into developmental and housekeeping promoters, which each have characteristic patterns of dinucleotide enrichment. Housekeeping promoters in Amphioxus have a novel promoter architecture, and a contain a high frequency of bidirectional promoters, which represents the ancestral vertebrate state. This set of genes highlight the malleability of promoter architecture during evolution. I developed a package in R/Bioconductor 'heatmaps' to enable effective visualisation of this, and other, data.

Taken together, these results suggest a second mode of regulation in vertebrates governing the regulation of developmental genes.

Acknowledgements

First, I would to thank Boris for the opportunity to undertake this PhD, and all the advice, technical or otherwise, experience, Croatian sweets, and most importantly scientific ideas he has shared we me over the last four years.

I would also like to think all the past and present members of the Lenhard group, particularly those who were there at the start of my PhD. Thanks also goes to those who helped with the writing of this thesis, especially Nathan and Liz, without whom you might not be reading this. Outside of the lab, others at Imperial have contributed enourmously to my experience during the PhD: ICCC gave me an excuse to leave London on the weekends, and ICSO and ULSO gave me something to do when I didn't. Finally, I would like to thank Aïda for keeping me sane and giving me a reason to finish.

Contents

1 Introduction **15**

1.1 Regulation of Gene Expression 15

1.1.1 Genes and Cellular Identity 15

1.1.2 Regulation of Gene Expression 15

1.1.3 Developmental Gene Regulation 16

1.2 Identification of Transcription Start Sites 18

1.2.1 CAGE 18

1.2.2 Pervasive Transcription 19

1.3 Promoter Architecture 19

1.3.1 Sequence Analysis 19

1.3.2 Core Promoter Motifs 21

1.3.3 CpG Islands and Dinucleotide Frequency 23

1.3.4 Promoter Types 23

1.3.5 Bidirectional Promoters 24

1.3.6 Nucleosome Positioning 25

1.3.7 Enhancers and 3D Genome Organisation 26

1.4 Regulation by Long-range Enhancers 27

1.4.1 Identification of Enhancers 27

1.4.2 Regulatory Landscapes 28

1.4.3 Enhancers and Disease 29

1.5 Genomic Regulatory Blocks 31

1.6 Aims of the thesis 33

2 Detecting Long-range Regulation **34**
2.1 Introduction . . . 34
2.2 Results . . . 38
2.2.1 Enhancers Act in Concert Across Broad Genomic Regions . . . 38
2.2.2 Predicting Target Genes . . . 44
2.2.3 GO Enrichment . . . 45
2.2.4 Enhancers can affect multiple genes in a TAD . . . 47
2.2.5 TAD boundaries restrict the effects of enhancers . . . 51
2.2.6 Genes under long-range regulation are enriched for long-range contacts . . . 53
2.3 Discussion . . . 55
2.4 Data . . . 60
2.5 Methods . . . 61
2.5.1 Modelling Enhancer-Promoter Interactions . . . 61
2.5.2 Measuring Enhancer Contribution . . . 65

3 The heatmaps.R Package for R/Bioconductor **69**
3.1 Introduction . . . 69
3.1.1 Existing Work . . . 70
3.1.2 Requirements . . . 71
3.2 Results . . . 72
3.2.1 Structure of the heatmaps.R packages . . . 72
3.2.2 Workflow . . . 73
3.2.3 Visualising Sequence Data . . . 75
3.2.4 Visualising Functional Genomics Data . . . 76
3.2.5 Clustering . . . 79
3.2.6 Performance . . . 80
3.2.7 Smoothing . . . 80
3.2.8 Plotting Options . . . 81
3.2.9 Multi-panel Plots . . . 82
3.3 Discussion . . . 83
3.4 Methods . . . 85

3.4.1 Zebrafish Promoters 85
3.4.2 Pattern Matching 86

4 Amphioxus Promoterome **87**
4.1 Declaration 87
4.2 Introduction 88
4.3 Results 89
4.3.1 Data 89
4.3.2 CAGE Data Processing 89
4.3.3 Categorising Promoters by Expression Profile 92
4.3.4 Promoter Analysis 92
4.3.5 De novo Motif Analysis 96
4.3.6 Bidirectional Promoters 100
4.3.7 Bidirectional Promoters in other species 101
4.3.8 Estimating the the rate of loss of bidirectional promoters 103
4.3.9 Promoter Evolution 108
4.4 Discussion 110
4.5 Methods 112
4.5.1 CAGE-seq 112
4.5.2 CAGE alignment 112
4.5.3 CAGE Tag Clustering 113
4.5.4 Expression clustering 114
4.5.5 Feature enrichment and visualisation 114
4.5.6 ATAC-seq Data 115
4.5.7 Gene Annotation and GO Enrichment 115

5 Discussion **116**
5.1 Developmental Promoters in Amphioxus are Defined at the Sequence Level 117
5.2 Developmentally regulated genes in humans respond preferentially to enhancers 117
5.3 Revisiting the GRB model of gene regulation 119
5.4 Promoters have characteristic features which change over time 120

5.5 Understanding biological sequence 121

List of Figures

1.1 Control of Gene Regulation by *cis*-regulatory elements 17
1.2 The GRB Model for Gene Regulation 30

2.1 Enhancers at the FZD7 locus 40
2.2 Enhancers at the IKZF1/GRB10 locus 41
2.3 Enhancers at the IRX3/5 Locus 42
2.4 Enhancers at the RUNX2 Locus 43
2.5 P-values for long-range regulation and the distribution of long-range regulated genes within TADs. 45
2.6 Expression patterns of genes involved in heart development . . 49
2.7 The distribution of long-range regulated genes within TADs. . 50
2.8 Average contributions of enhancers inside and outside of TADs 52
2.9 Contributions of enhancers by distance to their target gene . . 53
2.10 Significant chromatin contacts with enhancers for LRR and Non-LRR genes 54
2.11 Model Output for the KLK8 gene. 63
2.12 Model Output for the RNF180 gene. 64

3.1 The structure of a heatmap. 73
3.2 The workflow for creating a heatmap. 74
3.3 Sequence Features of Zebrafish Promoters 75
3.4 ChIP-seq and ATAC-seq at Zebrafish Promoters 77
3.5 Clustering phased ChIP-seq signal at zebrafish promoters . . . 78
3.6 Performance of heatmap plotting and smoothing 79
3.7 Shifting promoter in zebrafish 84

4.1 CAGE tag clusters in Amphioxus 90
4.2 SOM of gene expression by CAGE 91
4.3 Properties of expression-clusters amphioxus promoters I 93
4.4 Properties of expression-clusters amphioxus promoters II . . . 95
4.5 Mononucleotide frequency of ubiquitously-expressed promoters and IQ width of YY1-containing promoters 96
4.6 YY1 motifs affect Interquantile width in Ubiquitous Promoters 98
4.7 Distribution bidirectional promoters widths in amphioxus . . . 99
4.8 Dinucleotide density and nucleosome occupancy in amphioxus bidirectional promoters 101
4.9 Schematic of bidirectional promoter rediploidisation 105
4.10 Estimating the rediploidisaton of bidirectional promoter genes 107
4.11 Dinucleotide plots and nucleosome positioning at bidirectional promoters in other species. 109

List of Tables

2.1 GO BP Enrichment for genes under long-range regulation . . . 46
2.2 GO MF Enrichment for genes under long-range regulation . . 47
2.3 GO CC Enrichment for genes under long-range regulation . . 48
2.4 TADs containing many genes under long-range regulation . . . 51
2.5 Cell types used in the experiment for RNA-seq and Histone Acetylation data 67
2.6 Summary of data sources used in the experiment 68

3.1 Example options available for heatmaps 82

4.1 GO BP enrichment and exclusion in amphioxus bidirectional promoters 102
4.2 Biderectional promoter counts in other species 103
4.3 GO BP enrichment for bidirectional promoters in other species 104
4.4 Read count and alignment efficiency for CAGE samples 113

Chapter 1

Introduction

1.1 Regulation of Gene Expression

1.1.1 Genes and Cellular Identity

Multicellular organisms consist of many types of specialised cells, from simple colonial organisms such as slime moulds, to higher vertebrates with huge numbers of highly specialised structures. In all multicellular organisms, almost every cell contains a full copy of the genome within the nucleus, so cellular identity is maintained by expressing only those genes required by each cell. For development to unfold successfully, the genes determining cellular identity must be precisely expressed at the correct stages and at the correct location in the organism (Gurdon, 1992). Cells must integrate diverse signals from their environment to commit to one of many possible paths of differentiation, and this process is tightly regulated and generally irreversible.

1.1.2 Regulation of Gene Expression

Expression of most genes results from the binding of RNA Polymerase II (Pol. II) to the promoter region directly upstream of the main body of the transcript, and the subsequent transcription of DNA to RNA (Thomas and Chiang, 2006). RNA can itself can regulate the transcription of other genes or take part in catalytic processes (Nissen et al., 2000; Ernst and Morton,

2013; Bartel, 2004), but many genes require translation to form proteins, which govern the majority of cellular processes.

Proteins influencing transcription at gene promoters are known as transcription factors (TFs) (Maston et al., 2006). These include the general transcription factors which form the pre-initiation complex alongside Pol. II and are required for general transcription (Kadonaga, 2004), but the term more often refers to cell-type specific DNA-binding proteins, which play a major role in controlling cell-type specific expression. Transcription factors frequently bind directly to *proximal* gene promoter regions, as distinct from the *core* promoter region which binds Pol. II. Transcription factors binding at the promoter can also direct the repression of particular genes (Kadonaga, 2004; Gaston and Jayaraman, 2003; Perissi et al., 2010).

Enhancers are short, *cis*-acting DNA sequences that promote transcription of nearby genes independent of their orientation and position (Blackwood and Kadonaga, 1998). Identified more than thirty years ago, they are generally located in introns or intergenic space. Due to gaps displacement of enhancers from promoters, which can be over a megabase (Lettice et al., 2003), enhancers must form loops of chromatin in order to contact their target genes (Fraser et al., 2015). Elements which block this loop formation are called insulators (Gaszner and Felsenfeld, 2006), although recent evidence shows that canonical insulators themselves take part in looping interactions, forming a complex overall picture of 3D genome organisation.

1.1.3 Developmental Gene Regulation

Genes required for cellular homeostasis, such as energy metabolism and transcription of RNA, are expressed by all cells in an organism. Expression of the genes that control cellular identity is mediated by the activation of cell-type specific transcription factors (Spitz and Furlong, 2012). These are frequently downstream of signalling pathways which transduce signals from ligands binding to the cell surface. Cell-cell communication is vitally important in development for the formation of intricate patterns (Hafen et al., 1984) based on relatively simple principles of regulation (Ilsley et al., 2013).

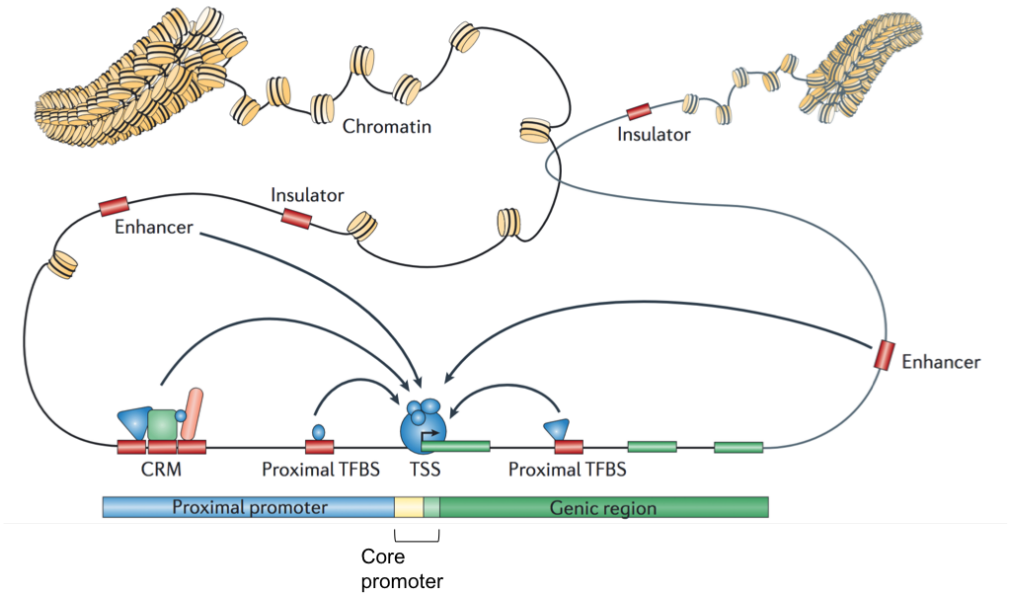

Figure 1.1: An overview of gene regulation by *cis*-regulatory elements. Transcription factors bind at both proximal and distal elements to influence transcription from the core promoter. 'CRM' refers to a *cis*-regulatory module, a set of closely linked TFBSs. This figure is adapted from Lenhard et al. (2012)

Cell-type specific proteins may have specific roles for cellular identity, such as myosin in muscle cells, or may themselves be transcription factors or signalling molecules expressed to communicate with nearby cells.

Enhancers play an important role in orchestrating precise spatio-temporal gene expression (Long et al., 2016), and in the thirty-plus years since their discovery, enhancers have been discovered in all metazoa studied (Sebé-Pedrós et al., 2017). In fact, the evolution of enhancers may underlie the huge increase in the complexity of body plans observed in the metazoan lineage (Sebé-Pedrós et al., 2017).

1.2 Identification of Transcription Start Sites

1.2.1 CAGE

Cap Analysis of Gene Expression (CAGE) (Shiraki et al., 2003; Kodzius et al., 2006) is a genome-wide method for identifying transcript abundance. This is achieved by first performing "cap-trapping" (Carninci et al., 1996): the 7-methylguanosine cap of mature mRNAs is biotinylated, following reverse transcription of the RNA, and the resulting complex is treated with RNAse I to degrade RNA not protected by synthesised cDNAs and isolated using streptavidin-coated beads. A linker sequence is added to the 5' end of the transcript before second strand synthesis, and, utilising a restriction site in the linker sequence, a short fragment of the 5' end of the transcript is cleaved. These fragments are then sequenced using next-generation sequencing technology, giving base-pair resolution of the start of each mRNA after they are mapped to the reference genome. CAGE also provides quantification of the number of transcripts initiated at each base pair, or CAGE-defined transcription start site (CTSS).

Gene promoters typically initiate transcription over a range of base pairs, rather than from a single nucleotide position. In order to estimate transcript abundance, nearby tags at each locus are clustered to produce *tag clusters* (TCs), effectively recapitulating the promoter for a single transcript. Each of these clusters also has a *dominant TSS*, the single base pair which has the highest number of initiation events (as determined by tags mapping to it). This is useful as it provides a single point for use in analyses which depend on spatial features of promoters. The width of these clusters is commonly estimated after first removing spatial outliers; typically, the first and last 10% of transcripts in the cluster, which would be referred to as the 10-90 Interquantile (IQ) width. In order to combine clusters across samples, an additional clustering procedure is performed to combine TCs into *consensus clusters*.

1.2.2 Pervasive Transcription

Modern high-throughput technology has facilitated genome-wide identification of promoters, replacing older methods which relied on identifying the transcription start sites (TSSs) of known genes. These have produced extensive maps of promoters in metazoa (for example, FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al. (2014)) which have led to increased study of promoters, and given insight into general properties of promoters. However, these have identified many more transcription start sites than can be accounted for by annotated genes (Carninci et al., 2006).

Many regions of the genome are actively transcribed, without any known biological purpose (ENCODE Project Consortium et al., 2007). In fact, some researchers contend that no specific sequence is needed, but simply that any open chromatin will be transcribed to some degree (Young et al., 2016). In this case, any sequence made accessible by transcription factors binding and opening up chromatin would classify as a promoter. Many of these transcripts are unstable and are rapidly degraded, making detection difficult. Active enhancers frequently act as sites for transcriptional initiation, and that bidirectional transcription at enhancers may be a hallmark of active regulatory elements (Andersson et al., 2015b).

1.3 Promoter Architecture

The exact, minimal components necessary to form promoters in Metazoa are not known. However, large-scale surveys of gene promoters using techniques such as CAGE have identified many features of gene promoters that have been robustly observed across many species. The term "Promoter Architecture" is used to refer to these observations.

1.3.1 Sequence Analysis

Mapping the precise regions which make up gene promoters allows us to investigate properties of promoters at the level of DNA sequence. There are

two complementary approaches to studying promoter sequence, and understanding biological sequences more generally.

A position weight matrix (PWM) represents a set of biological sequences probabilistically (Stormo et al., 1982; Stormo, 2000). The set of sequences is transformed into a position frequency matrix, which is simply a count of how many times each DNA base (**A**, **C**, **G** or **T**) appears at each position in the matrix. This is then transformed in a position probability matrix, representing the probability of finding each base at each position. This is then log-transformed to create the position weight matrix. PWM matches are scored by summing the log-probability of each observed base across the width of the PWM, and this score approximates a normal distribution (Claverie and Audic, 1996) allowing for statistical analysis of binding sites. Bases which are unrepresented in the original set of sequences can cause problems for this model, which would give a match a probability of zero and make the log-probability score incalculable, so psuedocounts can be added to the position frequency matrix to correct for this (Nishida et al., 2009). Motifs can be visualised as sequence logos (Schneider and Stephens, 1990) based on the information content at each position in the PWM.

PWMs are frequently used to represent the binding motif of particular transcription factors. Curated databases of PWMs, both free (Sandelin et al., 2004a; Mathelier et al., 2016) and commercial (Matys et al., 2003), allow the downloading of known PWMs in order to identify potential binding sites for specific factors. Another approach is to learn PWMs from collections of sequences (Bailey et al., 2009), and then compare these with known motifs, although this is computationally intensive, limiting the number of sequences that can be used, and does not guarantee finding enriched motifs even when they are known to be present. These methods will not always correctly identify binding sites, partly due to technical difficulties in identifying motifs (Simcha et al., 2012) and partly because motifs themselves are limited in the complexity of features they can represent. Scoring sequence matches using PWMs implicitly assumes each base in the motif is independent. Recent approaches have sought to rectify this problems (Mathelier and Wasserman, 2013; Siebert and Söding, 2016) but are yet to gain widespread acceptance,

and are harder to interpret. Additionally, PWMs do not take into account the genomic context of matches, such as DNA shape or chromatin accessibility.

Another approach to studying biological sequences is to look at the distribution of short, exact sequence matches, known as k-mers. There are aspects of promoter architecture that are not well captured by motifs; in particular, PWM discovery methods focus on motifs that occur more frequently than expected across whole sequences, rather than motifs that are enriched in certain positions. Many short DNA sequences, such as dinucleotides, have spatial patterns around promoters. **CG** dinculeotides are enriched around mammalian promoters (Bird, 1986) in a phenomenon known as 'CpG islands' (Gardiner-Garden and Frommer, 1987). Periodic enrichment in **WW** (**A** or **T**) and **SS** (**C** or **G**) dinculeotides contributes to the positioning of nucleosomes ((Segal et al., 2006), see section Nucleosome Positioning). Quantification of spatial patterns is difficult, and consequently the focus of many statistical methods has been on identifying enriched regions of k-mers(Gardiner-Garden and Frommer, 1987), an approach that has been refined using hidden Markov models (Wu et al., 2010) which capture spatial patterns accurately. An alternative statistical approach is to ignore the spatial dimension of k-mer frequency and focus on k-mer occurrence, allowing the use of powerful classifiers such as Support Vector Machines (Lee et al., 2011; Ghandi et al., 2016). Direct visualisation of k-mer patterns is also an effective method for analysing promoter sequence (Haberle, 2015).

1.3.2 Core Promoter Motifs

The core promoter is the DNA region immediately surrounding the TSS, where the pre-initiation complex is assembled. A number of motifs were identified at promoters before the advent of high-throughput technologies, such as the TATA-box (Lifton et al., 1978). The TATA-box is positioned at -32 to -29bp from the dominant transcription start site (Ponjavic et al., 2006), where it recruits TATA-binding protein (TBP) to the pre-initiation complex (Kadonaga, 2004). This results in 'sharp' initiation of transcription with the location of the TATA box as the chief determinant of TSS

usage (Ponjavic et al., 2006). TATA-box promoters are associated with tissue specific genes (Plessy et al., 2012), and mediate regulatory interactions in *Drosophila* (Butler and Kadonaga, 2001). The old 'textbook' view of TATA-box motifs accompanying most Pol. II. mediated transcription (Kadonaga, 2004) is flawed, however, since by no means all Poll. II. promoters contain a TATA-box motif.

The INR element is present in most human genes (Yang et al., 2007) and directly overlaps the TSS, and is also present in a modified form at the same location in Drosophila (Kutach and Kadonaga, 2000). In humans, INR consists of a pyrimidine base (**C** or **T**) followed by a purine (**A** or G); in Drosophila, the INR consensus is TCA[G/T]TY. More recently, the TCT was found to mark the transcription start sites ribosomal protein genes (Parry et al., 2010), replacing the INR element at these promoters, which may represent a 'high performance' TSS needed for highly transcribed genes (Lenhard et al., 2012).

The downstream promoter element (DPE) was discovered in TATA-less promoters in *Drosophila* (Burke and Kadonaga, 1997), but does not appear conserved to be present in vertebrate genomes. It is located at +28 to +32 relative to the INR element. High-throughput analysis of *Drosophila* cDNAs identified several other core promoter elements in Drosophila (Ohler et al., 2002), bringing the total count of known *Drosophila* motifs to 10. Despite numerous motifs, however, classifiers trained at the time could only detect promoters with 50% sensitivity, underscoring the fact the motifs alone do explain promoter function. Only the TATA, INR, TCT and BRE elements are shared between *Drosophila* and vertebrates (Lenhard et al., 2012).

Other TFs are frequently bound at promoters, but are thought to be cell-type specific factors directing expression, rather than core parts of the transcriptional machinery. The YY1 (Ying-yang 1) protein a universally expressed transcription factor, named after its dual role in both repressing and activating transcription (Shi et al., 1997). It has been implicated in diverse processes including polycomb-mediated repression (Wilkinson et al., 2006), neural induction (Satijn et al., 2001) and B-cell differentiation (Kleiman et al., 2016). It has also been shown to restrict downstream initiation of

transcription in human LINE elements (Kleiman et al., 2016), hinting a role in transcription beyond simple regulation.

1.3.3 CpG Islands and Dinucleotide Frequency

The first genome-wide studies of promoters revealed two classes of promoter, determined by the relative density of CpG sites (Carninci et al., 2006). High-CG promoters were distinguished by their overlap with CpG Islands. CpG sites are generally depleted in vertebrate genomes, possibly due to greater tendency to mutate: CpG islands are regions of the genome with a markedly greater concentration of CpG sites. Low-CG promoters were associated with narrower patterns of transcription initiation, however later studies showed that high CG content was not a requirement for broad promoters across all Metazoa, and that equally low-CG promoters can have broad initiation (Rach et al., 2011).

CpG sites are important sites of methylation in mammalian genomes, and 70-80% of CpGs are methylated genome-wide (Jabbari and Bernardi, 2004). DNA methylation is a key repressive epigenetic mark in vertebrates. Large domains of demethylated CpGs are associated with developmental genes (Jeong et al., 2014). Other dinucleotide patterns are enriched around promoters, although not as significantly as CGs, and may be involved in biological processes such as nucleosome positioning.

1.3.4 Promoter Types

The promoter features listed above do not occur independently, but segregate into recurring patterns which form the basis for promoter 'types', and these types are further associated with specific expression signatures. Lenhard et al. (2012) identified three main types of Pol. II promoter in vertebrates, with analogous types in Drosophila which do not always share the same features.

Type I promoters are tissue specific, but not developmentally regulated, representing genes stably expressed in terminally differentiated tissues. The

have a sharp TSS but disordered nucleosomes, and are enriched in TATA-box motifs (Yamashita et al., 2005; Carninci et al., 2006; Rach et al., 2011). Type II promoters are broadly expressed through the life cycle of an organism, including most 'housekeeping' genes, and have a broad TSS but well-positioned nucleosomes. They are enriched in CpGs but depleted in TATA-box motifs. Type III promoters are developmentally regulated, often by the polycomb pathway, and have large CpG islands extending into the gene body (Engström et al., 2007; Akalin et al., 2009). It is worth noting that is not the individual tissues in which a gene is expressed that govern promoter architecture, but the breadth of tissues. Minor promoter types include the aforementioned TCT promoters driving highly expressed genes essential for translation, and non-Pol. II promoters transcribing functional RNAs.

Overlapping promoters provide a particularly striking example of promoter architecture affecting gene regulation in the developing zebrafish embryo. (Haberle et al., 2014). One set of promoters, with a TATA-box, drives expression in oocytes, while an independent promoter, marked by a broad CG tract, drives expression after the mid-blastula transition, the point at which developing embryos begin to express their own genes. This second class of promoter aligns with upstream and downstream bands of TA enrichment, with a CG-rich core promoter.

1.3.5 Bidirectional Promoters

Promoters occur in closely-spaced, back-to-back arrangements much more frequently than would be expected by chance (Trinklein et al., 2004). However, these genes are not generally co-regulated (Engström et al., 2006), and so gene regulation does not explain why such an arrangement would be favoured. There is an ongoing debate concerning the intrinsic directionality of gene promoters: Duttke et al. (2015) argue that human promoters are intrinsically directional, i.e. generate transcripts primarily in a single direction along the DNA, while others have argued the exact opposite (Andersson et al., 2015a). This debate may partly focus on terminology (Andersson

et al., 2015a), as it is generally accepted that most promoters generate short, upstream transcripts (PROMPTs) that are rapidly degraded (Xu et al., 2009; Wei et al., 2011), and the directionality of promoters observed in functional genomics data is mediated chiefly by RNA stability. These antisense transcripts may then act as an opportunity for the emergence of new genes during evolution (Gotea et al., 2013).

1.3.6 Nucleosome Positioning

Due to the quantity of DNA in eukaryotic cells, it must be efficiently packaged to fit inside the nucleus. The most basic unit of the chromatin fibre is the nucleosome, a histone octamer which wraps 147bp of DNA in two loops (Luger et al., 1997). These assemble into higher-order structures, aided by the linker Histone H1, achieving compaction of 30-40 times over naked DNA (Widom, 1989). This compacting process can occlude DNA-binding proteins from contacting the DNA, making nucleosomes the primary factor regulating DNA accessibility (Bassett et al., 2009).

Nucleosome positioning is commonly assayed by first digesting the DNA using micrococcal nuclease, then sequencing the resulting fragments (MNase-seq, (Cole et al., 2012)). This preferentially digests the linker regions between nucleosomes, so the position of nucleosome centres can be recreated from the position of the fragment centres. Recent studies have suggested that treatment by varying concentrations of MNase reveals different patterns of nucleosome positioning (Chereji et al., 2016), indicating variation in the sensitivity to MNase digestion between nearby nucleosomes.

NucleoATAC (Schep et al., 2015), a novel way of analysing ATAC-seq (Buenrostro et al., 2015) data, has recently emerged as an alternative method for identifying the positions of nucleosomes genome wide. Transposons are added to cells and integrate into positions with accessible chromatin. Specific primers are used to sequence out from the insertions, generating short fragments of DNA which are sequenced. Both the position and the length of fragments are taken into account using ATAC-seq, which makes its estimates of nucleosome position robust.

Nucleosome positioning on the DNA is driven by a number of factors. Several studies (Segal et al., 2006; Kaplan et al., 2009; Locke et al., 2010) have identified DNA sequences which preferentially bind nucleosomes. The strongest signal is produced by bendable **WW** (**T** or **A**) dinucleotides with ∼10bp periodicity which directly contact the histones with each twist of the double helix (Segal et al., 2006). However, active processes also control the position of nucleosomes, leading to differences between *in vivo* and *in vitro* nucleosome positioning (Kaplan et al., 2009). These include ATP-dependent remodellers, transcription factors and RNA Polymerase II (Struhl and Segal, 2013).

The correct position of nucleosomes at promoters is required for gene expression. The first nucleosome downstream of the TSS at a promoter, the '+1' nucleosome, is generally positioned precisely (Rach et al., 2011), with following +2 and +3 nucleosomes showing increasing disorder independent of sequence (Rube and Song, 2014). There are indications that the relative order of nucleosomes around the TSS depends on the architecture of the promoter (Rach et al., 2011; Nozaki et al., 2011; Lenhard et al., 2012).

1.3.7 Enhancers and 3D Genome Organisation

It is widely accepted that enhancers physically contact the promoters of genes whose transcription they control. Given the large distances separating some pairs of enhancers and promoters, some form of chromosome looping seems to be the only plausible explanation for how enhancers can exert a regulatory effect on distal genes. However, physical contact between enhancers and promoters does not necessarily lead to activation (Ghavi-Helm et al., 2014).

Ligation based methods assay chromatin contact frequencies by cross-linking DNA, shearing the DNA using restriction enzymes and analysing the resulting fragments for pairs of sequences which show a greater frequency of ligation than expected by random chance, and are therefore closer together in space (de Wit and de Laat, 2012). 3C (Dekker et al., 2002) assayed the interaction frequency of two predetermined fragments. Ligation-based methods have since been extended to one-vs-many (4C, Zhao et al. (2006)),

many-vs-many (5C, Dostie et al. (2006)), and combined with antibody purification (ChIA-PET, Li et al. (2010)). Ligation-based methods have been used successfully to link promoters and enhancers in disease contexts (Lettice et al., 2003; Ragvin et al., 2010) and investigate the regulatory effects of promoter-enhancer loops (Guo et al., 2012; Ghavi-Helm et al., 2014).

Hi-C (van Berkum et al., 2010) is a ligation based method which identifies interactions genome wide. A key discovery of Hi-C is the existence of topologically associating domains (TADs), regions of the genome which represent discrete compartments of self-interacting chromatin (Dixon et al., 2012). TADs have been shown to restrict the action of enhancers, and disruption of TAD boundaries can lead to ectopic gene activation and developmental abnormalities (Lupiáñez et al., 2015).

The mechanisms underlying TAD formation are not yet fully understood. CTCF is known to be associated with loops in chromatin (Guo et al., 2012; Merkenschlager and Nora, 2016), but the 'contact domains' identified by (Rao et al., 2014) are not sufficiently numerous to explain the phenomenon of TADs genome-wide. CTCF is enriched at TAD borders (Dixon et al., 2012), but TADs appear to be more complex structures than simple loops, based on Hi-C data. It has been postulated that CTCF acts as a boundary element for loops formed by cohesin, which are then 'extruded' by motor proteins, either cohesin itself (Goloborodko et al., 2016) or by other factors, possibly RNA Pol. II (Sanborn et al., 2015a). It has been shown in *D. melanogaster* that transcription of the zygotic genome precedes the formation of TADs (Hug et al., 2017).

1.4 Regulation by Long-range Enhancers

1.4.1 Identification of Enhancers

Next-generation sequencing technologies have produced many methods for identifying enhancers genome-wide. Chromatin immunopreciptation followed by sequencing (ChIP-seq) identifies enhancers bound by known transcription factors. This identifies many sites which are not evolutionarily conserved,

even in closely related species (Villar et al., 2015), questioning the usefulness of purely sequence-based methods for identifying enhancers. It stands to reason that DNA must ultimately define the location of enhancers, but we may not fully understand the exact sequence determinants of enhancers, so we cannot identify conserved functionality without homologous sequence.

ChIP-seq has also been applied to studying histone modifications. Histones, which wrap most DNA in eukaryotic cells, have flexible tails which can be chemically modified, usually by the addition of an acetyl or methyl group. Several studies have reported mono-methylation of Histone H3 Lysine 4 (H3K4me1) and acetylation of lysine 27 (H3K27ac) as important for enhancer activation. Other marks have been identified as important for certain cell-type specific enhancers (Taylor et al., 2013), or contributing to enhancer 'priming' and 'poising' (Calo and Wysocka, 2013).

Self-transcribing enhancer assays allow the identification of enhancers based on their ability to activate transcription from a nearby upstream promoter, in *Drosophila melanogaster* (STARR-seq, Arnold et al. (2013) and in embryonic stem cells (FIREWACh, Murtha et al. (2014)). This requires cultured cells and is highly cell-type dependent, but has the advantage of measuring enhancer activity directly rather than by a proxy such as TF binding or histone modification. There is also evidence from STARR-seq for multiple classes of enhancer, which enhance developmental and housekeeping genes respectively (Zabidi et al., 2015).

1.4.2 Regulatory Landscapes

Enhancers can regulate multiple genes, and genes themselves are frequently controlled by many enhancers (Lower et al., 2009; Montavon et al., 2011). Understanding these complex loci has led to several models of gene regulation by enhancers. The human β-globin locus has been studied extensively as a model for gene activation by multiple enhancers (Levings et al., 2002). "Super-enhancers" have been identified as clusters of nearby enhancers with extremely high binding of mediator (Hnisz et al., 2013) that form around key developmental genes (Whyte et al., 2013). Similarly, "stretch enhancers",

which are defined as long histone acetylation peaks which are enriched for disease risk variants (Parker et al., 2013). It is not known whether these regions represent a distinct regulatory mode (Pott and Lieb, 2014), or whether the properties super- or stretch-enhancers are shared more generally by any closely spaced, dense clusters of enhancers.

Lorberbaum et al. (2016) identified a number of enhancers at the Ptch1/patched locus in both mouse and fly, which drive expression in independent tissues, creating a complex overall pattern of Ptch1 expression. This behaviour may underlie complex expression patterns in many genes, and explain why clusters of enhancers, such as super-enhancer, stretch enhancers or Genomic Regulatory Blocks (see below), surround important developmental genes, which are often expressed across many tissues.

1.4.3 Enhancers and Disease

Deletions or mutations within enhancers can have serious consequences for gene expression, and there are many known examples of Mendelian genetic disorders caused by mutated enhancers. Lettice et al. (2003) mapped separate inherited mutations in multiple families to a very long-range (1Mb) to an enhancer regulating SHH, in these cases leading to pre-axial polydactyly. Other studies have linked mutations in enhancers regulating SOX9 to Pierre Robin syndrome (Benko et al., 2009), and mutations in TBX5 enhancers to congenital heart disease (Smemo et al., 2012).

The recent increase in genome-wide association studies has led to the identification of a large number of genetic variants implicated in disease, many of which are found outside of coding genes (Maurano et al., 2012). Non-genic mutations may be in linkage disequilibrium with causal genic variants absent from single nucleotide polymorphism (SNP) panels, however it is likely a significant fraction of these variants represent causal mutations in regulatory elements, since many variants overlap with enhancer-associated chromatin features, such as DNase I sensitivity (Maurano et al., 2012). This has created difficulties in correctly assigning the variant to the misregulation of a particular gene or process. A variant strongly associated with obesity in humans

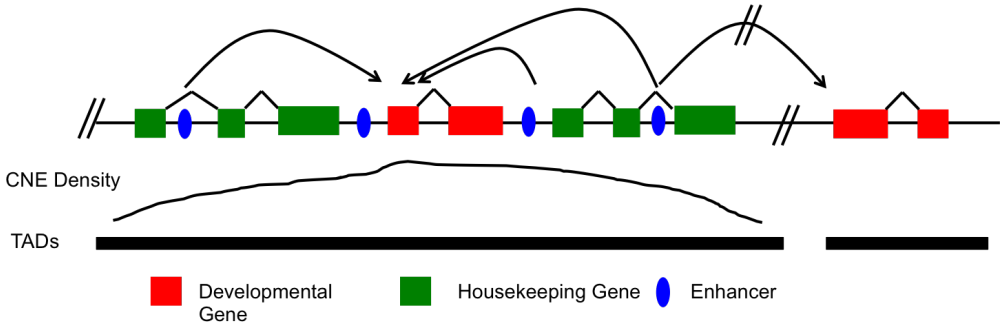

Figure 1.2: An overview of the GRB model of gene regulation. Enhancers target developmental genes, and do not regulate housekeeping genes. Regions of high non-coding conservation (GRBs) coincide with TADs, which restrict enhancers from regulating neighboring developmental genes

(Frayling et al., 2007) discovered in the first intron of the FTO gene, was initially thought to regulate FTO because of its spatial proximity. However, conserved enhancers in linkage disequilibrium with the lead SNP drive expression patterns similar to IRX3 (Ragvin et al., 2010). A later study used 4C to show long-range functional connections with IRX3 (Smemo et al., 2014). Weedon et al. (2014) investigated variants causing isolated pancreatic agenesis at the PTF1A locus by searching for overlaps with regulatory features, linking several novel variants to PTF1A function. Systematic approaches to studying enhancer mutations may represent an important method for understanding distal disease-causing variants (Miguel-Escalada et al., 2015).

Until recently, methods for studying chromosome conformation were limited to individual loci at high resolution. Novel methods have been developed based on Hi-C, adding an additional step to enrich for specific fragments prior to sequencing. These have substantially increased the resolution possible in genome-scale analyses. These include Capture-C (Platt et al., 2016) and promoter-capture Hi-C (Schoenfelder et al., 2015; Mifsud et al., 2015), which has been used to study the 'interactome' of risk loci (Jäger et al., 2015).

1.5 Genomic Regulatory Blocks

Conserved non-coding elements are an important feature of vertebrate and insect genomes (Bejerano, 2004; Sandelin et al., 2004b; Siepel et al., 2005). They were first identified as ultraconserved regions between mouse and humans, which were 200bp or more in length with perfect sequence identity (Bejerano, 2004). The definitions were relaxed and extended to more distant species, including invertebrates (Sandelin et al., 2004b; Siepel et al., 2005). Many of these elements act as developmental enhancers in experimental assays (Visel et al., 2007), but the enhancer activity of CNEs does not explain the full extent of their conservation (Harmston et al., 2013).

The elements do not occur randomly along the genome, but in large, syntenically conserved arrays (Kikuta et al., 2007; Akalin et al., 2009), termed "Genomic Regulatory Blocks", or GRBs. The location of these arrays, and CNEs in general, are strongly associated with important developmental genes (Sandelin et al., 2004b; Woolfe et al., 2005), which suggests that CNEs act as enhancers for these developmental genes. Other genes are present at the loci, but these are frequently housekeeping genes, and in many cases are lost from one copy of the GRB following genome duplication (Kikuta et al., 2007). GRBs are deeply conserved, even when the individual elements within them are not: high levels of non-coding conservation are observed around homologous genes in *D. melanogaster* and humans, even though very few conserved enhancers are identifiable between deuteresomes and protostomes (Clarke et al., 2012; Maeso et al., 2012). GRBs are also associated with TADs throughout evolution (Harmston et al., 2017).

Taken together, these observations lead to a model for developmental gene regulation known as the GRB Model. Developmental transcription factors require precise spatio-temporal regulation, and this is provided by the large number of enhancers at GRB loci. Each enhancer may only be active in very specific tissues, but the sum of these activities creates the complex pattern of expression required to correctly form the developing embryo. Nearby housekeeping genes are thought to be unresponsive to regulation by these enhancers. Ectopic activation of nearby genes is prevented by TAD formation

around GRBs.

1.6 Aims of the thesis

Multiple lines of evidence reviewed here suggest the regulation of developmental genes and tissue-specific genes differs from the regulation of ubiquitously expressed genes. In this thesis I investigate the role of both enhancers and promoters in the regulation of developmental genes.

In Chapter 2, I plan to investigate how enhancer activation affects gene expression by analysing functional genomics data, using both visualisation and statistical modelling or both. My main question is whether all genes are regulated by enhancers, or if only a subset are. I am also interested in how enhancer action is co-ordinated across genomic domains.

In Chapter 3, I plan to develop a new R/Bioconductor package, 'heatmaps', with the aim of facilitating the analysis of spatial patterns in genomic data, such as sequence features, histone acetylation and nucleosome occupancy.

In Chapter 4, I will analyse CAGE-seq data in the European Amphioxus, in order to test whether promoters that are regulated differently also show differences in their architecture. I will then compare this with CAGE-seq data in other organisms to see whether these patterns are stable throughout different lineages.

Chapter 2

Detecting Long-range Regulation

2.1 Introduction

Gene expression is controlled by the binding of transcription factors (TFs), which bind at gene promoters, or at regulatory elements which are brought into contact with promoters through chromosome looping (see Introduction). Enhancers can be located hundreds of kilobases, or more than even a megabase away from their target gene. In this chapter I will refer to this as 'long-range regulation' (LRR), to distinguish regulation by distal enhancers from regulation by TFs binding at or near the gene promoter. Long-range regulation poses a challenge in studying regulation, because it is not always possible to identify all enhancers for a given gene. Even in situations where enhancers are known, we cannot be sure which gene is being regulated, since enhancers frequently regulate genes at long distances and across other genes.

Methods for linking genes to enhancers include eQTL mapping, 3C-based technologies which assay physical interactions, and enhancer knockouts. Genetic evidence provides a gold standard, but does not provide genome-wide answers to the question. Direct genetic manipulation is still generally low-throughput, despite advances in CRISPR technology. eQTL studies require large cohort sizes, and consequently very many experiments, to have suffi-

cient power to detect associations, and they are also dependent on known SNPs located in regions of interest. Ultimately, we still lack methods for assigning enhancers to genes genome wide, and a solution to this problem would represent a significant breakthrough in regulatory genomics.

Interaction datasets provide a clear target for machine-learning tasks, allowing researchers to use complex models and avoid overfitting using cross-validation, although it has not been shown that physical interactions alone are sufficient for promoter activation. In fact, it has been shown that loops can form without gene activation (Ghavi-Helm et al., 2014), although in this case the regulatory interactions were true, and gene expression followed later in development. If the only mechanism for avoiding ectopic promoter-enhancer interactions is that loops can only be formed between the correct pairs, this raises a further question: what regulates these interactions? It is not implausible that each enhancer also contains binding sites for architectural proteins. Many architectural proteins are indeed present at the base of chromatin loops, including CTCF (Phillips-Cremins et al., 2013) and ZNF143 (Bailey et al., 2015), which implies that these chromatin looping is actively regulated. However, these proteins are not present at every enhancer, or every loop anchor, and the formation of many static loops in chromatin is not commensurate with current models for genome folding (Goloborodko et al., 2016). If interactions are unregulated and permissive, then another explanation is needed to explain enhancer-promoter specificity.

If a set of enhancers all regulate the same gene, they must all be able to physically contact the gene promoter. It follows that they are likely to be located in the same Topologically Associating Domain (TAD). TADs have been proposed to constrain enhancer-promoter actions to within each domain (Dixon et al., 2012), and disruption of TAD boundaries has been shown to disregulate nearby genes (Lupiáñez et al., 2015). Additionally, a subset of the strongest TADs coincides with of conserved non-coding elements (CNEs) (Harmston et al., 2017) which can act as regulatory elements, further strengthening the case that TADs represent important regulatory units.

Beginning with the ENCODE project (ENCODE Project Consortium et al., 2007), there have been several efforts to generate large datasets com-

prising ChIP-seq and RNA-seq in many cell types. This provides us with matched samples containing data which can be used to investigate both gene and enhancer activity. This has made statistical and machine-learning approaches possible, which generally seek to correlate epigenetic features to predict regulatory interactions.

Cismapper (O'Connor et al., 2016) directly correlates histone acetylation and gene expression. It uses the P-value of these correlations to compute the positive predictive value (PPV) for links between enhancers and promoters at various thresholds and distances, based on a 'gold standard' of Promoter Capture Hi-C (Mifsud et al., 2015) contacts. The authors note that the recall is low and that there is a significant tradeoff between PPV and FDR, but show a significant improvement over simple distance-based metrics, where the score is solely dependent on the distance between promoter and enhancers. Ernst et al. (Ernst et al., 2011) used a similar strategy, but additionally trained classifiers based on correlations of multiple marks with gene expression, and reported good agreement with known eQTLs.

Zhu et al. (Zhu et al., 2016) used a tensor-based algorithm to identify spatial associations between 16 histone modifications, DNase-seq, and RNA-seq. The authors identified associations after decomposing these 16 values into 'eigenloci', a form of dimensionality reduction, calculating significance through permutation tests. They reported that these associations were good predictors of interactions, based on Hi-C data, and further validated selected predictions using 3C.

Both Roy et al. (Roy et al., 2015) and He et al. (He et al., 2014) trained random forest classifiers, using interaction datasets as a gold standard. Features included epigenetic marks, correlations between marks and expression and additional features derived from the underlying sequence, such as co-occurring transcription factor motifs and synteny.

These methods all work to some degree, but both correlation- and interaction-based approaches have some drawbacks. One problem faced by many methods is multiple testing. There are many genes and many enhancers at most loci, therefore in all-vs-all comparisons, the number of tests scales as the product of these two numbers. This means that the threshold for signif-

icance is very difficult to reach, and with noisy data detecting any signal above background noise can be tricky.

Rationale

I aim to re-investigate the problem from the perspective that enhancers work in concert, across set domains, to regulate specific target genes. Correlations provide an alternative line of evidence to physical interactions: one which complements method measuring physical chromatin contacts, as neither method alone can provide conclusive, causal evidence of regulation. Testing enhancers acting together across broad genomic domains will help to reduce the multiple testing burden, thereby increasing the power considerably. There is also a strong biological rationale for this approach: in cases where a single gene has multiple enhancers and a complex expression pattern, we would not expect a single enhancer to correlate strongly with the overall expression pattern. Instead, each enhancer might drive expression in a single tissue, and combining these effects creates the overall pattern of gene expression, which is a mosaic of individual contributions from enhancers. I will look specifically at the co-ordination of enhancers across TADs, since TADs are thought to restrict the limits of enhancer action (see Introduction).

I will apply this novel approach to selected datasets to produce a robust list of genes under long-range regulation in these conditions. I also hope to explore general mechanisms of enhancer action. My goal is to link enhancers to their target promoters by correlating enhancer activity and gene expression.

Aims

1. Confirm that enhancers do in fact work together across broad domains.
2. Derive a statistical model for enhancer activation and gene expression.
3. Investigate the properties of genes identified as targets.
4. Contrast these results with 3D genome organisation data.

2.2 Results

A large panel of cell types with matching ChIP-seq and gene expression data is needed to investigate correlations between enhancer activity and gene activation. I used data from the Roadmap Epigenome Consortium (Roadmap Epigenomics Consortium et al., 2015). The Roadmap project generated epigenomic data, mostly in the form of ChIP-seq experiments, in over 100 human cell types. They also produced RNA-seq data in a subset of these cell types. I chose to use Histone H3, lysine 27 acetylation (H3K27ac) ChIP-seq data to measure enhancer activity, since this is reported in many studies to be a mark of active enhancers (Calo and Wysocka, 2013), and defined enhancers by calling peaks in this data (see Methods for details).

I validated the chosen enhancer set by comparison to the enhancers defined using ChromHMM (Ernst and Kellis, 2012), a multivariate hidden Markov model which uses a combination of epigenetic marks to define discrete epigenetic states. The enhancers I defined overlapped with at least one ChromHMM defined enhancer in 91.1% of cases. It is possible the minor discrepancy between the two is due to the fact that my defined set of enhancers is slightly more permissive, because it is defined only H3K27ac peaks, rather than a broad spectrum of epigenetic marks. In ChromHMM, active enhancers are defined H3K27ac in combination with H3K4 mono-methylation (and an absence of other marks).

2.2.1 Enhancers Act in Concert Across Broad Genomic Regions

First, I established whether or not enhancers act together, to make sure that this is a valid assumption to use in later models. I started by calculating the Spearman Rank Correlation Coefficient (ρ) between H3K27ac levels at enhancers, and expression levels (by RNA-seq) at gene promoters, across the 38 cell types. I did not assume a direct linear correlation between variables, which would not be expected except in situations where a single enhancer controlled a single gene, therefore I used Spearman correlation. There is no

need to normalise signal between enhancers, since each correlation is independent. I calculated these values across TADs, since I expected co-ordination of enhancers to be present at this level (see Introduction).

To visualise these correlations, I plotted the value of ρ on the Y axis, and genomic coordinates on the X axis. Each correlation value is plotted at the coordinate of the *enhancer* (specifically, the centre of the acetylation peak). Each TAD contains many genes, and so there are multiple values plotted at each enhancer location, with the *gene* distinguished by colour. Therefore, spatial patterns of enhancer/gene coactivation will be visible as stretches of high or low correlations of the same colour.

Figure 2.1 shows the Frizzled-7 (FZD7) gene locus, including the TAD containing FZD7 and 200kb flanking the TAD boundaries. Panel a illustrates the difficulties in assigning genes based on individual correlations alone, because the signal-to-noise ratio is so high. Many enhancers show large positive and negative correlations with multiple genes, and these are distributed considerably outside the expected values for random variables, due to the correlation structure between the expression of genes in the TAD. The 95% confidence interval for the Spearman correlation of 38 random variables is ± 0.32 (permutation test, $n = 10^6$)).

However, if we simplify this graph to show only the enhancers for a single gene, a clear pattern emerges: in this case for the FZD7 gene. There are consistently high correlations between enhancer acetylation and FZD7 expression across most of the body of the TAD, and strikingly only two enhancers showing negative correlations in this region. This supports the idea that the vast majority, if not all, of the enhancers in this region are indeed enhancers for a single gene, FZD7, and that while there are individually high correlations between enhancers and other genes, these are unlikely to be significant. Significantly, FZD7 is an important developmental regulator (Finch et al., 1997) and is surrounded by non-coding conservation, overlapping with the GRBs identified in (Harmston et al., 2017).

Figure 2.2a illustrates many of the same features as Figure 2.1. It shows the genomic region around two TADs, containing the genes Ikaros (IKZF1) and Growth factor receptor-bound protein 10 (GRB10). Ikaros is also sur-

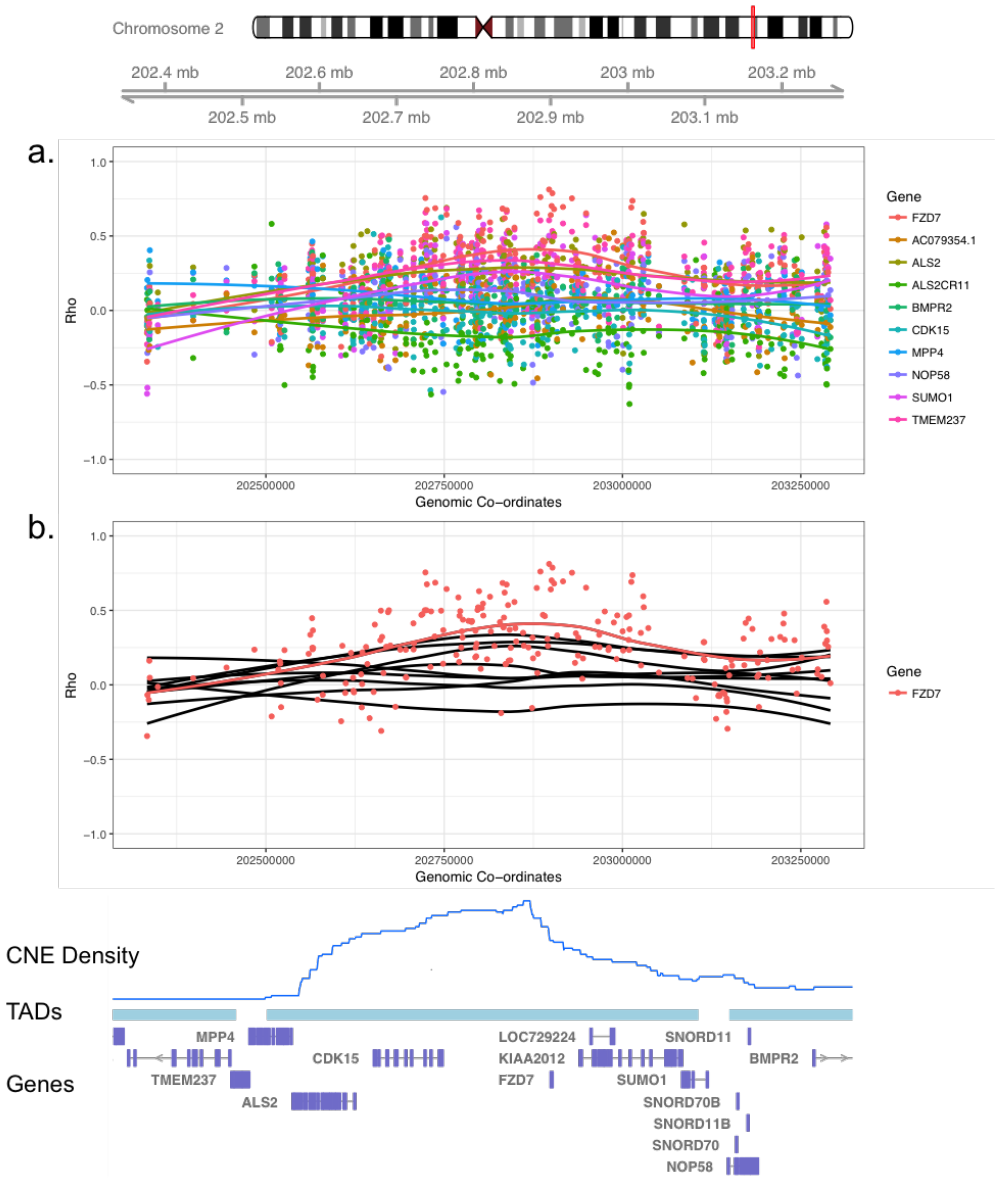

Figure 2.1: Correlations between promoters and enhancers at the FZD7 Locus a. Correlations between enhancers and all promoters within the FZD7-containing TAD. The dots represent individual correlations coloured by gene, and positioned according to the enhancer. Coloured lines show LOESS smoothed correlations for each gene. b. The same region, but with correlations only displayed for the FZD7 gene. Black lines show LOESS smoothed averages for other genes. CNE density is taken from Ancora ((Engström et al., 2008), see Methods)

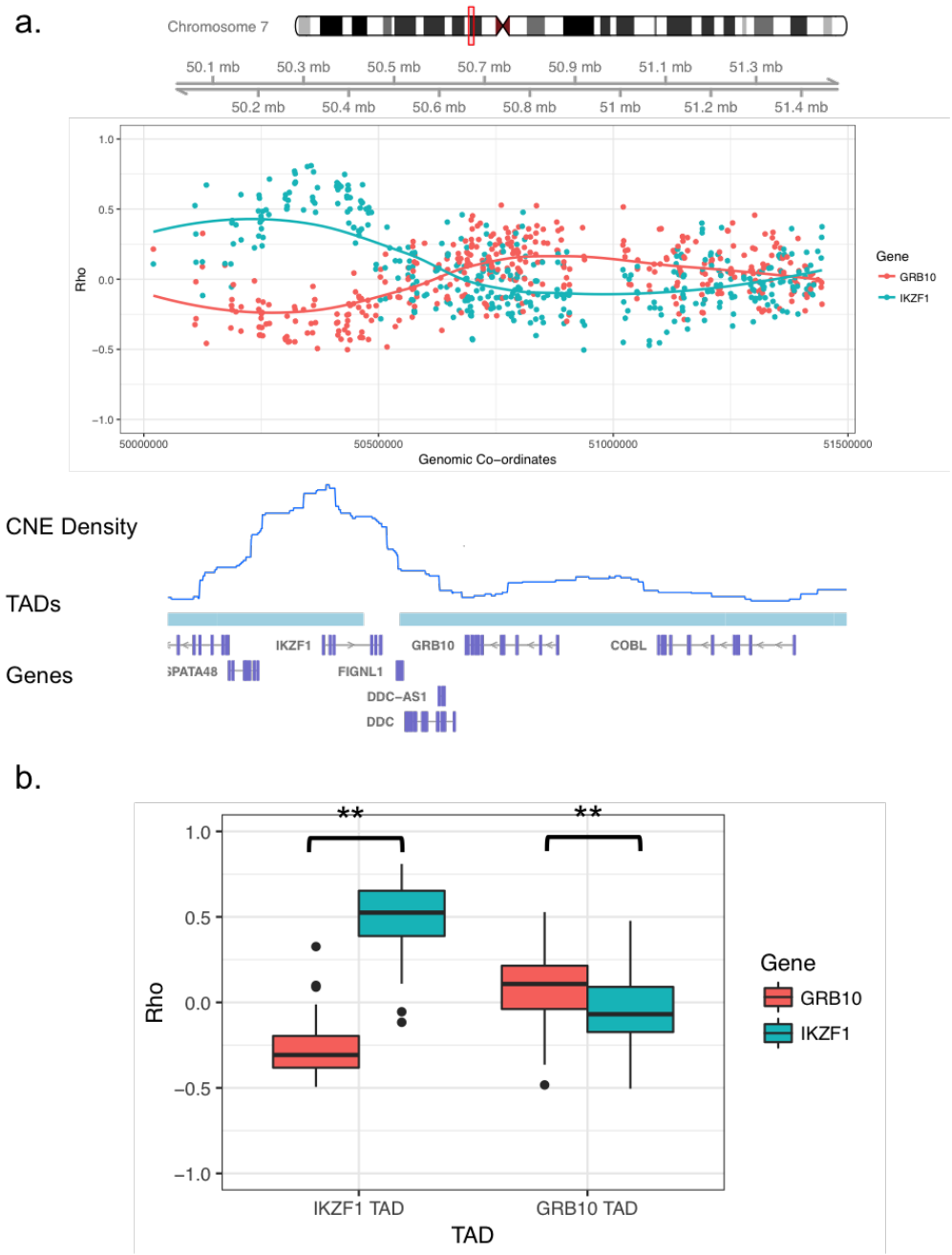

Figure 2.2: a. Correlations between enhancers and promoters at the IKZF/GRB10 locus, and annotation as for Figure 2.1. Only correlations for these two genes are shown. b. Boxplot of enhancer correlations for IKZF1 and GRB10, split by TAD.

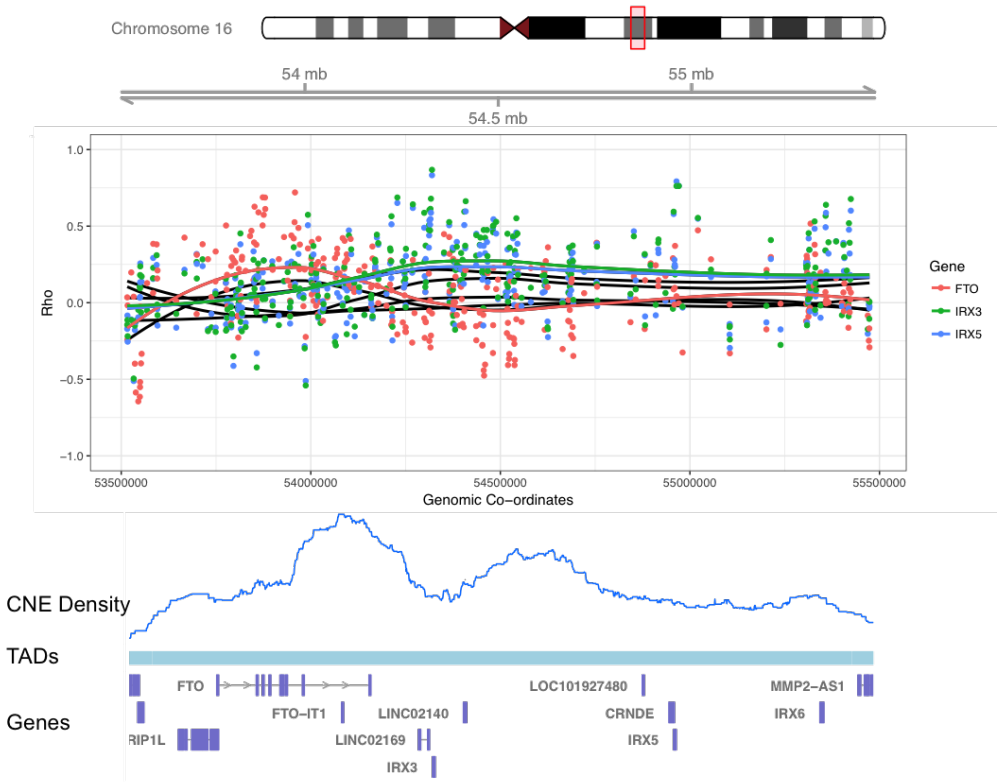

Figure 2.3: Correlations between enhancers and promoters at the IRX3/5 locus, plus annotation as for Figure 2.1. Only correlations for IRX3, IRX5 and FTO are shown.

rounded by high levels of non-coding conservation (see the "CNE Density" track). Around both genes (specifically, within their TADs), there is a striking local pattern in the correlations between enhancer activation and gene expression. This pattern changes as the TAD boundary is crossed.

Most of the enhancers show high positive correlations for the gene in the same TAD, and within the TADs there is again a strong reduction of negative correlations. This is shown in Figure 2.2b., with a highly significant different in correlations between enhancers and each gene (IKZF TAD, $p < 2.2x10^{-16}$), GRB10 TAD, $p < 1.2x10^{-12}$).

Figure 2.3 shows the IRX3/5 locus, highlighting the genes IRX3, IRX5

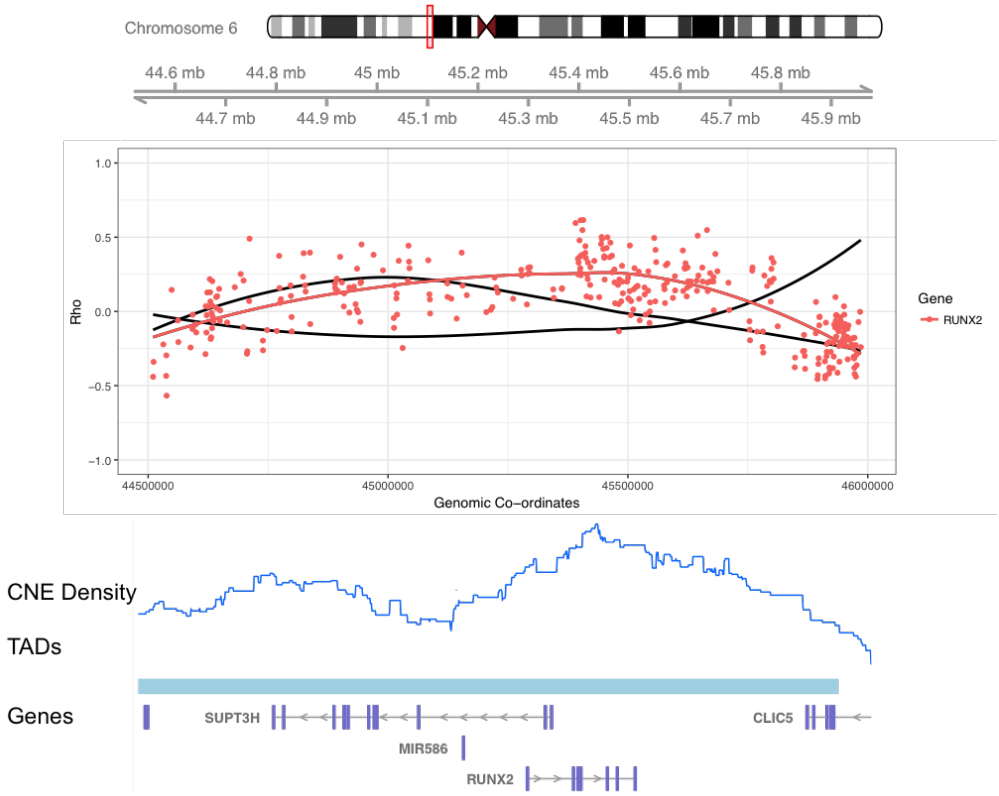

Figure 2.4: Correlations between enhancers and promoters at the RUNX2 locus, plus annotation as for Figure 2.1. Only correlations for RUNX2 are shown.

and FTO genes. IRX3 and 5 are members of the Iroquois family of transcription factors, which are important developmental regulators. FTO is an RNA demethylase that was implicated in obesity through a variant present in its first intron (Frayling et al., 2007), which was later linked to IRX3 by long-range physical contacts (Ragvin et al., 2010). Across most of the TAD, enhancers correlate more strongly with the two IRX genes, however across the body of the FTO genes, many enhancers correlate better with expression of FTO, therefore it is not clear from this analysis alone which genes are likely to be the targets of long range regulation.

Figure 2.4 shows the RUNX2 locus. RUNX2 is a transcription factor involved in osteoblast development osteoblast development (Lee et al., 2000)

and implicated in cancer (Pratap et al., 2005). RUNX2 is present at a relatively gene-poor locus, and across the length of its TAD shows an over-representation of positive correlations with enhancers.

Based on visual inspection of Figures 2.1-2.4, and at many genomic loci, I concluded that enhancer activity seems to be frequently co-ordinated to the expression of particular genes across TADs, and that this would be a reasonable assumption to add into models of enhancer activity.

2.2.2 Predicting Target Genes

Manual plotting of TADs and inspection of the results can provide qualitative evidence, but I wanted to test my hypothesis using a quantitative model to discover targets genes genome-wide. Using the same data as above, I fitted a regression model between locus-wide acetylation levels at enhancers within TADs and gene expression (see Methods). This model does make predictions of expression, but I was less interested in the accuracy of these predictions and more interested in the relationship between acetylation and expression: can it be shown that locus-wide acetylation has a statistically-significant correlation with gene expression?

Figure 2.5 shows the results of running the model on the Roadmap Epigenome panel of cell lines in terms of the predicted p-values. A low p-value indicates that the relationship between locus-wide enhancer activity and gene expression is stronger than would be expected by chance. There is a clear enrichment for low p-values above background, as indicated by the dotted line. After multiple testing correction (Benjamini-Hochberg, FDR ¡ 0.05), I identified 3086 genes under long-range regulation.

Genes under long regulation are more likely to overlap CpG islands (56% vs. 52% for non-LRR, $p < 3.35x10^{-4}$, Fisher test) and CpG islands overlapping genes under long-range regulation are longer on average (1284bp vs. 1023bp, $p < 2.2x10^{16}$, Wilcoxon test). This is of note because previous studies (see (Lenhard et al., 2012)) have associated longer CpG islands with developmentally regulated genes.

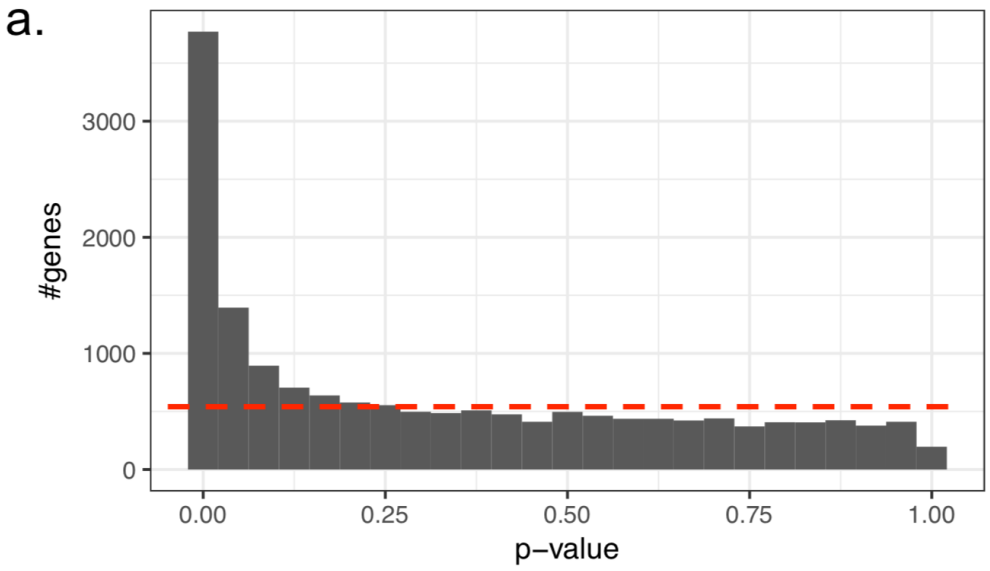

Figure 2.5: a. Unadjusted P-values for the correlation of enhancers with gene expression across TADs. The red dotted line indicates "excess significance", above results that would be expected by random chance.

2.2.3 GO Enrichment

I decided to investigate which genes made up the set of predicted target genes, and I started by looking at Gene Ontology enrichment. The 'Biological Process' ontology clearly shows that this gene set is very significantly enriched for developmental genes. This pattern of enrichment is strikingly similar to GO enrichment of GRB-associated genes (Kikuta et al., 2007; Akalin et al., 2009).

As is common with GO enrichment analysis, the most clearly enriched terms are very general. Further down the in the list are "embryo development" ($p = 2.01 \times 10^{-13}$), "regulation of developmental process" ($p = 2.30 \times 10^{-14}$), and intriguingly "cell adhesion" ($p = 8.65 \times 10^{-15}$). There are also many specific terms relating to system development, such as "nervous system development" ($p = 5.4 \times 10^{-12}$), "skeletal system development" ($p = 1.07 \times 10^{-11}$) and "heart development" ($p = 9.83 \times 10^{-8}$, see below).

The *Molecular Function* ontology paints a similar picture, with the high-

GOBPID	Ratio	Count	Size	Term	P-value
GO:0009653	1.72	890	2172	anatomical structure morph.*	7.49×10^{-25}
GO:0048731	1.55	1458	3879	system development	7.37×10^{-24}
GO:0044707	1.50	1900	5258	single-MCO§ P.†	9.89×10^{-24}
GO:0048856	1.51	1728	4719	anatomical structure dev.‡	1.31×10^{-23}
GO:0044767	1.50	1808	4979	single-organism dev. P.	4.34×10^{-23}
GO:0048869	1.56	1297	3403	cellular developmental P.	5.98×10^{-23}
GO:0032501	1.49	2041	5722	MCO P.	7.23×10^{-23}
GO:0032502	1.50	1831	5057	developmental P.	8.95×10^{-23}
GO:0030154	1.56	1240	3246	cell differentiation	5.81×10^{-22}
GO:0007275	1.50	1595	4336	MCO dev.	8.73×10^{-22}

Table 2.1: GO Enrichment (Biological Process ontology) for genes under regulation by enhancers. P-values are Bonferroni corrected. *morphogenesis §multi-cellular organism †process ‡development

est ranked terms indicating either transcription factor activity (e.g. "RNA Pol II. Transcription Factor activity", Adj. P-value $6.06x10^{-6}$) or receptor activity ("Signal Transducer Activity", Adj. P-value $2.79x10^{-4}$).

The *Cellular Component* ontology provides information about the location of the protein within the cell. Many of the highly enriched terms in this category relate to membrane-bound proteins, including the most significant ("Plasma Membrane", Adj. P-value $1.31x10^{-19}$) and the largest represented category, "Membrane" (2369 genes, Adj. P-value $1.40x10^{-4}$), which indicates that over half the genes controlled by long-range regulation are present in the cellular membrane. This likely refers in the most part to receptor proteins, as per the Molecular Function ontology.

In order to further investigate gene regulation within specific tissue, looking at heart development. The Roadmap Epigenome dataset contains 3 heart tissues: Aorta, Right Ventricle, and Right Atrium. Figure 2.6 shows 3 genes with varying expression in all 3 tissues which are all key for heart development. GATA6 is expressed throughout the heart at moderate levels, and is a key transcription factor involved in heart development, and GATA6 mutations have been linked to heart defects(Chao et al., 2015). S1PR1 is a G-protein coupled receptor which is important for vascular develop-

GOMFID	Ratio	Count	Size	Term	P-value
GO:0000981	1.81	170	566	Pol. II TF†, DNA-binding	2.15×10^{-06}
GO:0001228	2.17	99	290	Pol. II DNA-binding	7.07×10^{-06}
GO:0008092	1.65	212	755	cytoskeletal protein binding	1.13×10^{-05}
GO:0005198	1.79	148	496	structural molecule activity	4.37×10^{-05}
GO:0004871	1.50	291	1116	signal transducer activity	5.63×10^{-05}
GO:0003779	1.89	110	353	actin binding	2.69×10^{-04}
GO:0004872	1.44	263	1035	receptor activity	2.92×10^{-03}
GO:0060089	1.44	263	1035	molecular transducer activity	2.92×10^{-03}
GO:0015026	5.15	20	36	coreceptor activity	4.30×10^{-03}
GO:0038023	1.49	213	816	signalling receptor activity	4.55×10^{-03}

Table 2.2: GO Enrichment (Molecular Function ontology) for genes under regulation by enhancers. P-values are Bonferroni corrected. †Transcription Factor

ment (Chae et al., 2004). MYL2 is the human smooth muscle myosin heavy chain(Matsuoka et al., 1993), and so is vitally important for heart function, and is very highly expressed in the Right Ventricle.

2.2.4 Enhancers can affect multiple genes in a TAD

The majority of TADs analysed contain only a single gene predicted to be regulated by enhancers. In fact, there are more TADs containing only a single gene under long-range regulation than a random sample obtained by permuting TAD labels ($p < 10^{-5}$), which indicates there may be an underlying biological reason. Intuitively, this arrangement would mean that each enhancer only regulates one gene, permitting fine-grained control of gene expression.

However, significant numbers of TADs contain more than one gene predicted to be regulated by enhancers. There are many possible reasons for this, which I will explore below. First of all, this effect might be due to statistical artefacts. Using a relatively permissive FDR of 0.05, we would expect to see some TADs with more than one gene predicted simply by chance. We would expect such false positives to occur more or less randomly throughout the data, so we can calculate the expected baseline by repeatedly re-sampling

GOCCID	Ratio	Count	Size	Term	P-value
GO:0071944	1.66	1054	4087	cell periphery	5.69×10^{-27}
GO:0005886	1.67	1032	3986	Plasma Membrane (P.M.)	7.03×10^{-27}
GO:0044459	1.62	589	2202	P.M. part	1.33×10^{-15}
GO:0005887	1.68	376	1339	integral comp. of P.M.	5.02×10^{-12}
GO:0031226	1.64	385	1395	intrinsic comp. of P.M.	3.83×10^{-11}
GO:0043292	2.67	76	194	contractile fibre	2.58×10^{-07}
GO:0005578	2.18	107	311	proteinaceous ECM†	6.05×10^{-07}
GO:0016021	1.31	950	4153	integral comp. of M.	9.63×10^{-07}
GO:0031012	1.94	136	428	extracellular matrix	1.34×10^{-06}
GO:0030016	2.60	71	184	myofibril	2.17×10^{-06}

Table 2.3: GO Enrichment (Cellular Component ontology) for genes under regulation by enhancers. P-values are Bonferroni corrected. †Extracellular Matrix

our data after permuting TAD labels. This is shown in figure 2.7, as an average of 100 resamples: TADs containing 3 or more target genes are in fact *depleted* in the data when compared to the values which would be expected were target genes distributed randomly, which indicates that it is extremely unlikely that we see so many multi-target TADs because of statistical effects alone.

Secondly, it might be the case that two genes are truly co-regulated and are generally expressed in the same tissues. In this case, the method would correctly identify both genes as targets. One possible example of this are the IRX3/5 genes, which are shown in Figure 2.3. These genes are closely related transcription factors which are active in a wide range of tissues, with functions conserved across vertebrates and invertebrates, making their co-regulation highly plausible (Kerner et al., 2009). Their expression is tightly coupled across cell types, with an R^2 value of 0.86 (Pearson correlation coefficient). However, using correlation alone cannot distinguish this from the possibility that are independently regulated, and that this effect is not seen by chance in this specific case. Many other highly correlated gene pairs also share evolutionary history, such as DLX1/2 (Homeobox transcription factors, $R^2 = 0.98$), CD5/6 (T-cell surface receptors, $R^2 = 0.97$) and CAV1/2 (voltage-dependent calcium channel sub-units, $R^2 = 0.84$).

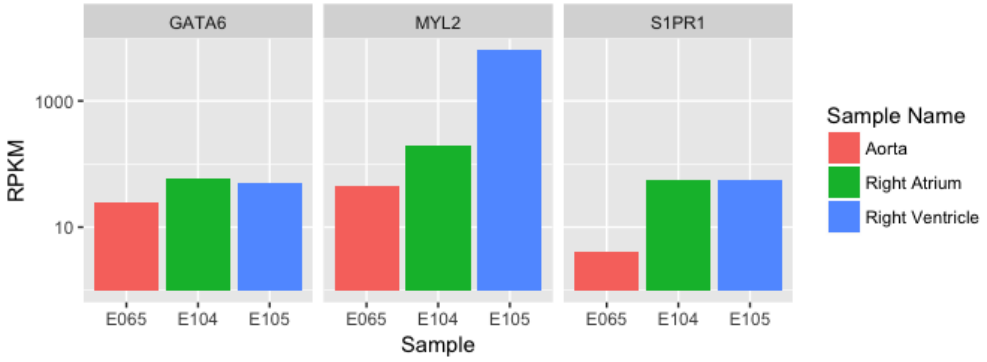

Figure 2.6: Expression patterns of 3 genes involved in heart development, across the 3 Roadmap samples from heart tissue.

In order to test co-regulation more generally, I compared the correlation between target gene pairs within TADs with exactly 2 target genes predicted (n=430), with the correlation values obtained from gene pairs chosen at random from each TAD (n=2431). The median correlation value for predicted target genes was 0.38, vs 0.03 for random genes, and based on subsampling the randomly selected genes, $p < 10^{-5}$ for this comparison. However, this is probably to be expected given that pairs of target genes are already known to correlated independently with acetylation levels at the locus.

Despite the over-representation of TADs containing 1 or 2 target genes, there are also regions which contain high numbers of genes are predicted as targets. According to 2.7, there are no more of these regions in the dataset than would be expected by chance. However, The regions with the highest number of coregulated genes (and frequently the most significant p-values as well) are in many cases arrays of closely related genes, which are thought to be coregulated based on existing literature. This suggests there may be a biological rationale behind these arrangements, even if they are not enriched statistically.

The Protocadherin-α gene cluster, containing 16 putative target genes, is of particular interest. These genes have been extensively studied due to their striking organisation (Wu and Maniatis, 1999) and association with CTCF binding (Monahan et al., 2012; Guo et al., 2012). Protocadherins are

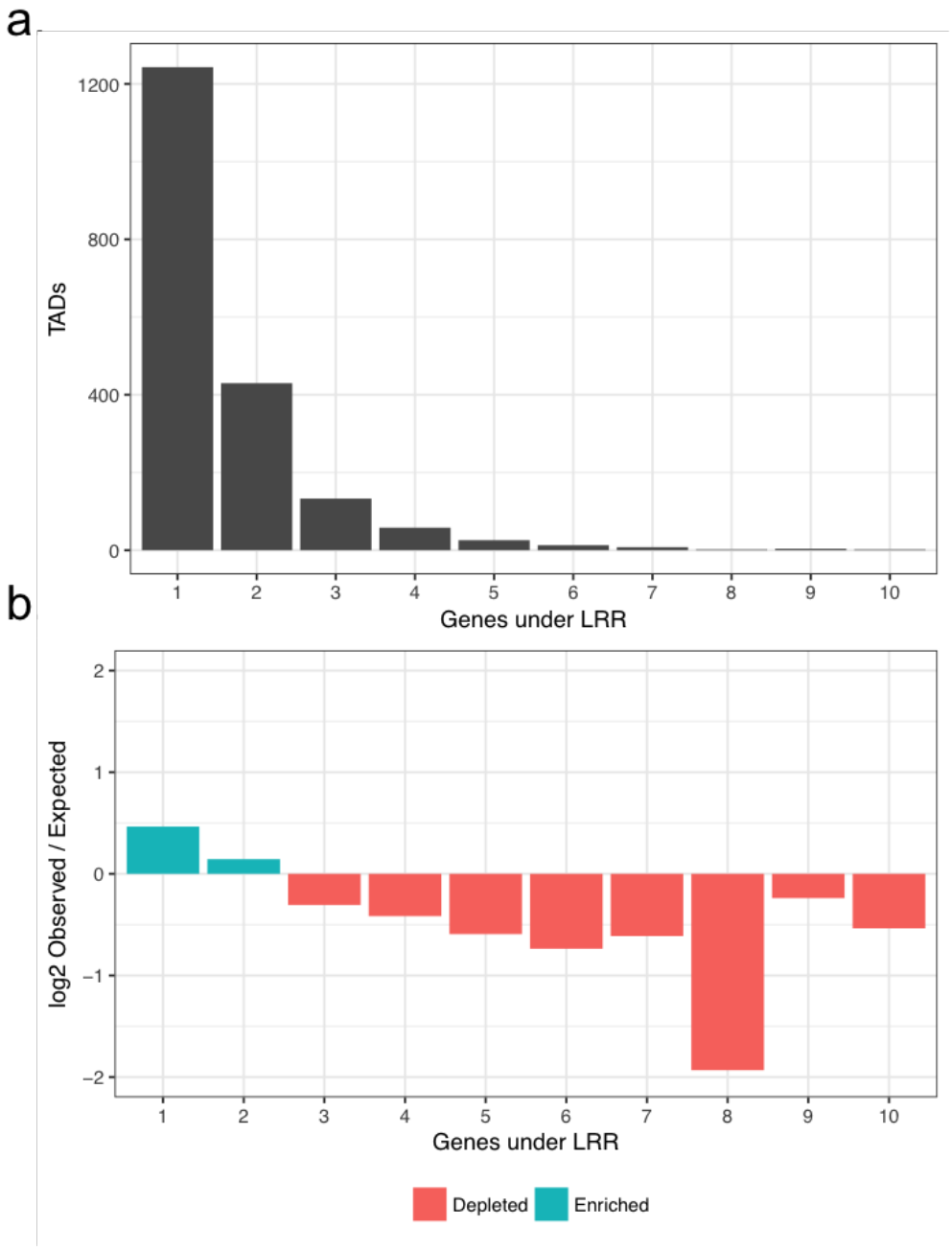

Figure 2.7: a. The distribution of the number of genes under long-range regulation in individual TADs. A single TAD containing 16 genes under long-range regulation is not shown. b. The log-ratio of of the observed values in (a) to the expected value if target genes were randomly assigned (calculated from 100 permutations).

TAD Co-ordinates	Predicted Targets	Gene Family
Chr5:140.1Mb-140.5Mb	16	Protocadherin-α
Chr1:247.7Mb-248.9Mb	10	Olfactory Receptor
Chr14:104Mb-106Mb	10	Unknown
Chr1:152.9Mb-153.3Mb	9	Small Proline Rich Protein Family
Chr6:31.5Mb-32.9Mb	9	Human Leukocyte Antigen
Chr17:38.5Mb-39.8Mb	9	Keratins
Chr19:19.7Mb-21.3Mb	9	Zinc Fingers
Chr1:154.1Mb-155.2Mb	8	Unknown
Chr11:118Mb-118.5Mb	8	Unknown
Chr1:152Mb-152.9Mb	7	Unknown

Table 2.4: The 10 TADs containing the largest numbers of genes under long-regulation. The predominant gene family at the locus is indicated, although other genes are present in some cases.

stochastically expressed, with each cell only expressing a limited number of Protocadherins, and this enables them to act as cell surface markers which give cellular identity to developing neurons. Activation by enhancers is key to promoter selection (Guo et al., 2015), and all three key enhancers correlate with the expression of all Protocadherin-α genes. In this particular example, it is unlikely that all of the genes are expressed in a single cell due to the switch-like regulatory system elucidated in (Guo et al., 2015). However, they are all expressed in the same tissue (developing neurons in this case), and since the resolution of our data is well below the single-cell level there is no way to distinguish this from the current data.

2.2.5 TAD boundaries restrict the effects of enhancers

Running the model as described provides results on a *per-gene*, not a *per-enhancer* basis, which means that it is not possible to investigate directly how regulation by enhancers varies across tads, or across TAD boundaries. To do this, I derived a statistic for the contribution of each individual enhancer within a TAD, across the cell types in the Roadmap dataset (see Methods).

I re-ran the model using TADs extended by 200kb either side to test whether TAD boundaries had the expected effect on enhancer regulation.

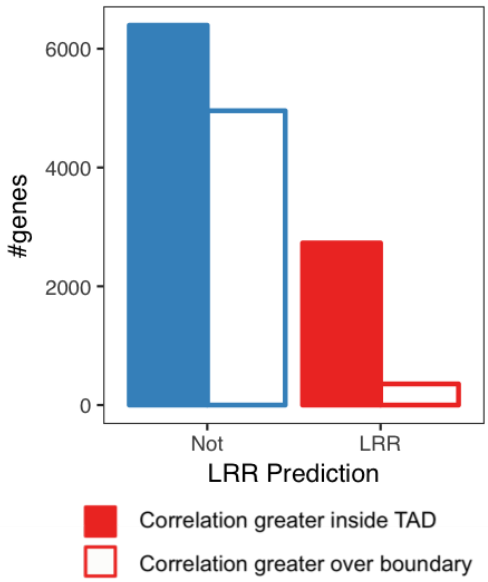

Figure 2.8: Average enhancer contributions, Δr, were calculated for enhancers inside TADS vs. those up to 200kb from the boundary, outside of the TAD. Genes are divided into those which had a greater average contribution from enhancers *within* the TAD, and those which had a greater contribution from enhancers *outside* the TAD. Genes are also split by LRR class.

The vast majority of genes under long-range regulation showed a greater average contribution by enhancers inside their respective TADs than those outside, but with little difference for those not predicted to be regulated by enhancers Figure 2.8.

However, this result could easily be explained as purely a function of distance, rather than TAD boundaries. To check for this possibility, I plotted enhancer contribution for enhancers inside and outside of TADs, which is shown in Figure 2.9. This clearly shows a greater contribution of within-TAD enhancers to the regulation of LRR genes, compared to those outside, at all distances, demonstrating that TAD boundaries exert a considerable effect on the ability of enhancers to regulate genes across those boundaries. The negative values stem from the fact that a small random change in prediction is more likely to have a negative effect on the correlation (r) for a good prediction than a bad prediction. Enhancers very close to the gene, but over

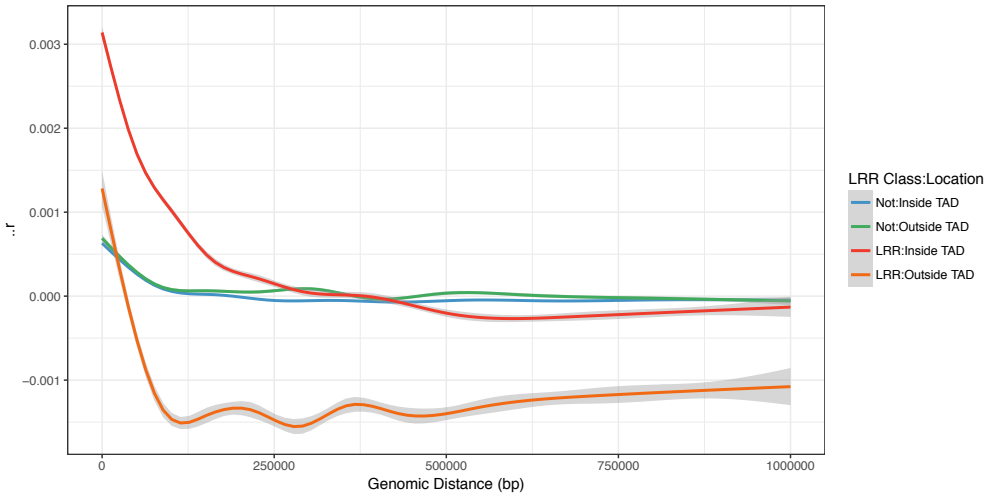

Figure 2.9: The average contribution of enhancers shown as a function of distance from promoters of possible regulatory partners. Different lines indicate the LRR status of the putative target gene, and the position of the enhancer relative to the TAD containing that gene.

a TAD boundary, have a more positive impact on the model than we expect. This is possibly due to poorly called TAD boundaries.

2.2.6 Genes under long-range regulation are enriched for long-range contacts

As discussed in the introduction, a popular explanation for the specificity of enhancer-promoter regulation is that the promoters only form 3D contacts with specific enhancers, so enhancers cannot regulate other genes. I decided to explore using high-resolution interaction data from Promoter-capture Hi-C (see Methods). Promoter-capture Hi-C is an experimental technique which allows enhancer-promoter contacts to be examined genome-wide through the use of hybrid-capture baits to enrich contacts of interest surrounding promoters (Schoenfelder et al., 2015), which gives greater resolution for interactions at regions of interest without requiring additional sequencing.

Promoter Capture Hi-C is not available in a wide variety of cell lines, but it is available in CDC34+ cells (Mifsud et al., 2015), which is fortunately also

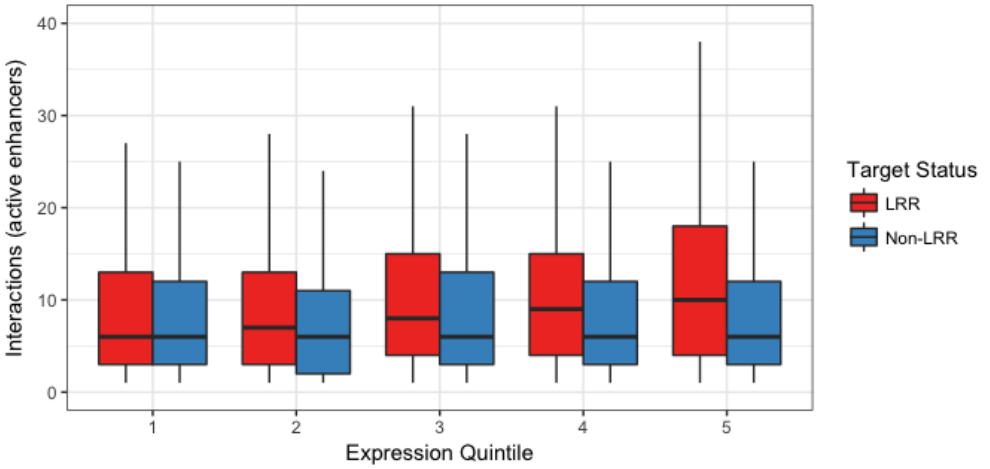

Figure 2.10: Number of significant chromatin contacts with enhancers for LRR and Non-LRR genes.

present in the Roadmap Epigeome data used in this analysis, which means that the predictions made should be valid. *CD34+ cells* in this context refers to hæmatopoietic progenitor cells which express the CD34 transmembrane protein. These are generally found in the umbilical cord and bone marrow, and can give rise to a number of differentiated cell types. CD34 itself may be a marker with importance outside of the hæmatopoietic lineage (Sidney et al., 2014).

Figure 2.10 shows the number of unique interactions between enhancers (as defined in the earlier analysis) and genes predicted to be either targets of or unresponsive to long-range regulation (designated "LRR" and "Non-LRR" respectively) using Promoter Capture Hi-C data from (Mifsud et al., 2015). Genes are split by expression quintile (expression data also from (Mifsud et al., 2015)), which guards against potential confounding (for example, it is possible that this particular assay enriches fragments with open chromatin) and allows us to investigate the relationship between enhancer contacts and expression.

It is clear from the figure that target are, in general, enriched for enhancer contacts compared to non-target genes, and that this effect is stronger for highly expressed target genes. This is confirmed using a Wilcoxon rank-sum

test: for highly expressed (quintiles 4 and 5) genes, target genes are enriched in contacts vs non-target genes (mean 13.4 vs 9.2, $p < 10^{-15}$).

Equally, non-target genes do not show any trend for increased (or decreased) interactions across expression quintiles, suggesting that this set of genes is *not* responsive to the effect of enhancers (or at least that, for these genes, contact with enhancers is not the dominant form of gene regulation). The differences between the 1st quintile of expression and the 5th quintile are not statistically significant despite the large number of genes tested ($p = 0.12$, $n = 2298$).

Interestingly, there is no difference in the number of contacts for target and non-target genes in the lowest expression quintile ($p < 0.55$). This may represent an average background level of chromatin contacts, in the absence of any regulatory events (e.g. distal binding and activation of enhancers by transcription factors). This relatively high background level of chromatin contacts (shown in the lowly-expressed target genes and in all non-target genes) suggests that 3D genome conformation alone might not be sufficient to explain why some promoters, and not others, are sensitive to regulation by enhancers, which is discussed further below.

2.3 Discussion

In this chapter I have presented a new method for using functional genomics data to link enhancers to promoters, a long-standing problem in regulatory genomics. This has produced a list of putative target genes, which are very strongly enriched for developmentally regulated genes, which appear to be the primary targets of long-range regulation. I have also shown that the actions of enhancers are co-ordinated across TADs, and that TADs restrict enhancer activity to inside their borders. Genes inside TADs, however, contact enhancers but do not show evidence of regulation.

This poses an interesting question for current models of gene regulation. We know that the 3D structure of the genome is highly regulated, It is well established that TADs restrict enhancer action across boundaries (Lupiáñez et al., 2015). However, TADs represent broad domains of association, rather

than tightly regulated loops, and the current, widely supported theory of loop extrusion fits this model (Sanborn et al., 2015b; Goloborodko et al., 2016). Rather than specific anchors fixed together in space, this model supports a continuous process whereby loop of chromatin are continuously broken and reformed. Very high-resolution Hi-C has identified examples of CTCF-bound loops, but these appear to be exceptions rather than the rule (Rao et al., 2014).

This creates a 'last mile' problem of how enhancers differentiate between neighbouring genes within TADs. Many previous papers on enhancer-promoter interactions have focused on physical interactions exclusively, and suggest a model where physical interactions are causally responsible for selecting enhancer targets (Fraser et al., 2015). For loci with many genes and enhancers in close proximity, this model of individually regulated interactions starts to seem overly complex. Loop extrusion does not create topologies where overlapping loops create distinct territories; rather, this would simple create a highly compacted domain. Cross-referencing my results with the available high-resolution Hi-C data indicates that, at least in CDC34+ cells, enhancers are not in fact physically prevented from contacting the promoters of other genes.

The regulation of specific classes of genes offers one solution to this problem. I propose a model in which strong TAD borders prevent enhancers from contacting promoters outside of their domain, but within TADs interactions can take place at random, perhaps driven in part by open chromatin regions. In this model, the 'last mile' problem is not such an issue, because promoters which are not regulated by long-range regulation are not responsive to the contact from enhancers, whereas other promoters, primarily developmentally regulated genes, can respond to input from enhancers. The mechanisms underlying this process are not at all clear, but there are several lines of relevant evidence.

Firstly, we know that the promoters of different classes of genes have markedly different characteristics when considered *en masse* (Lenhard et al., 2012), even if these differences are hard to detect at the level of individual promoters. This might simply be because we lack sufficient understanding

of the biological sequence at promoters, as studies are generally restricted to looking at enriched motifs. PWM-based motifs do not always accurately characterise transcription-factor binding sites, and promoters may have significantly more complexity, since initiation requires simultaneous binding of a great number of factors. Even if motifs explain a good deal of the variation, current methods may lack the power to detect them in many situations (Simcha et al., 2012), particularly if they are rare.

Secondly, *in vitro* models have shown that promoters can indeed have a differential response to regulation by enhancers (Zabidi et al., 2015), and although the work was carried out in *Drosophila melanogaster* it is relevant to humans as a proof of concept. Their discoveries also follow the housekeeping vs. developmental gene dichotomy which I have shown here. Fundamental mechanisms of gene regulation are generally conserved even between distant metazoan clades, and there are indications that this is true for some mechanisms of long-range regulation (Harmston et al., 2017).

Finally, the conclusions of this chapter are in agreement with models of regulation derived from GRBs (Harmston et al., 2013). GRBs are large arrays of conserved non-coding elements, of which many have been shown to function as enhancers (Visel et al., 2007), and there is evidence from both genome-wide association and genome duplications that these arrays regulate developmental genes, ignoring nearby "bystanders" which are frequently housekeeping genes.

GRBs also have an underlying connection to TADs (Harmston et al., 2017), which supports the idea that enhancers must be confined to a local topological compartment, both to ensure contact with their target and prevent ectopic gene activation. GRBs are frequently larger in size than TADs, although this may simply be an artefact of the algorithm used to call TADs, which can be sensitive to small, local changes in interactions. These errors in calling TAD boundaries may explain the result in Figure 2.9, which showed that even over TAD boundaries, some enhancers appear to contribute to gene expression.

As we any statistical method, correlations within the data present a problem for this method, and are to have consequences for this experiment. Gene

expression patterns are not random, so the chance of two genes having a correlation higher than expected by the model is probably underestimated. For example, a transcription factor active in heart cells that is (assuming our model to be correct) regulated by enhancers, would likely appear to be co-regulated with a nearby gene if that gene was also active in heart tissue, even if there was no direct causal link. Without perturbation experiments we cannot see if genes with similar expression patterns are truly coregulated. There are regions, such as the proto-cadherin and olfactory receptor loci, which have strong evidence for coregulation, but this is unlikely to be the case universally.

The model assumes total independence of enhancers, and that each enhancer contributes equally. This is a key statistical assumption, as it stops the model from overfitting, however it is somewhat unrealistic. There is some evidence that enhancer do act independently in a flexible, modular fashion (see Introduction), however there may well be cases where more complex relationships exist. It is also very likely that different enhancers exert more or less influence on gene expression, possibly based on proximity (in either linear genome space or in 3 dimensions), the various factors which bind to the enhancer, or some unknown factor affecting "enhancer strength".

The metric we are using for enhancer activation (H3K27 acetylation) might also lead to false positives. I have tried to filter as conservatively as possible for acetylation peaks around promoters (using Ensembl data), but unannotated transcripts or small RNAs could be present in the peak set used, and it is likely that these genes would produce false correlations in some circumstances due to the issues outlined above.

Equally, the correlations I have analysed might also produce false negatives in some cases. It has been conjectured that H3K27 acetylation does not mark all enhancers, and other marks have been proposed as being important (for example (Taylor et al., 2013)), which could result in false negatives in regions with many such enhancers. The method outlined is also unable to detect long-range regulation outside of the cell types assayed, if the genes are not active in our panel of cell types which, while extensive, is nowhere near exhaustive. In fact, genes which are only active in a single cell (within the

panel) may not be called as target genes due to a lack of statistical power. TAD boundaries may also shift between cell types, which I am not able to account for in the model, nor is Hi-C data, required to define TADs, available in most of the cell types assayed.

There is also a question of what exactly the correlation between gene expression and acetylation is measuring. Due to the modular nature of enhancer activity (see Introduction), individual enhancers are likely to be active only in single tissues. This is not a problem for the method, since in the case of a gene which is active in several tissue but driven by different enhancers, the activity of these various enhancers would be considered together. However, in this case we would expect each enhancer to be either *on* (in one particular tissue) or *off* (in all other tissues), and this kind of binary data is not well approximated by regression modelling. It is possible that in this data, what we actually see are binary features (as described) which appear to be continuous, because we are using ChIP-seq and RNA-seq performed on bulk samples. Therefore, "high" expression and "high" enhancer acetylation might be better interpreted as a sample containing a high proportion of cells in which the genes and enhancers are active. In any case, it does appear that the model is finding signal from these features, even if it is not certain exactly what we are looking at at the single-cell level. It would be very interesting to re-examine this question as more and more single-cell data becomes available.

These results may be of great interest to researchers who want to identify the effects of distal, non-coding SNPs. Many examples exist of disease associations in which SNPs regulate a gene which is not the closest, even if the SNP is located within an intron of that gene (Ragvin et al., 2010; Lettice et al., 2003). The provisional list of genes targeted by enhancers provides a strong set of first candidates for further experiments, and is preferable to the frequently-employed nearest-neighbour rule, or *ad-hoc* theories relating genes to diseases. In complex loci with tens of genes located within megabase-scale regions, it could prove invaluable for prioritising targets for downstream analysis.

2.4 Data

I defined enhancers by calling peaks on the ChIP-seq data, using the MACS peak caller (Zhang et al., 2008). Any peak identified in any of the samples would be used as an enhancer region, and these regions were merged across cell types using the DiffBind package (Ross-Innes et al., 2012), and normalised using Trimmed Mean of M-values (TMM) normalisation, as implemented in the package.

RNA-seq data was obtained from Roadmap as RPKM (Reads Per Kilobase of transcript per Million mapped reads), which I converted to TPM (Tags per million mapped reads), which has been shown to be more appropriate for comparisons between samples (Wagner et al., 2012). I defined promoters as the transcription start site (TSS), using annotation from Ensembl 82 (Aken et al., 2016). H3K27ac is also present at active promoters, and so retaining peaks at active promoters would confound my results by effectively correlating two different measures of promoter activation (Histone acetylation and expression). It is difficult to exactly define the exact extent of promoter regions, and transcript models can also have some margin of error, so I excluded all peaks within 5kb of an annotated TSS.

This resulted in a dataset with defined enhancers and promoters for 38 cell types, listed in table 2.5. In order to investigate spatial patterns of enhancer action, I also needed pre-existing, independent genomic domains to define regions of co-operating enhancers. I used TADs defined in (Dixon et al., 2012). TADs are thought to vary little between cell types (Dixon et al., 2015; Battulin et al., 2015), so I used only the embryonic stem cell data to represent TADs for all cell lines. These data are summarised in table 2.6.

The "CNE density" track used for plotting is taken from the Ancora browser (Engström et al., 2008). Conserved Non-coding elements (CNEs) are short streteches of genomic DNA outside of protein-coding genes which exhibit very high levels of conseravtion. The track is formed by taking CNEs with a minimum identity of 96% over 50bp (i.e. 48 out of 50 bases matching exactly) from human (hg19) to mouse (mm10), and then applying a smoothing algorithm (see (Engström et al., 2008) to visualise the distribution of

these elements: higher density areas indicate a greater local concentration of conserved elements.

2.5 Methods

2.5.1 Modelling Enhancer-Promoter Interactions

The rationale for an automated approach here is clear: there are far too many TADs to sensibly examine by hand, and that approach risks introducing bias. Without a statistical, unbiased approaches, researchers might limit their investigation to genes already implicated in disease processes, or though to be more likely to be under long-regulation by prior knowledge, missing alternative possibilities. It is possible to perform statistical analyses on individual correlations, but calculating significance using positive/negative correlations alone throws away potentially useful data, reducing power. It is possible to calculate the mean correlation across TADs, but this is very susceptible to outlying high correlations. It is not known exactly how gene expression depends on activation by enhancers, so in order to model co-ordinated enhancer action I had to make several assumptions.

Activation of enhancers is correlated across TADs, but I assumed that the effect of each enhancer on gene expression was *independent*. The biological rationale for this is that each enhancer drives expression in a specific population of cells, and it is the combination of these individual patterns of enhancer activation which combine to form the overall expression of a gene (for examples, see (Lorberbaum et al., 2016)). I also assumed this affect is linear on expression. Enhancers, represented by H3K27ac peaks, varied considerably in size, so comparing acetylation signal directly is not reasonable. I normalised H3K27ac signal at enhancers to between 0 and 1, as a proportion of the maximum observed signal across all cell types. However, this assumes that the effect of each enhancer is equal. This is certainly an approximation of the biological reality, but allowing enhancer strength to vary means that the system is underdetermined in most cases, because there are more enhancers than conditions. Variable selection methods, such as

LASSO (Tibshirani, 1996), can address this, but this is also a distortion of biological reality, since we do not expect a large fraction of enhancers to have zero effect.

Posed in this way, we can model total gene expression (E_{Pred}) as a function of basal expression (E_{Basal}) and expression driven by enhancer acetylation. Expression driven by enhancers is modelled as a gene's sensitivity to long-range regulation (S_{LRR}), times the H3K27 acetylation levels (Ac) at all n enhancers within the same TAD:

$$E_{Pred} = E_{basal} + S_{LRR} * \sum_{1}^{n} Ac_i$$

This way we estimate both E_{basal} and S_{LRR}, which is the sensitivity of a promoter to enhancer activation within its TAD. S_{LRR} provides a measure of how sensitive genes are to regulation by enhancers, but this is not normalised between genes. We can instead calculate the statistical significance of the fit, which is straightforward since the model is effectively a simple linear regression. These p-values still require correction, since many tests have been performed, but only for the number of genes, not the number of possible gene-enhancer interactions. This results in a list of genes that show statistically significant evidence of regulation by enhancers.

An example of the output of this model for one of the best-predicted genes, KLK8, is show in Figure 2.11. The top panel shows the predicted RPKM values plotted against the observed values, which show a very strong correlation. The model has predicted the expression pattern both qualitatively and quantitatively, correctly predicting high expression in 2 cell types, with one much higher than the other. The lower plot visualises the contribution of each enhancer in a separate colour. The sum of each contribution defines the final prediction, as outlined in the equations above. The observed expression is indicated by a black outline. It is interesting to note that more individual enhancers are active in the highly expressed cell type than in the lower-expressed cell type, it is not that the same enhancers are active at lower levels (this is particularly obvious for the "pink" enhancers towards the bottom of the graph)

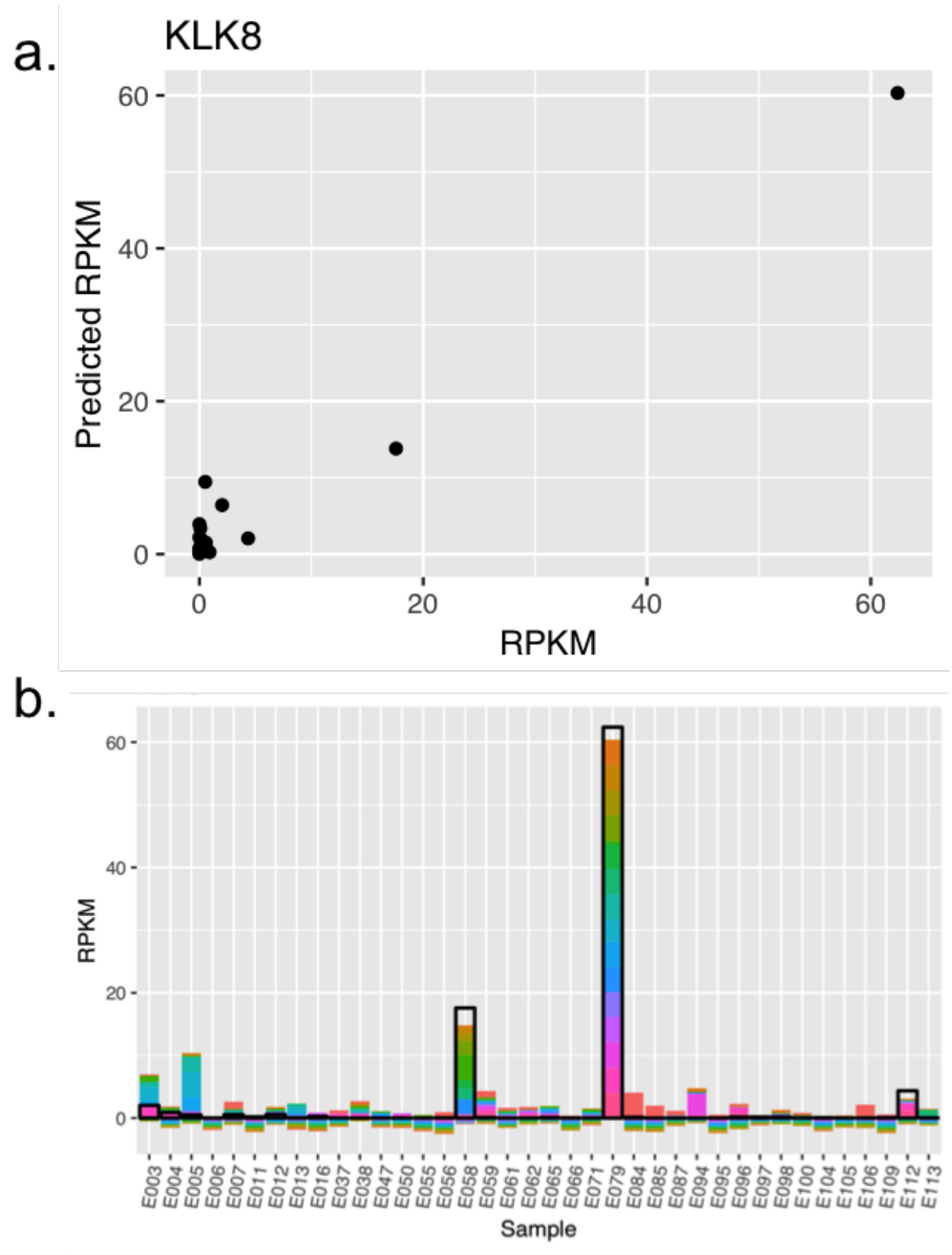

Figure 2.11: a. Predicted RPKM vs. Observed RPKM b. Predictions visualised as contributions of individual enhancers

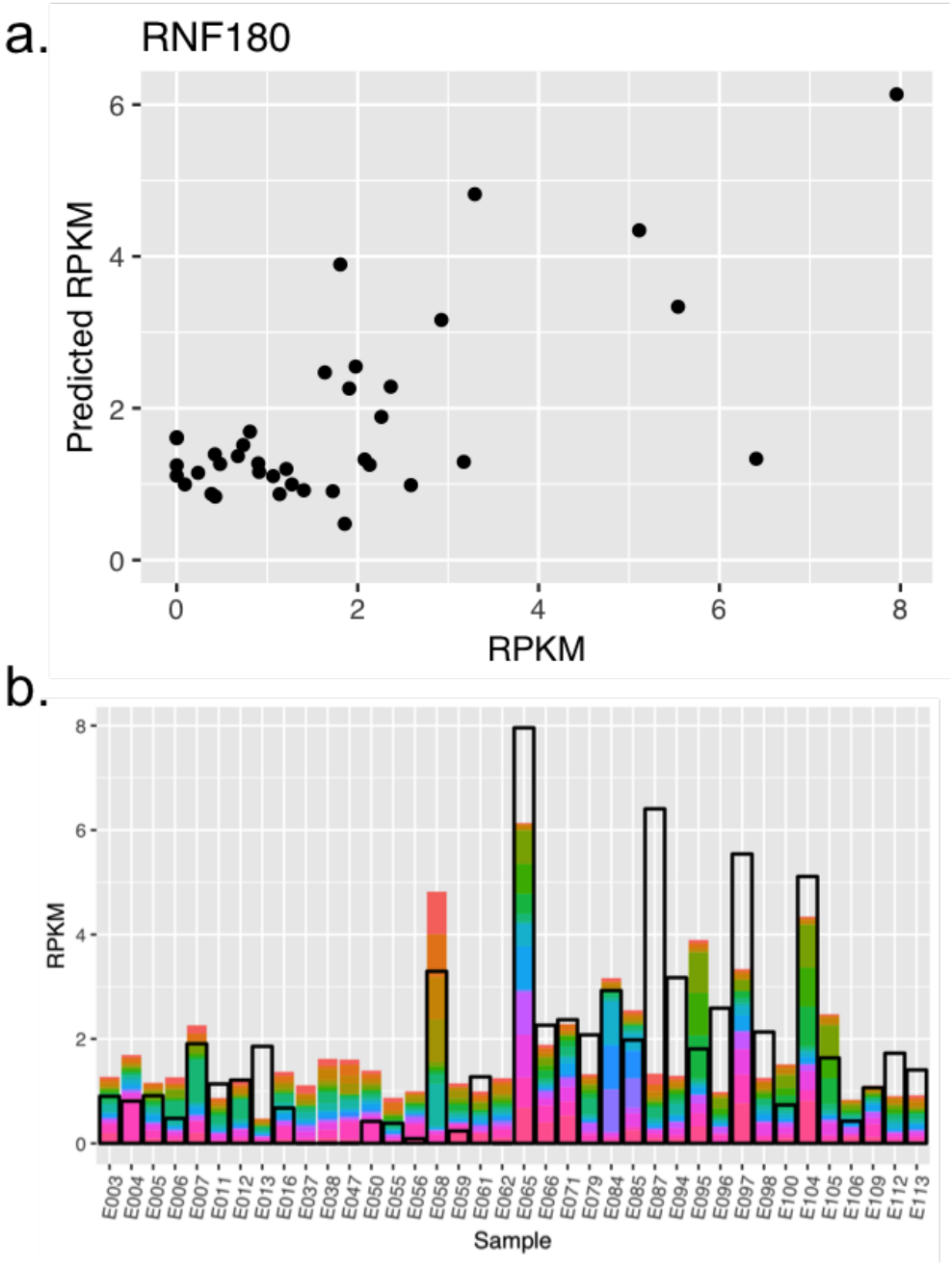

Figure 2.12: a. Predicted RPKM vs. Observed RPKM b. Predictions visualised as contributions of individual enhancers

Figure 2.12 shows a different gene, RNF180, which is still predicted as a target gene but with a much lower significance than KLK8. Here, the model broadly makes the correct *qualitative* predictions (i.e. high vs low expression), but also makes several mistakes, and is not accurate in predicting exact expression levels (except in cases which are most likely due simply to luck). Again, the lower enhancer contribution figure does show accurate predictions being made with disjoint subsets of enhancers, in E058 and E065 (the two leftmost highly-predicted genes).

2.5.2 Measuring Enhancer Contribution

We can assess the relative individual contributions of enhancers under this model by subtracting their contribution from the predictions and comparing the modified prediction to the original. The modified prediction (E') for gene expression leaving out enhancer j is therefore:

$$E'_{PRed} = E_{basal} + S_{LRR} * \sum_{1}^{n} Ac_i - S_{LRR} Ac_j$$

From here we can then calculate the difference in correlation, should the contribution of that enhancer be removed from the model. This is similar to leave-one-out testing used to assess the contribution of individual features in complex, hard-to-interpret models such as random forests.

$$\Delta r = Cor(E_{Actual}, E_{Pred}) - Cor(E_{Actual}, E'_{Pred})$$

In theory, this represents the effect on gene expression of removing a single enhancer. For genes under long-range regulation, this should make a large difference, whereas for genes which do not respond to enhancers, this would make little difference. Therefore, the fit between the model and observed gene expression should decrease much more for LRR genes, measured as the change in correlation (Δr). Note that I have not re-calculated the predictive equation, as would be normal in other circumstances, such as calculating variable significance in a machine learning task. This is for the same reasons as stated earlier: I am not trying to select the most predictive enhancers, but

estimate the contribution of each. Since enhancers often appear redundant (Hong et al., 2008; Cannavò et al., 2016), interpretation would be difficult.

Table 2.5: Cell types used in the experiment for RNA-seq and Huston Acetylation data

ID	Name	Anatomy	Type
E003	H1 Cells	ESC	Primary Culture
E004	H1 BMP4 Mesendoderm	ES-deriv	ESC Derived
E005	H1 BMP4 Trophoblast	ES-deriv	ESC Derived
E006	H1 Derived Mesenchymal	ES-deriv	ESC Derived
E007	H1 Derived Neuronal Progenitors	ES-deriv	ESC Derived
E011	CD184+ Endoderm	ES-deriv	ESC Derived
E012	CD56+ Ectoderm	ES-deriv	ESC Derived
E013	CD56+ Mesoderm	ES-deriv	ESC Derived
E016	HUES64 Cells	ESC	Primary Culture
E037	Primary T helper	Blood & T-cell	Primary Cell
E038	Primary T helper naive	Blood & T-cell	Primary Cell
E047	Primary T CD8+ naive	Blood & T-cell	Primary Cell
E050	Primary hematopoietic stem cells	HSC & B-cell	Primary Cell
E055	Foreskin Fibroblast	Epithelial	Primary Culture
E056	Foreskin Fibroblast	Epithelial	Primary Culture
E058	Foreskin Keratinocyte	Epithelial	Primary Culture
E059	Foreskin Melanocyte	Epithelial	Primary Culture
E061	Foreskin Melanocyte	Epithelial	Primary Culture
E062	Primary Mononuclear Cells	Blood & T-cell	Primary Cell
E065	Aorta	Heart	Primary Tissue
E066	Liver	Other	Primary Tissue
E071	Brain Hippocampus Middle	Brain	Primary Tissue
E079	Esophagus	Digestive	Primary Tissue
E084	Fetal Intestine Large	Digestive	Primary Tissue
E085	Fetal Intestine Small	Digestive	Primary Tissue
E087	Pancreatic Islets	Other	Primary Tissue
E094	Gastric	Digestive	Primary Tissue
E095	Left Ventricle	Heart	Primary Tissue
E096	Lung	Other	Primary Tissue
E097	Ovary	Other	Primary Tissue
E098	Pancreas	Other	Primary Tissue
E100	Psoas Muscle	Muscle	Primary Tissue
E104	Right Atrium	Heart	Primary Tissue
E105	Right Ventricle	Heart	Primary Tissue
E106	Sigmoid Colon	Digestive	Primary Tissue
E109	Small Intestine	Digestive	Primary Tissue
E112	Thymus	Thymus	Primary Tissue
E113	Spleen	Other	Primary Tissue

Table 2.6: Summary of data sources used in the experiment

Data Type	Total	Post-filtering	Data Source
Enhancers	486902	382021	H3K27ac
Genes	19443	19443	RNA-seq
TADs	3117	3117	Hi-C

Chapter 3

The heatmaps.R Package for R/Bioconductor

3.1 Introduction

Data visualisation is an important tool at all stages of the scientific process. It is frequently used at the start of an experiment, for quality assurance; in early stages, as a means of generating hypotheses; and, of course, in clearly showing the results of a study in an effective manner. A graph, or image, is almost always preferable to a table or description conveying the same information.

Visualisation is most important in cases where it is difficult to neatly summarise the data in a statistical form. This can be in the preliminary stages, when it is not known if the data conform reasonably to a given model (see Anscombe's warnings for statisticians who fail to graph their data (Anscombe, 1973)), or as a first look to see if it is worth investing time in generating a complex model. In some cases, however, patterns in the data that are easy to spot with the human eye can be exceedingly tricky to model accurately. This applies particularly to spatial patterns in data. For example, CAGE data is often represented by clusters of nearby TSSs (see Introduction), but this is a simplification, and often more features are visible at the level of individual loci.

The term 'heatmap' confusingly refers to several plots commonly used in genomics. The term in general refer any bivariate plot in which x and y are independent variables, which can categorical or spatial, and the value of z is mapped to a colour. The first, and arguably more common usage, is in plotting correlations between many variables, where each correlation is represented by the colour (or shading) of the panel at the confluence of two variables. Another use is for a plot which displays data from genomics experiments over many separate windows as a single plot. For example, a heatmap could show the histone H3 Lysine 27 acetylation levels at a set of enhancers, sorted by the strength of the signal. This directly visualises spatial patterns in the data, and often combines multiple experiments (for example, different histone modifications) to show common patterns. For the rest of this chapter, 'heatmap' refers exclusively to the second type of plot. These plots can provide valuable insights which would difficult to discover through summary statistics alone. However, their usage is perhaps not as common in the literature as it should be.

3.1.1 Existing Work

Packages do already exist for producing a wide variety of heatmap plots, from diverse data sources. Genomation (Akalin et al., 2015) produces plots very easily, but lacks an easy way to include user-defined data. EnrichedHeatmap (Gu, 2017) produces complex plots, combining many different panels, but this process is tricky and the results would generally require editing before inclusion in a manuscript. 'ngs.plot' (Shen et al., 2014) produces many useful metrics for analysing NGS data alongside the ability to create heatmaps, but the plots produced are not of high quality. 'deepTools' (Ramírez et al., 2016) provides a comprehensive suite for analysing NGS data, producing attractive multi-panel figures, but is embedded with the Galaxy platform (Goecks et al., 2010), making use difficult for those who use R/Bioconductor. In addition, none of these packages provide functions for manipulation of biological sequence, which I believe is an important factor to consider in many analyses.

'seqPattern' (Haberle, 2015) provides functions to manipulate sequence data, but does not provide tools for plotting NGS data in the same package. The plots are effective, since the package provides smooths the raw image (a binary matrix of hits), allowing easier visual identification of patterns in the data. However, they are always produced as individual files so multi-panel figures require combining in an editor to visualise more than one pattern simultaneously.

In summary, many packages effectively plot heatmaps, and taken together cover almost all potential use cases, and a wide variety of plot styles. However, this means that users must choose, and learn, different packages for working with different kinds of data. Beyond mere inconvenience, this complicates data integration considerably: for example, plotting transcription factor binding sites alongside ChIP-seq data for the same factor would involve the use of multiple packages.

3.1.2 Requirements

I set out to write a new package in R/Bioconductor to fulfil what I perceived as a gap in the currently available packages. The core requirement is to have a package that enables data integration, particularly between functional genomics and sequence-led approaches. There are also a number of other concerns that are common to most plotting packages, which I have outline below.

The ideal for any plotting package is to produce publication-ready plots programmatically, removing the need for extensive modification using image-manipulation programs and encouraging rapid exploration and validation of new hypotheses. This requires effort on the part of the package author to ensure that layouts, text scaling and colour schemes are chosen effectively. In addition, the package must be flexible, so the user is not forced into particular design choices by the author. This can be a problem for packages that focus on initial ease of use: while they can produce one particular design of figure with very small amount of code, modifying this design is not possible.

Creating multi-panel figures programmatically (i.e., without collating im-

ages after plotting using an external program such as Adobe Illustrator) is also an important goal, since visualising multiple sources of data alongside each other is a standard use case for heatmaps. This functionality is available using the graphics engine in R, but control of these functions is not always available to the user.

This flexibility should also extend to the input data. Functions should be provided for common operations, but it should also be simple to plot arbitrary data, in case certain use cases are not anticipated. Visualising biological sequence data requires functions to interrogate biological sequences. This includes both exact matching of sequences, and scanning sequences using PWMs. In addition, the results of these analyses need to be processed to be displayed informatively. which most easily achieved by smoothing the results of pattern matching.

Any program dealing with the next-generation sequencing data, or the results of genome-wide analyses needs to be efficiently written, so that simple operations do not take large amounts of time to complete. Writing the package in R (Ihaka and Gentleman, 1996), and using the packages and software infrastructure for manipulating genomic sequence and annotations available through the Bioconductor platform (Gentleman et al., 2004; Huber et al., 2015) provides fast performance for bioinformatics workflows.

3.2 Results

3.2.1 Structure of the heatmaps.R packages

The core philosophy of the heatmaps package, that will help to realise these ideas, is that the heatmaps class should only try to represent the plot itself, not the underlying data. This has a number of advantages.

Most importantly, this allows the raw input data, potentially spanning megabases of sequence, to be compressed down into a smaller, representative image, making plotting much faster. This allows new hypotheses to be tested more easily, and is invaluable for combining plots into multi-panel figures and experimenting with formatting. While this may seem obvious, some packages

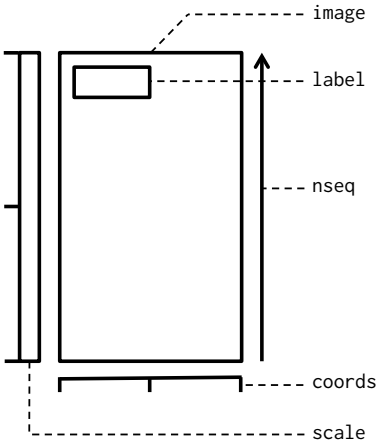

Figure 3.1: The structure of a heatmap.

tie data extraction to plotting so that even small graphical changes require considerable data processing. See the Performance section for benchmarks.

It also provides easier manipulation of the data, since the image is represented by a single matrix. This makes it simple to apply arbitrary transformations to the data, such as smoothing, logarithm or winsorization (capping values at a certain quantile), without losing the plotting abilities of the package, or having to 'hack' the package's source code.

For programming purposes, a heatmap is just an image matrix (`image`), with metadata. I based my packages around a central 'Heatmap' class, which would provide a self-contained data structure with the all the information needed to produce a plot. A heatmap needs co-ordinates for the X-axis (`coords`), and to keep track of the number of rows in the image for Y-axis co-ordinates (`nseq`). Since we might want to normalise the values between images, there is also a slot to store the numerical scale used for plotting (`scale`). Finally, there is also a label field (`label`). Figure 3.1 illustrates this structure.

3.2.2 Workflow

The raw data required to plot a heatmap consist of a set of windows onto the genome (Figure 3.2a), representing the regions to be plotted, and the

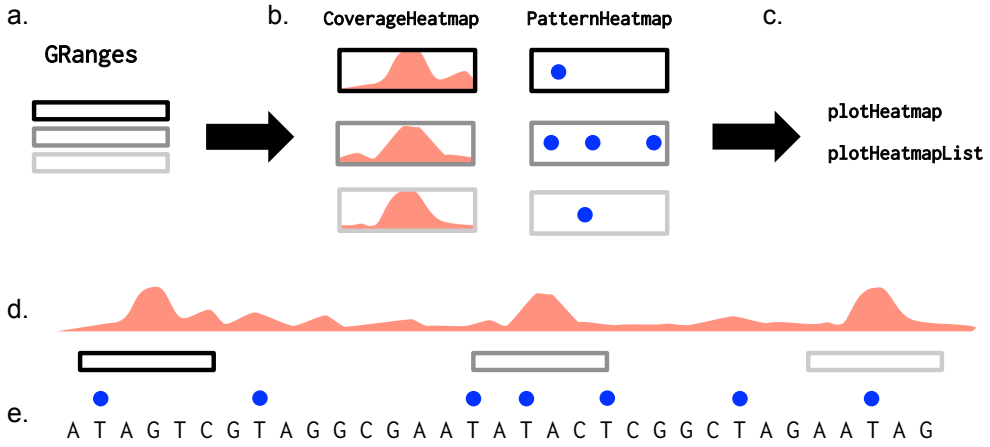

Figure 3.2: The workflow involved in creating a heatmap. Predefined genomic intervals (a) and used to select regions of the genome and retrieve data on functional genomics experiments (d) or sequence data (e), combined into matrices (b) and then displayed using functions in the package (c).

annotations to be plotted. These generally takes the form of a linear track along the genome (Figure 3.2d), such as would be displayed in a genome browser, or annotation of the underlying sequence (Figure 3.2e). These are then combined into heatmaps (Figure 3.2b), which can display any data, and can be plotted with the `plotHeatmap` or `plotHeatmapList` functions (Figure 3.2c).

Bioconductor provides libraries which allow efficient storage and manipulation of genomics data. This includes reading in the data represented in panels d and e, and combining this linear data into matrices as shown in panel c (Figure 3.2). These manipulations are not always simple for the user, however, and in any programming environment, small mistakes in code can lead to hundred-fold decreases in speed. The R language is particularly vulnerable to this, since operations carried out by the interpreter are many orders of magnitude slower than those delegated to compiled subroutines. The heatmaps package includes high-level functions for many common operations which remove obstacles from the user, so results are always generated in the fastest possible time.

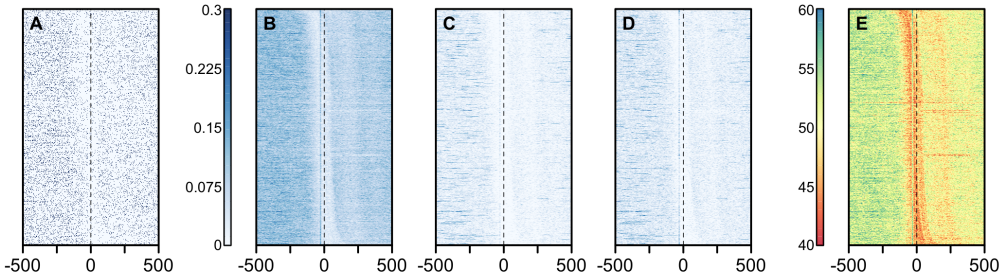

Figure 3.3: The sequence of Zebrafish promoters at 24h. a. Matches to **TA** visualised at dots. b. Matches to **TA** visualised by smoothing out those dots. c. Exact matches to **TATA**, smoothed. d. Matches to the consensus TATA-box PWM. e. Visualisation of the TATA-box PWM score at every point across the loci.

3.2.3 Visualising Sequence Data

Sequence data is rarely displayed directly, as raw ACGTs, as the resulting figures would be all but impossible to interpret. Sequence features are usually summarised using either k-mers or PWMs. k-mers are exact matches with a specified sequence of k base pairs. These are generally short, since longer sequences become increasingly rare. They often include degenerate base pairs, specified using the IUPAC ambiguity codes: **Y** for a pyrimidine (either an **A** or **G**), a **W** for 'weak' bases (**A** or **T**), **S** for 'strong' bases (**C** or **G**), or an **N** for any base at all. Longer or more complex features are commonly specified as PWMs, since they allow for probabilistic matches. Both of these inputs can be used to create heatmaps directly, either from sequence data input by the user or fetched automatically from Bioconductor's genome packages (`BSgenome`, Pagès et al. (2017)).

Exact matches to k-mers, or thresholded matches to PWMs, produce binary data. It is difficult for the eye to see patterns in binary data represented as a series of dots on the page. In addition, no figure will have sufficient resolution to accurately visualise individual base pairs within kilobases of sequence. Applying a smoothing function to this binary data produces an image with continuous values representing the local density of matches, a far more effective way of displaying sequence data. This also solves the

resolution problem.

Figure 3.3 shows all of these methods, applied to a set of zebrafish promoters defined by CAGE (see Methods). In order to investigate the presence of the TATA motif at promoters, I have looked at exact matches to **TA** and **TATA**, PWM matches at over 80% of the maximum score, and additionally the PWM score over each locus.

Figure 3.3a shows the raw data for an exact match to **TA**, before smoothing. Each dot represents a match to the exact pattern. While all the information is displayed here, it is very difficult for the eye to make out the patterns. Figure 3.3b shows this data smoothed using kernel smoothing (see Smoothing section for details), which is more easy to interpret. Figure 3.3c shows an exact match to the extended sequence, **TATA**, which is has markedly fewer matches, even at the expected TATA-box location. By comparing panels b and c, it appears that many TATA-box motifs actually do not contain the exact TATA sequence.

Figure 3.3d shows sites where the PWM score exceeds a specified threshold. In this case, the threshold is 80% of the maximum possible score. This is more sensitive than the exact matching method. This is expected, since there are a greater number of informative bases in the TATA motif, allowing for greater flexibility in matching true TATA-boxes.

Figure 3.3e shows how this value varies across each window. This is a `PWMScan` heatmap, where the individual log-likelihood of a match is evaluated at each point in the sequence and displayed according to a colour legend. This second plot is more informative, since it not only shows the enrichment of TATA at the expected location, but clearly shows exclusion outside this region.

3.2.4 Visualising Functional Genomics Data

Functional genomics data is often specified as a linear track along the genome, for example counting the number of ChIP-seq reads mapping to each location along the genome. This is handled in Bioconductor by the `RleList` class, which in run-length encoded, so that runs of repeated values do not take

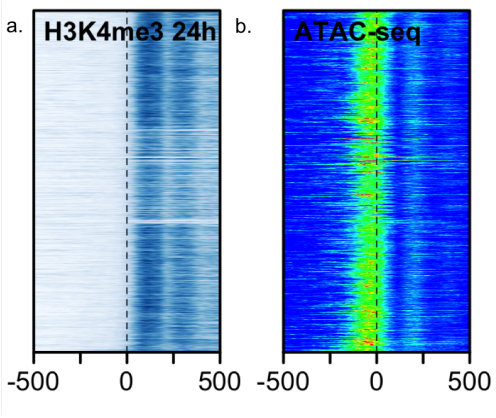

Figure 3.4: a. H3K4me3 ChIP-seq signal at a subset of Zebrafish 24h promoters. b. ATAC-seq signal at the same set of promoters.

additional memory. This data can be directly converted to a heatmap by intersecting the track with a `GenomicRanges` object. Any data that can be displayed as a continuous track along the genome can be handled in this way, as can binary data specified as ranges, such as CpG islands. File handling is not provided by the package, since Bioconductor already provides methods for reading in almost all commonly used formats.

Figure 3.4 shows an example of functional genomics data plotted at the same promoters as Figure 3.3. The panel a shows ChIP-seq for H3K4me3, a histone modification associated with active promoters. The two nucleosomes downstream of each promoter are clearly visible as the source of the signal. The panel b shows ATAC-seq, an assay for accessible chromatin. The nucleosome free region upstream of the dominant TSS is visible as an enrichment in ATAC fragments, which is also visible, but much less clear, around +200bp, between the nucleosomes. Both panels have been smoothed, and the ATAC-seq signal has been cut off at 80% of the maximum observed, so a greater dynamic range can be visualised. This is easily accomplished by modifying the `scale` property in a heatmap, rather than changing the original data.

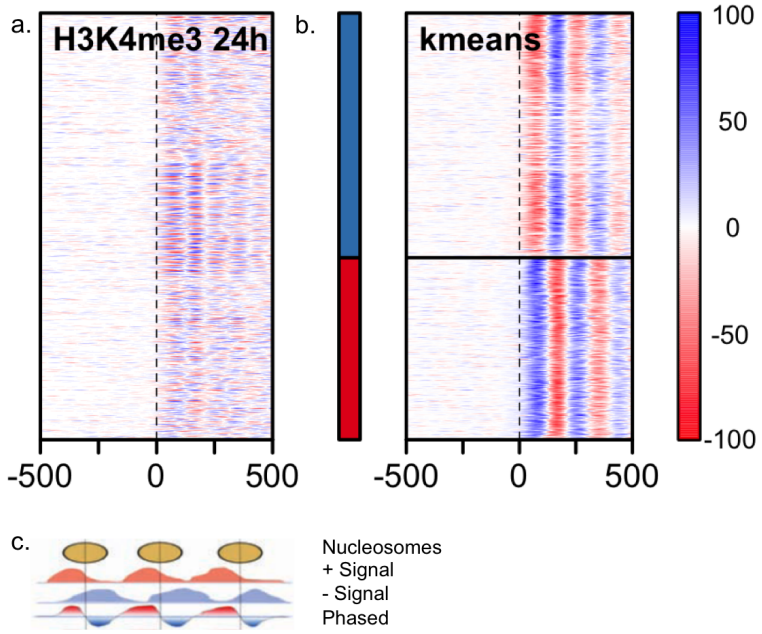

Figure 3.5: a. H3K4me3 ChIP-seq at 24h Zebrafish promoters. The value displayed is the coverage on the plus strand with the negative strand coverage subtracted. b. The same signal, but following k-means clustering. c. An illustration of how the phased ChIP-seq signal is calculated.

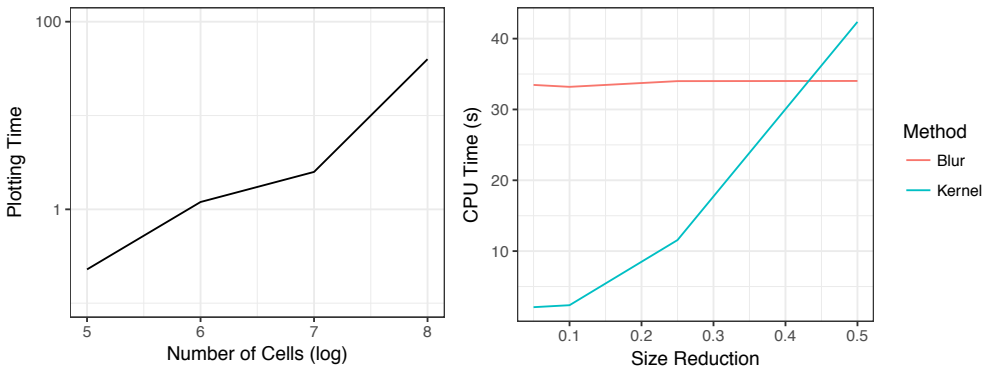

Figure 3.6: a. Time taken to plot a heatmap, in seconds, versus the number of pixels in the image. b. The time taken to smooth and downsample a 40 Megapixel image, based on the smoothing method used and the degree of downsampling.

3.2.5 Clustering

The package provides a method for displaying clusters in the data, although it does not provide clustering methods. This is a deliberate choice, because clustering is a complex process and limited, built-in routines may not be the most appropriate choice for many examples. Instead of implementing many common clustering algorithms, users are free to cluster their data using any approach, and `heatmaps.r` will display the clusters they have defined. This also allows users to display clusters derived from external data, such as prior annotation, for example differentially expressed genes or ChIP-seq peaks.

It is possible to visualise nucleosome positioning by subtracting plus-strand ChIP-seq reads from negative-strand ChIP-seq reads (Figure 3.5c), rather than using ATAC-seq data. This is shown in Figure 3.5a, however the signal is unclear. Figure 3.5b shows the same data following kmeans clustering, which reveals the pattern in the data. This is actually equivalent to clustering based on positive and negative strand promoters, which is a somewhat contrived example, but illustrates the functionality of the package.

3.2.6 Performance

Performance is a critical concern for plotting heatmaps. A heatmap covering 20,000 windows of 10kb each contains 200 million individual observations, which approaches gigabyte of data. This makes it very difficult to perform operations on such data in memory on anything other than a high-performance cluster. Keeping operations in memory is desirable, since it is much faster and simplifies programming. Even on such machines, the time taken to perform even basic operations can exceed the patience of a researcher carryout out interactive data analysis. The plotting of these figures also takes a significant amount of time, which can be particularly frustrating when trying to change cosmetic details or comparing multiple plots.

The package takes several steps to combat performance concerns. With images of the size described above, it is not possible for practical purposes to view the figure at full resolution, either on the page or on screen. Therefore, with the proper tools to reduce the size of these images, performance can be dramatically increased without suffering any loss in final image quality. In fact, using the package it is possible to perform such analysis easily on most recent commercially available laptops. The standard approach to downsizing images used in image processing is to first smooth the data, and then reduce the size, referred to as downsampling in image processing. The smoothing avoids unwanted artefacts created by the downsampling process, which can adversely affect the quality of the final image.

Figure 3.6a shows that the plotting time increases roughly linearly with the number of pixels in an image. As we can see, larger plots can take up to minute to plot. However, 10^6 pixels, which plots very quickly, is a high enough resolution to fill an average figure panel at the 300dpi resolution required by most journals. Performance would deteriorate further at larger sizes, since the dataset would increase beyond the size easily stored in RAM.

3.2.7 Smoothing

Two smoothing methods are included in `heatmaps.r`. These are vital for displaying sequence data, as shown above, and are also useful in plotting

functional genomics data.

The first is a binned kernel density smoothing, which operates on binary data. Kernel smoothing replaces individual points in data with a continuous-valued function, a "kernel". In this case, a Gaussian kernel is used. The implementation is taken from the KernSmooth (Wand, 2015) package, using a binned kernel density estimator as described in Wand (1994). The *binned* approach is key, because it computes the kernel density estimate only for a subset of points. This provides both downsampling and smoothing for binary data in a single step, and so under these conditions it is faster than the alternative Gaussian blur function (Figure 3.6b) when downsampling is required.

The Gaussian blur function provides smoothing for continuous-valued data, such as ChIP-seq. The implementation is provided by the EBImage (Pau et al., 2010) package. This is slower than the kernel smoothing method when downsampling is also required (Figure 3.6), but works on any data, and is faster in some cases when used on large images with no downsampling (Figure 3.6). Another option provided in `heatmaps.r` is to create the heatmap using binned data. This would also reduce the resulting size by a significant margin straightforwardly: if one value is calculated every 10bp rather than for every point, then the resulting image is a tenth of the size.

3.2.8 Plotting Options

In order to provide the maximum flexibility for users, I have provided two separate interfaces for drawing heatmaps. The more basic function, `plotHeatmap`, is responsible for drawing the image itself, and all the extra features required, such as axis ticks or labels. I have included as many options as is sensible so that these plots can be fully customised, and a selection of these is outlined in Table 3.2.8. However, the `plotHeatmap` is designed so that additional data or labels can be plotted by the user, whilst keeping the desired functionality of package. This could include highlighting specific features on the plot, which is aided by the fact that the co-ordinates used by `plotHeatmap` track the original windows used to create the plot, even after downsampling or smoothing.

Option	Function
color	Specifies the colour scale
label	Plot a label, e.g. "TATA"
label.xpos, ypos	Position the label precisely
label.col	Colour for label, useful for dark plots
legend	Include a legend indicating numerical values
partition	Specify ration of cluster sizes
partition.lines	Plot lines delineating clusters
box.width	Width of the box around the heatmap
refline	Draw dashed line at $x = 0$

Table 3.1: Example options available for heatmaps

Therefore, +200bp in the 1000th sequence in the list is represented by the point $x = 200$, $y = 1000$.

The other function provided is `plotHeatmapList`. This is a wrapper for `plotHeatmap` which also controls the layout of panels for multiple figures, as well as the margins and legends. This means that users can plot multiple heatmaps in one figure, without having to control the arrangement themselves. All of the options available to `plotHeatmap` can be passed to `plotHeatmapList`, and separate options can be passed to each figure by specifying these options as a list. `plotHeatmapList` also normalises figures that are grouped together, so separate datasets can be compared quantitatively.

3.2.9 Multi-panel Plots

A central aim of the package was to produce complex, multi-panel figures directly from R that would be suitable, with minimal modification, to appear as journal figures or in presentations. One such example, Figure 3.7, is shown below. This figure is reproduced from the central figure of Haberle et al. (2014). The arrangement of the panels, the labelling of individual features and the labelling of axes is all done directly from R. The figure shows the same promoters as in Figure 3.3, but aligned to the dominant TSS (as defined by CAGE) at two different time points during development, before and after zygotic genome activation (ZGA). The differences in promoter architecture

at each time point are visible from the dinucleotide frequencies: the maternal TSS in defined by a strong TA signal slightly upstream of the promoter (a TATA-box motif), and the zygotic TSS is defined by a broader band of CG enrichment.

3.3 Discussion

I have written a package, heatmaps, in R which fulfils all of the requirements I outlined in the introduction. It provides plotting methods for both next-generation sequencing data, such as ChIP-seq, alongside methods for plotting DNA sequence features such as exact k-mer matches and PWMs. To the best of my knowledge, this is the first package to include both these features "out of the box". All of this is built on a flexible environment which allows easy creation of multi-panel figures, including normalising between panels and plotting clusters in the data.

Additionally, the package allows users to control lower-level plotting functions, which simplify the creation of complex plots. This is targeted at those users whose needs are not met by the basic functions of the package. All of the data manipulation performed by the package uses Bioconductor infrastructure to maximise efficiency.

The package was accepted into Bioconductor (Version 3.5), which shows that it meets the rigorous quality requirements of Bioconductor (Bioconductor Core, 2017). These include guidelines for correct use of R classes, "robust and efficient code", and documentation covering all aspects of the package. I hope that, with the acceptance of the package into Bioconductor, more researchers will include heatmaps as part of their regular workflow, and in particular include sequence features in their analysis. I believe that this method of looking at biological data has advantages, and that heatmaps are currently under-utilised by the genomics community, and that my package can help to address this.

The heatmaps package removes many of the programming challenges involved in creating heatmaps. Even for experienced bioinformaticians, making these plots can be time-taking exercise with no code readily available,

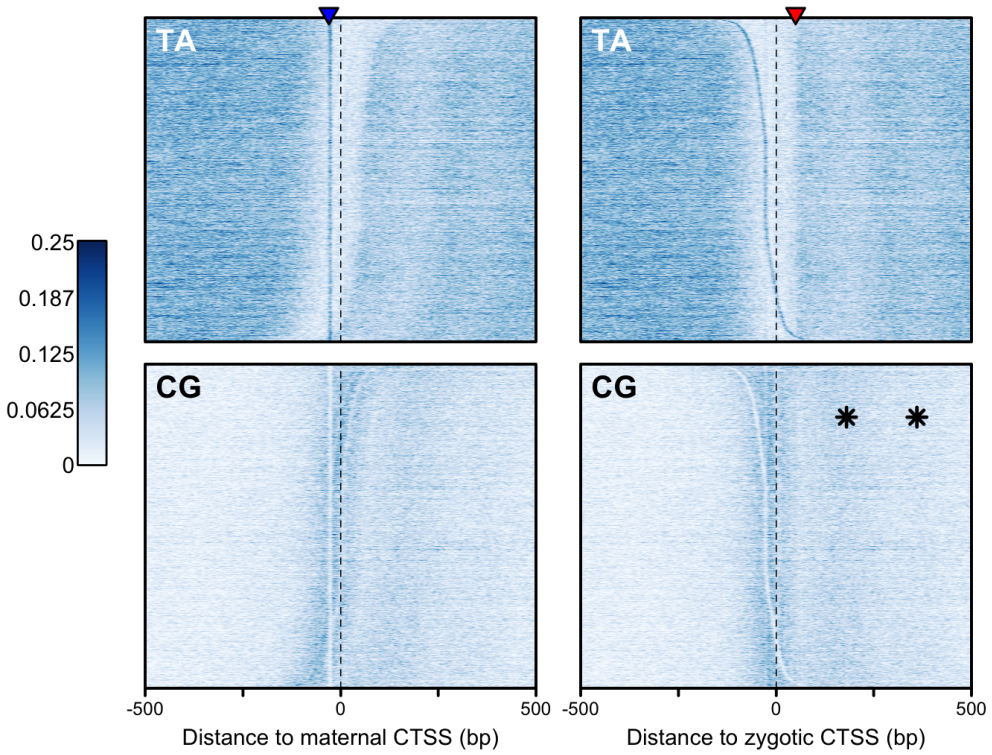

Figure 3.7: TA and CG frequencies around 8369 zebrafish promoters, aligned to CAGE transcription start sites for maternal transcripts (left) and zygotic transcripts (right). The blue arrowhead indicates the TATA-box motif. The red arrowhead indicates the CG-TA boundary. Asterisks indicate GC enrichment between nucleosomes.

for reasons outlined below. For primarily experimental researchers, this may be a step too far unless clearly demanded by the data. For any user, having a fully-featured package which handles the technical difficulties in making these plots will encourage quick experimentation with different hypotheses, and yield better-looking final figures.

The package also encourages users to look for spatial patterns in the data, particularly sequence data. This aspect of biological data analysis is often overlooked in favour of summary statistics. However, many common analysis methods are not capable to detecting certain patterns. ChIP-seq data is generally analysed using peak callers, which may ignore local depletions inside of larger peaks, which can be around important features. PWM motif enrichment, the most common sequence analysis tool, does not capture patterns of spatial enrichment for shorter sequences, such as those shown in Figure 3.7. The k-mers themselves may not be enriched overall in the sample, but patterns are clearly visible at the level of individual loci. In effect, it is possible to see patterns by eye which are difficult to detect with statistical methods. It should be noted that revealing these patterns often takes significant insight or intuition on the part of the researcher, since the correct ordering of windows, or selection of features, is paramount. This is something of a double-edged sword, since results based purely on visual patterns, rather than quantified data, may be seen as less rigorous. However, I believe that, until methods are developed to accurately capture these patterns, data visualisation remains a powerful tool for studying biological systems.

3.4 Methods

3.4.1 Zebrafish Promoters

Zebrafish promoters were taken from the set of shifting promoters in Haberle et al. (2014). Except where otherwise noted, they were aligned by the maternal dominant TSS, as defined by CAGE. H3K4me3 data were used from the same paper. ATAC-seq data were taken from Gehrke et al. (2015).

3.4.2 Pattern Matching

The TATA-box motif was taken from the JASPAR Vertebrate database (Mathelier et al., 2016).

Chapter 4

Amphioxus Promoterome

4.1 Declaration

Amphioxus tissue isolation, RNA extraction and genomic DNA extraction were performed by collaborators at the CABD institute in Seville. Analysis of genomic DNA and RNA-seq was performed by collaborators in Barcelona, led by Manuel Irimia and Ferdinand Marelatz. I have used both the genome assembly and the predicted gene models from RNA-seq, including homology prediction based on human proteins.

CAGE library preparation and sequencing were performed at the RIKEN institute in Japan. All further analyses were performed by me.

4.2 Introduction

Despite their central importance to gene regulation, it is still not known precisely what determines promoter activity at the sequence level in higher organisms. We have many reasons to believe that key aspects of promoter function are conserved across large evolutionary distances: deeply conserved enhancers, transcription factors, promoter motifs, and the conservation of the transcriptional machinery itself. Despite this, promoter features vary significantly between different clades (Lenhard et al., 2012). If we can pinpoint which DNA sequence features are conserved, they might provide clues to the mechanisms which underlie shared promoter biology across metazoa.

Branchiostoma lanceolatum, commonly known as the European amphioxus, is a small (up to 6cm) chordate which is local to the North East Atlantic and the seas around Europe. The amphioxus lineage diverged from vertebrates close to the root of the vertebrate clade. For a long time amphioxus was thought to be the closest extant relative of the vertebrates, but molecular methods have placed the Tunicate branch closer (Delsuc et al., 2006). Tunicates share little in morphology with vertebrates outside their larval stage, with adults taking on astonishingly diverse morphologies. Amphioxus shares many morphological features with vertebrates, and so may be more representative of the vertebrate last common ancestor. These features make amphioxus an attractive species for investigating vertebrate evolution, since it could allow us to distinguish vertebrate-specific innovations from earlier evolutionary changes.

In this chapter, I use CAGE data to investigate promoter architecture in the European amphioxus. This will identify promoter features that are common between amphioxus and vertabrates, features that are shared between all metazoa and lineage-specific features in amphioxus.

4.3 Results

4.3.1 Data

RNA was isolated from 7 amphioxus developmental stages and tissues by collaborators working at the CADB in Seville. In collaboration with RIKEN, we performed Cap Analysis of Gene Expression followed by sequencing (CAGE-seq, (Shiraki et al., 2003)), on these samples, producing genome-wide maps of transcription initiation at single base-pair resolution. Whole-embryo RNA was isolated from amphioxus at the 32 cell stage, 8 hours and 15 hours post fertilisation, and also for the neural tube, hepatic system, muscle and female gonads.

4.3.2 CAGE Data Processing

CAGE tags were aligned using Bowtie and processed using CAGEr (see Methods for details). I aggregated signal at individual nucleotides into tag clusters (TCs), which represent the range of initiation for a single promoter. Each promoter is also assigned a dominant transcription start site (TSS), the nucleotide at which initiation is most frequent. The widths of these clusters form a bimodal distribution in six samples 4.1a, which shows that amphioxus has the expected mixture of broad and sharp promoters as seen in other Metazoa (Carninci et al., 2006; Hoskins et al., 2011). This was not observed in the muscle sample, which means that many of the clusters observed in the data are likely to be artefacts; as such, the muscle sample was excluded from further analysis. Figure 4.1c illustrates how sharp and broad promoters appear at individual loci. To increase the robustness of this measurement, I used the distance between the 10th and 90th percentiles of initiation, known as the inter-quantile (or IQ) range. The data from the three embryonic stages, as well as female gonads, neural tube and hepatic system, passed quality control and were analysed further.

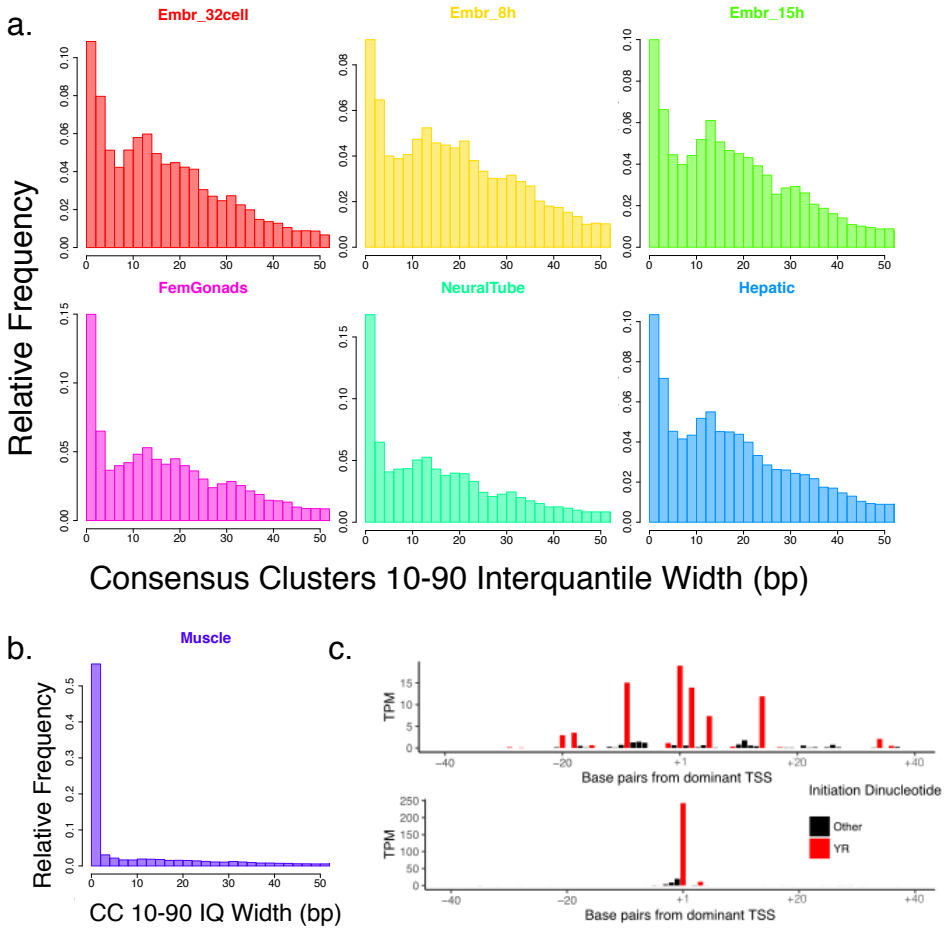

Figure 4.1: **a.** CAGE Clusters show a bimodal distribution in width between "broad" and "sharp" promoters, which have characteristic patterns of initiation in six samples. **b.** This distribution is not observed in the muscle sample. **c.** Top: CAGE tags in Embryo 15h from the region Sc0000001:7714823-7714902. Broad promoters initiate from multitude of sites over dozens of base pairs. Bottom: CAGE tags in Embryo 15h from the region Sc0000240:29994-30073. Sharp promoters usually initiate dominantly from a single base pair. In both cases, the strongest initiation nucleotides are purines (R) that are immediately preceded by a pyrimidine (Y).

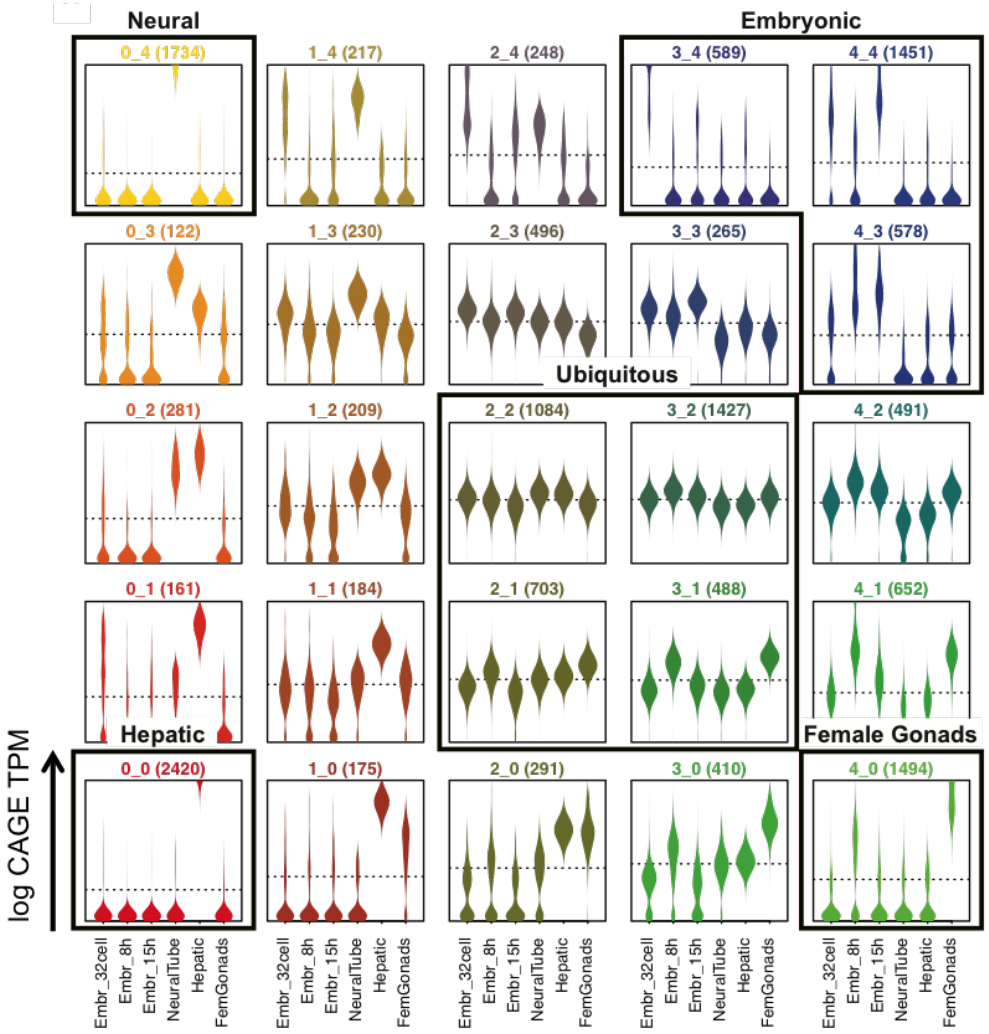

Figure 4.2: Consensus Clusters arranged in a self-organising map (SOM) according to their expression between samples, with representative clusters chosen for further analysis highlighted. Each box represents one SOM cluster, with series of beanplots showing distribution of scaled expression values (logarithm of normalised number of CAGE tags per million) at different time points for all promoters belonging to that cluster. The dotted line indicates the mean expression. The co-ordinates above each cluster illustrate the position of each cluster within the SOM, as spatial arrangement is significant for this method. The numbers in brackets indicated the number of promoters in each cluster.

4.3.3 Categorising Promoters by Expression Profile

Each box represents one SOM cluster, with series of beanplots showing distribution of scaled expression values (logarithm of normalised number of CAGE tags per million) at different time points for all promoters belonging to that cluster.

Tissue specificity is a major determinant of promoter architecture in Metazoa (Carninci et al., 2006; Hoskins et al., 2011). To compare gene expression between different developmental stages and tissues, TCs were aggregated between stages to form consensus clusters. This is necessary as the tag clusters from separate stages can have different coordinates, even if they represent the same gene promoter. Gene expression profiles across all stages were clustered using a self organising map (see 4.2), and I selected 5 representative clusters to analyse further: genes specifically expressed in the embryo, hepatic system, neural tube or female gonads, and genes expressed at stable levels across all samples (ubiquitous genes). The goal here was not to categorise all promoters, but to to find robust clusters to analyse further.

4.3.4 Promoter Analysis

I investigated the differences and similarities between these groups using a variety of data sources known to be important for promoters (see Introduction). Exact matches for short sequences of nucleotides, or k-mers, can neatly display general patterns in nucleotide content across regions. Matches to position weight matrices (PWMs) provided a more accurate picture of potential binding by transcription factors. The strength of a match can be measured in many ways, but here I used a threshold based on the maximum possible score. Gene models based on RNA-seq were used to show the first exons of each transcript. Interquantile (or IQ) range is a robust measure for the range of initiation at each promoter, and promoters were aligned by the dominant TSS, which is the single base pair with the largest number of transcripts initiating from it. This allows us to make precise spatial comparisons between promoters. Nucleosome occupancy was determined from NucleoATAC signal (see Methods), and is known to be important for promoter function (see In-

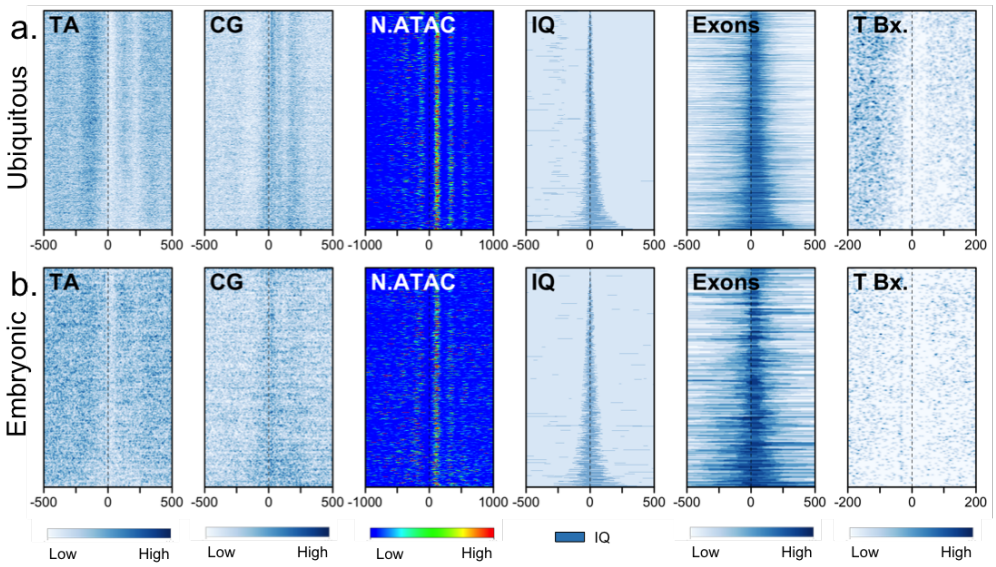

Figure 4.3: Heatmaps of the first two promoter clusters identified in by CAGE expression clustering, aligned by dominant TSS. **TA** and **CG** show the smoothed density of exact dinucleotide matches. N.ATAC is smoothed nucleosome occupancy in 15h Embryo. IQ is the 10-90 interquantile range of the consensus clusters. Exons are displayed as smoothed gene models predicted from RNA-seq. TATA and YY1 are matches to PWMs at 80% of the maximum score, again displayed as smoothed density. a. Ubiquitous cluster b. Embryonic

troduction). All of these data except IQ range are displayed following kernel smoothing (see Methods) to improve the ability to visualise overall trends in the data, and therefore the exact values displayed at any point in the image are arbitrary and dependent on the smoothing parameters. Interesting values are visualised more quantitatively in additional figures.

In the ubiquitously expressed cluster there is a well-defined band of **TA** dinucleotides upstream of gene promoters (Figure 4.3a, **TA**). Downstream of the dominant TSS, there are two periodic **GC** enrichment bands (Figure 4.3a **CG**), which become less clear at broader promoters. These promoters show strongly positioned nucleosomes at all widths. This contrasts with vertebrate promoters, which show strong **GC** and **CG** enrichment in the nucleosome free region from -60 to +40 (Carninci et al., 2006; Rach et al., 2011); in fact, this pattern is unique among species in which CAGE data is available. Additionally, broad promoters of increasing width are increasingly asymmetric, since their IQ range extends downstream from the most commonly used TSS position (Figure 4.3a), which is not observed in vertebrates.

The two GC enrichment bands are positioned either side of the +1 nucleosome. Closer inspection reveals repeating **WW** and **SS** dinucleotides at $\tilde{1}$0bp intervals, flanking the '+1' nucleosome (Figure 4.5a), which constitutes a very strong nucleosome positioning signal. It is also possible that the **TA** enrichment band upstream of the promoter contributes to positioning the -1 nucleosome.

Promoters that are specific to embryonic development, neural tube or the hepatic system share many features that contrast with ubiquitously expressed genes. They lack both the upstream **WW** enrichment band and the asymmetry of ubiquitously expressed promoters, and the precise +1 nucleosome positioning signal is not evident (Figures 4.3b, 4.4 TA and CG). Despite the loss of this signal, NucleoATAC shows that the nucleosome positioning is very similar to ubiquitous promoters in embryonic promoters (Figure 4.3b N.ATAC), where ATAC-seq is available.

Promoters specific to female gonads share features of both ubiquitous and tissue-specific genes. Based on the SOM clustering 4.2, they are closer to ubiquitous genes in terms of gene expression. There is some evidence of

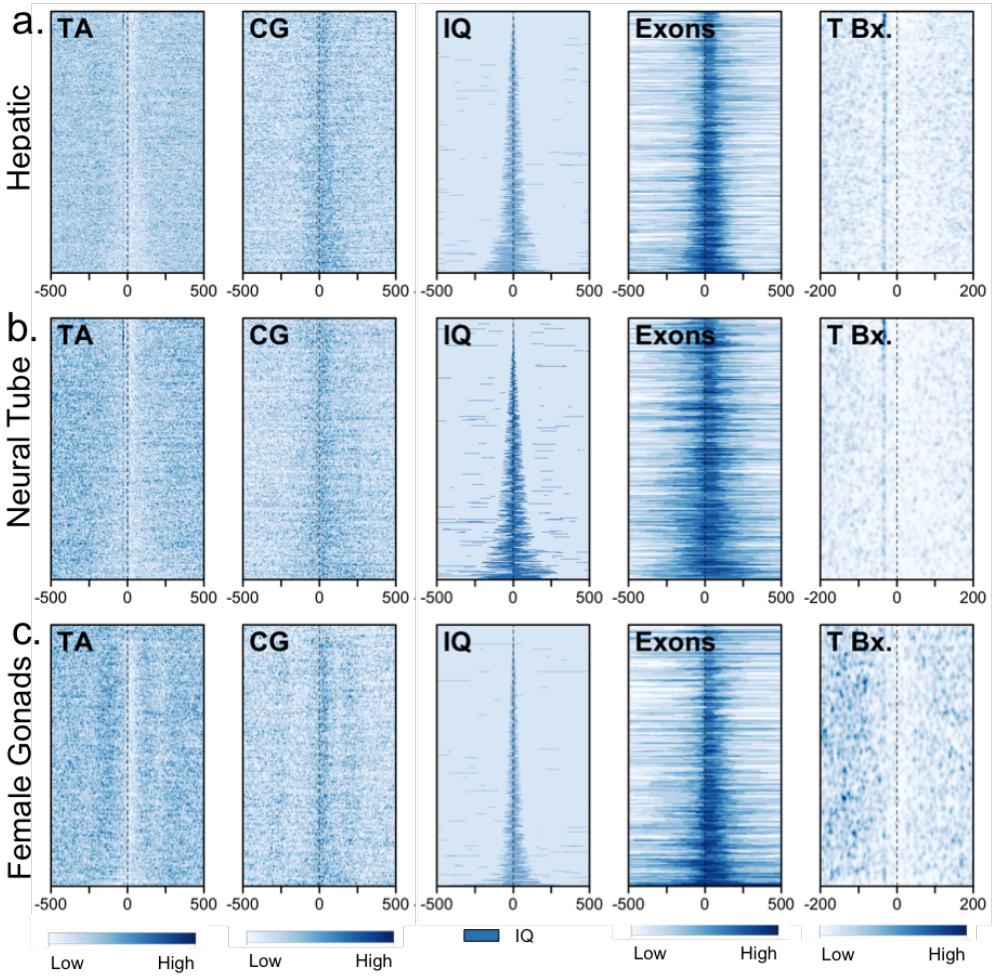

Figure 4.4: Heatmaps of the remaining promoter clusters identified in by CAGE expression clustering, aligned by dominant TSS. **TA** and **CG** show the smoothed density of exact dinucleotide matches. N.ATAC is smoothed nucleosome occupancy in 15h Embryo. IQ is the 10-90 interquantile range of the consensus clusters. Exons are displayed as smoothed gene models predicted from RNA-seq. TATA and YY1 are matches to PWMs at 80% of the maximum score, again displayed as smoothed density. a. Hepatic b. Neural Tube c. Female Gonads

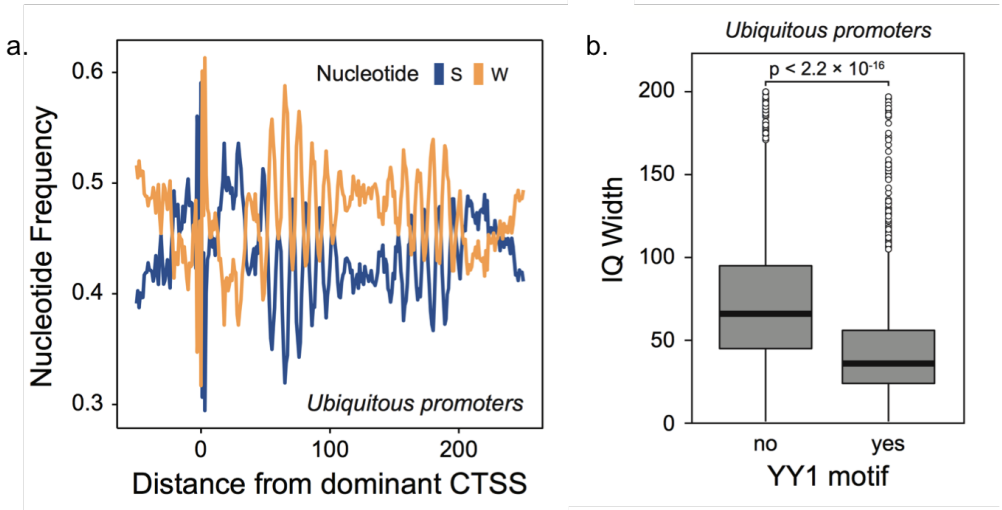

Figure 4.5: a. Meta-region plot visualising mono-nucleotide density (S and W) around the +1 nucleosome. b. IQ width of ubiquitous promoters containing a YY1 motif vs. those that do not.

the **TA** enrichment band and nucleosome-flanking GC bands, but not at the narrowest promoters, which may be have tissue-specific architecture.

4.3.5 De novo Motif Analysis

De novo motif analysis identified many motifs at tissue-specific promoters. These differed significantly between the tissue-specific clusters, and a number of motifs were recognisable by their similarity to tissue-specific vertebrate transcription factors, indicating that both transcription factor binding sites (TFBSs) and the roles of transcription factors (particularly those important for development) are highly conserved. However, no new core promoter motifs were discovered in any of the clusters.

The amphioxus homologue of Grainyhead, an important TF in early development, is found at over 300 promoters in the embryonic cluster, and the motif very strongly matches the motif in the database (Jolma et al., 2013). Neural tube promoters had a 140 very strong JDP2 (TGACGTCA) sites (with several positions invariant across all binding sites), as well as NFYA (CCAAT) motifs and YY1. Hepatic promoters contained many Hepatocyte

nuclear factor 4 alpha (HNF4) binding sites. Female Gonad promoters contain many NR2F1 binding sites, which is a steroid hormone receptor.

All three clusters contain a significant minority of TATA promoters, and these are concentrated at the top of the heatmap (Figures 4.3b, 4.4, TA), indicating that the sharpest promoters are frequently TATA-dependent like in other metazoan promoteromes (Lenhard et al., 2012). In general, however, they lack the fine structure seen in ubiquitous promoters, and it is not at all clear how the promoter is defined at the sequence level in the non-TATA promoters. No clear distinctions were observed in the heatmaps between promoters in the different clusters.

Many ubiquitous promoters are followed by a YY1 motif between 10 and 30 bp downstream from the dominant TSS (Figure 4.6a YY1), and YY1 motifs are concentrated at the top of the panel, indicating that YY1-containing promoters are also the narrowest. It appears that initiation is limited to the region upstream of the YY1 motif at YY1-containing promoters, as seen by the smaller IQ ranges of these promoters which do not extend far downstream of the dominant TSS (Figure 4.6a IQ, Figure 4.5b). These promoters also have shorter first exons (Figure 4.6a Exons), which suggests a role for YY1 in preventing ectopic initiation after the first splice junction. YY1 is also present at Female Gonad promoters, where they are also present at narrower promoters (Figure 4.6b YY1). However, YY1 promoters are not the narrowest: this may be due to YY1 promoters having ubiquitous promoter architecture, which is generally not as sharp.

The YY1 motif has previously been linked to the precise control of initiation of human LINE elements (Athanikar et al., 2004), and genes with short 5' untranslated regions (UTRs) where it may fulfil a similar role (Xi et al., 2007), but this is the first demonstration of these combined effects at non-repeat loci and the potential role of YY1 as a possible core promoter element throughout metazoa deserves further study.

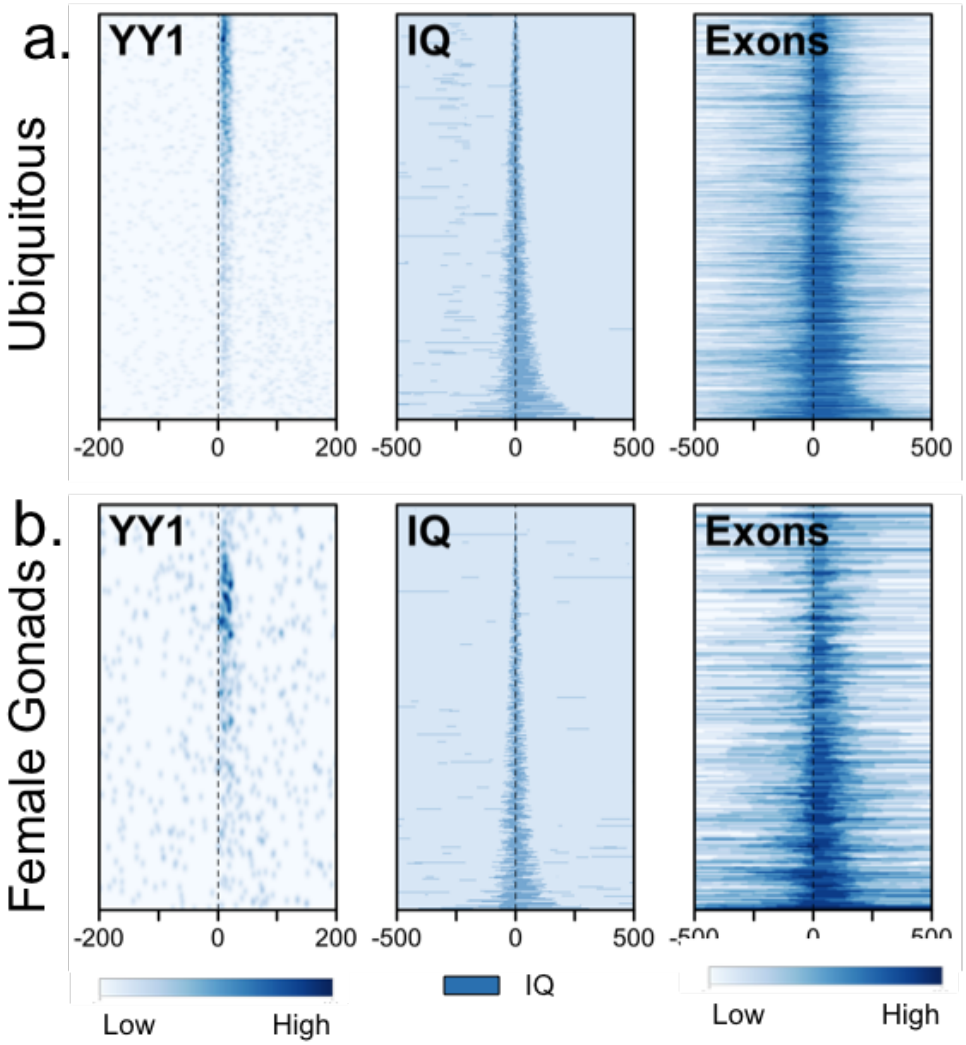

Figure 4.6: Heatmaps of promoter clusters identified in by CAGE expression clustering, aligned by dominant TSS. IQ is the 10-90 interquantile range of the consensus clusters. Exons are gene models predicted from RNA-seq. YY1 is the density of matches to the PWM at 80% of the maximum score. a. Ubiquitous b. Female Gonads

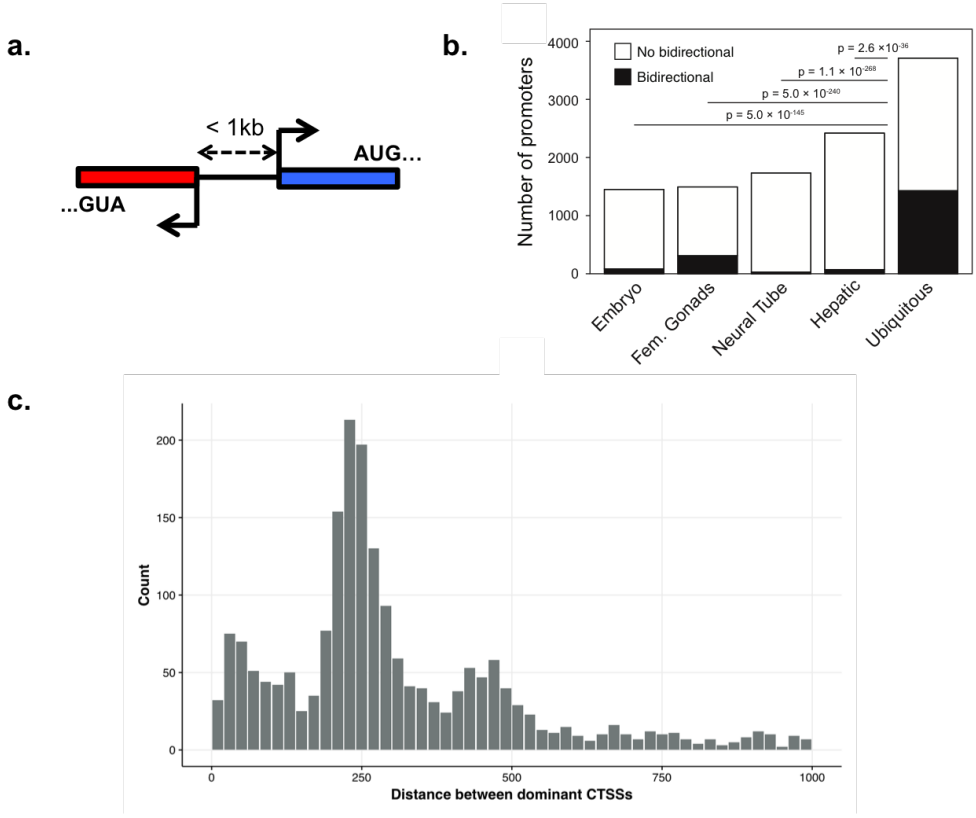

Figure 4.7: a. A schematic of a bidirectional promoter, showing the back-to-back arrangement of the plus-strand promoter (blue) and the negative strand promoter (red), the maximum distance between them (<1kb) and the requirement to be protein coding (AUG codon) b. Promoters distribution by representative CAGE expression clusters. The height of the bar shows the number of promoters in each category, and the fill denotes those which are bidirectional. c. Width between bidirectional promoters visualised as a histogram, showing the periodicity in width which corresponds to integral nucleosome positions.

4.3.6 Bidirectional Promoters

Bidirectional promoters are pairs of gene promoters arranged in a closely spaced back-to-back configuration (Figure 4.7a). I defined bidirectional promoters as head-to-head promoters with a distance of less than 1kb between dominant CTSSs, using consensus clusters. This is the furthest at which we see an enrichment of promoters close together in this orientation in Amphioxus. I restricted the analysis to protein coding genes, since bidirectional initiation is a common feature of many promoters, and does not always result in functional transcripts. Using these criteria, I identified 3950 bidirectional promoters in Amphioxus.

The majority of the genes in bidirectional promoters were from the ubiquitously expressed gene cluster (Figure 4.7b), with significant enrichment vs. all other clusters (binomial test, see Figure 4.7b). Bidirectional promoters are enriched in housekeeping genes and depleted in developmentally regulated genes (Table 4.1), which is expected given that many bidirectional promoters are ubiquitously expressed, a defining feature of housekeeping genes. This could be a result of the lack of space for proximal upstream regulatory elements at closely-spaced promoters, which are necessary for complex patterns of expression. The genes which require complex patterns of expression, such as developmental genes, are excluded from bidirectional promoters (Table 4.1). Bidirectional promoters in amphioxus are not significantly co-regulated, in line with previous studies (Engström et al., 2006).

The distance between bidirectional promoters in Amphioxus is strikingly periodic (Figure 4.7c). and the distance between peaks is roughly the space occupied by a single nucleosome.

Bidirectional promoters have a shared architecture with ubiquitously expressed genes, comprising of an AT-rich band upstream of the promoter, which is occupied by the -1 nucleosome, and a CG band downstream of the promoter, which positions the +1 nucleosome. The periodicity in inter-promoter distance can be seen in the AT-rich upstream signal (Figure 4.8), which changes from a single enriched region to two as the width between promoters increases. This is reminiscent of a 'phase-transition', as it is a discrete

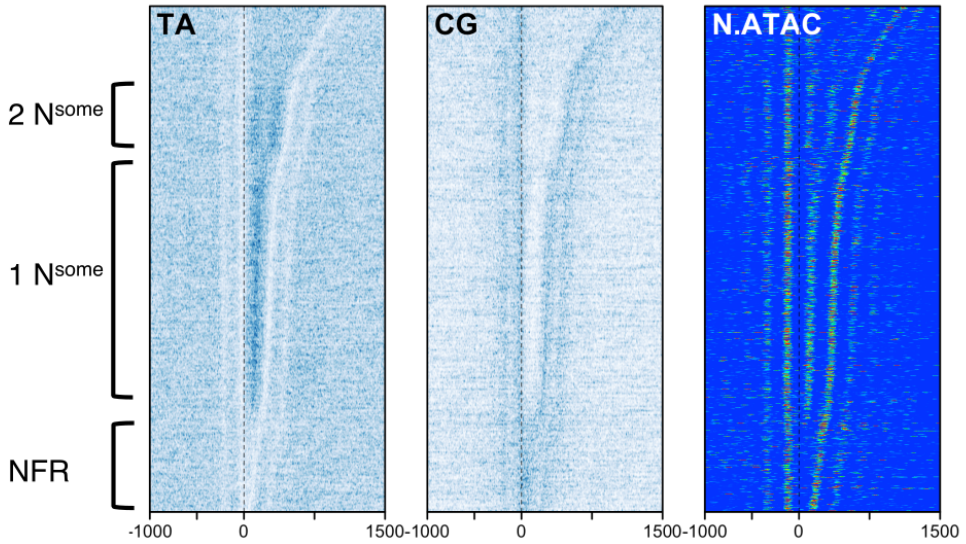

Figure 4.8: Heatmaps of TA, CG and nucleosome occupancy (by NucleoATAC) around bidirectional promoters in Amphioxus, arranged by distance between the two CTSSs. Both TA and NucleoATAC signal clearly indicate regions where 0, 1 or 2 nucleosomes separate promoters

change rather than a gradual one. This is recapitulated in the nucleosome positioning (Schep et al., 2015) derived from the ATAC-seq data (Figure 4.8), which shows that the AT-rich regions are indeed occupied by nucleosomes, and that there is a clear transition from 1 to 2 nucleosomes between the two promoters as the distance increases. The depletion of non-integral nucleosome spacing between promoters indicates that these promoters can "share" -1 or -2 nucleosomes, but configurations in which the nucleosomes are out-of-phase are disfavoured.

4.3.7 Bidirectional Promoters in other species

We have identified a larger number of bidirectional promoters than previously reported in mouse (Engström et al., 2006), despite a lower number of genes overall in Amphixous. We hypothesised that this could reflect a high number of bidirectional promoters in the vertebrate-amphixous last common ancestor, and that the rapid reduction in number of bidirectional promoters

Table 4.1: Top 10 Gene Ontology (GO) Terms enriched in Amphioxus bidirectional promoters, and top 5 excluded terms P-values are Bonferroni corrected. *Process §Ribonucleoprotein Complex †Single-organism organism ‡Multicellular organism

GOBPID	Ratio	Count	Size	Term	P-value
GO:0006396	2.39	332	559	RNA Pr.*	6.84E-19
GO:0034660	2.64	234	376	ncRNA metabolic Pr.	8.54E-16
GO:0022613	2.73	189	299	RNPC§biogen.	3.63E-13
GO:0034470	2.69	172	273	ncRNA Pr.	1.63E-11
GO:0016071	2.23	228	391	mRNA metabolic Pr.	1.34E-10
GO:0042254	2.81	136	212	ribosome biogenesis	2.22E-09
GO:0006364	2.84	113	175	rRNA Pr.	1.33E-07
GO:0016072	2.82	114	177	rRNA metabolic Pr.	1.38E-07
GO:0006397	2.26	161	273	mRNA Pr.	3.71E-07
GO:0006281	2.13	172	299	DNA repair	1.32E-06
GO:0044707	0.52	725	2414	Single-MCO†Pr.	1.63E-30
GO:0032501	0.53	804	2610	MCO Pr.	4.54E-29
GO:0007275	0.54	606	2028	MCO development	6.57E-24
GO:0048856	0.56	690	2232	Anatomical struct. dev.	9.51E-22
GO:0051239	0.48	270	1037	reg. of MCO Pr.	1.32E-19
GO:0032502	0.59	764	2401	Dev. Pr.	2.96E-19
GO:0048731	0.56	542	1804	system dev.	2.98E-19
GO:0044767	0.59	754	2371	SO‡developmental Pr.	6.01E-19
GO:0044700	0.61	706	2203	SO signalling	1.49E-15
GO:0023052	0.61	708	2207	signalling	1.90E-15

Table 4.2: The number of bidirectional promoters in several species, and the number of whole genome duplications between each species and the vertebrate last common ancestor

Organism	Bi-proms	WGD to LCA
Amphioxus	3950	0
Mouse	1752	2
Zebrafish	1098	3
Fly	4036	N/A

in the vertebrate lineage could be a result of the two rounds of whole-genome duplication (WGD) which occurred at the base of the vertebrate tree, after Amphioxus diverged from the vertebrate last common ancestor (LCA). Many genes rediploidise following WGD, and if the genes comprising each bidirectional pair rediploidise independently then we can expect a significant drop in the number of bidirectional promoters following WGD. Using the same criteria, we found 4036 promoters in fly (*Drosophila melanogaster*), 1752 in mouse, and 1098 in zebrafish (see Table 4.2), which is consistent with this hypothesis.

These bidirectional promoters also share similar gene ontology enrichment as those in amphioxus (Table 4.3), which is to say they are predominantly housekeeping genes involved in basic cellular processes. This provides further evidence that these distinct sets of bidirectional promoters represent a single set of genes through evolution.

4.3.8 Estimating the the rate of loss of bidirectional promoters

The hypothesis that bidirectional promoters will reduce in number following successive rounds of WGD is based on the observation that most genes rediploidise following WGD (Kikuta et al., 2007). We can use the relative numbers of bidirectional promoters across the vertebrate lineage to estimate both the rate of loss of bidirectional promoters, and the rediploidisation rate of the genes within them.

Immediately following WGD, each bidirectional promoter will exist in two

Table 4.3: Top 5 GO Ontology Biological Process terms enriched in bidirectional promoters in mouse, zebrafish and fly.

GOBPID	OddsRatio	Count	Size	Term	P-value
			Mouse		
GO:0006396	2.39	332	559	RNA processing	6.84E-19
GO:0034660	2.64	234	376	ncRNA MP	8.54E-16
GO:0022613	2.73	189	299	RNPC biogenesis	3.63E-13
GO:0034470	2.69	172	273	ncRNA processing	1.63E-11
GO:0016071	2.23	228	391	mRNA MP	1.34E-10
			Zebrafsh		
GO:0006396	3.04	70	337	RNA processing	1.53E-09
GO:0034470	4.12	42	157	ncRNA processing	1.84E-08
GO:0034660	3.49	48	204	ncRNA MP	8.73E-08
GO:0042254	3.74	39	156	ribosome biogenesis	9.57E-07
GO:0022613	3.09	45	209	RNPC biogenesis	7.78E-06
			Fly		
GO:0044260	2.12	1414	2712	cell. macromol. MP	3.39E-52
GO:0009987	2.10	2496	5493	cell. process	1.59E-45
GO:0044237	1.82	1696	3523	cell. MP	1.32E-35
GO:0010467	1.99	904	1692	gene expression	1.61E-32
GO:0034641	1.88	1117	2178	cell. nitrogen MP	4.10E-32

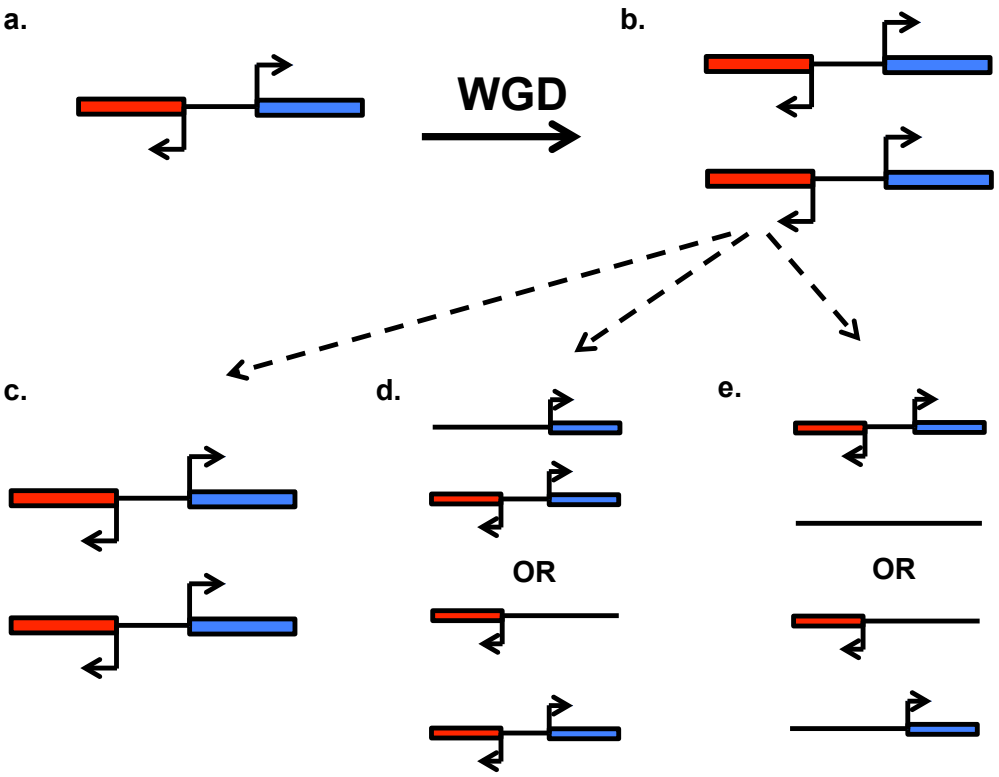

Figure 4.9: A single ancestral bidirectional promoter (a) is duplicated following a WGD event (b). If neither gene rediploidises, both bidirectional promoters are preserved (c), if one gene rediploidises, 1 bidirectional promoter is preserved, and if both genes rediploidise (e), there is a probability of $\frac{1}{2}$ that one bidirectional promoter remains. f. The observed number of bidirectional promoters (bp) after each WGD, vs the number expected under our model (line).

copies in the new, tetrapoid genome (Figure 4.9b). After a period of time, each of the two genes comprising the original bidirectional promoter may have rediploidised. If neither gene rediploidises, then 2 bidirectional promoters will remain after WGD, as shown in panel c. If one gene rediploidises, then 1 bidirectional promoter will remain (Panel d). If both genes rediploidise, there is a 1/2 chance of a bidirectional promoter remaining, which is the probability that the remaining copies of both genes are at the same locus (Panel e).

Given a reploidisation rate R, the probabilities of 0, 1 or 2 genes rediploidising are $(1-R)^2$, $2R(1-R)$ and R^2 respectively. Combining this with the expected number of bidirectional promoters remaining after each possible event (2, 1 or $\frac{1}{2}$), the ratio of bidirectional promoters remaining (or rate of loss, L) as a function of the rediploidisation rate (R) will be:

$$L = 2(1-R)^2 + 2(1-R)R + \frac{R^2}{2}$$

which simplifies to:

$$L = \frac{R^2}{2} - 2R + 2 \tag{4.1}$$

If this hypothesis is correct, and the ration of bidirectional promoters lost at each WGD is constant, then the number of bidirectional promoters in each generation will follow a geometric distribution with the equation:

$$n = a \times L^{WGD} \tag{4.2}$$

where n is the number of bidirectional promoters, a is the ancestral number and L is the rate of loss. Using the numbers from Table 4.2, $L = 0.66$. The curve was fitted assuming the number of bidirectional promoters in Amphioxus to be the ancestral number and using the non-linear least squares (`nls`) optimisation function in R. The resulting fit is shown in Figure 4.10b.

With L equal to 0.66, we estimate R to be 0.85 from the quadratic equation above. This means that most genes in bidirectional promoters rediploidise, which is in line with previous studies reporting that housekeep-

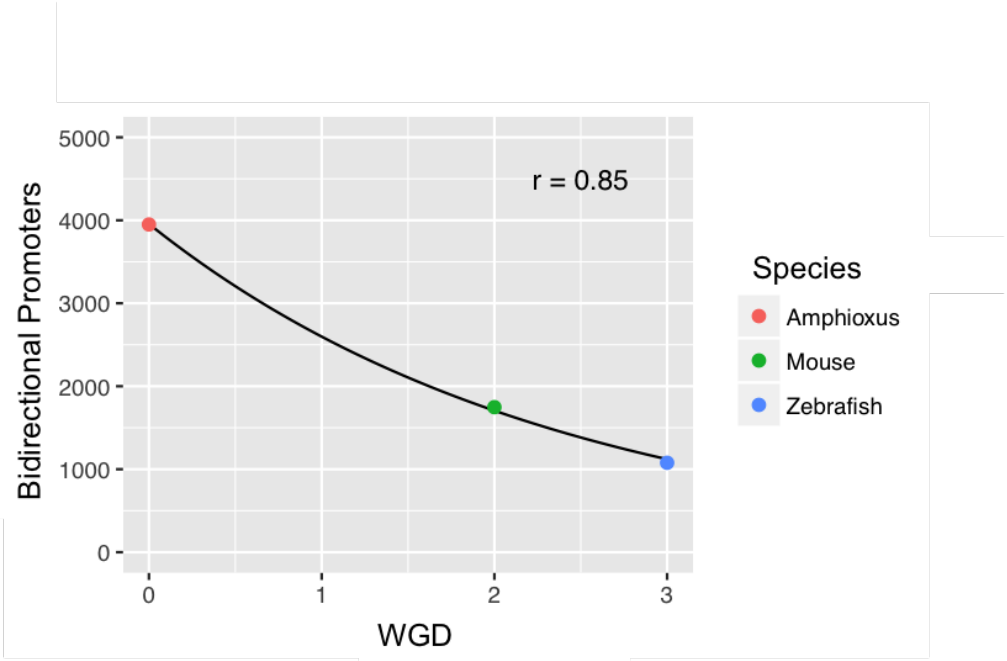

Figure 4.10: Number of bidirectional promoters as a function of WGD since the vertebrate LCA. The line is equation 4.2 fitted to the three points. r is the rediploidisation rate inferred from this analysis from equation 4.1

ing genes are more likely to undergo rediploidisation following WGD (Kikuta et al., 2007). The comparison of this estimate with our observed values shows a good fit with the observed data (Figure 4.9f), supporting the hypothesis that amphioxus is representative of the ancestral state.

4.3.9 Promoter Evolution

Following the set of genes at bidirectional promoters through vertebrate evolution also allows us to trace the changing architecture of housekeeping promoters. Neither zebrafish nor mouse promoters show the same precise nucleosome positioning signal or constrains on bidirectional promoter width that is seen in Amphioxus, even though within each genome the promoter structure is very consistent (Figure 4.11).

Zebrafish bidirectional promoters (Figure 4.11b) show a strong TA enrichment between bidirectional promoters, for those pairs of promoters more than a nucleosome width apart. This correlates with nucleosome binding, according to ATAC-seq data. There is also a slight CG band at the promoter itself. Mouse bidirectional promoters (Figure 4.11a) show no TA enrichment, but have a broad CG band extending into the transcript of each promoter. *Drosophila melanogaster* bidirectional promoters (Figure 4.11c) have a similar TA enrichment between promoters, but also have a second TA band downstream of the dominant TSS. All promoters show disordered, rather than in-phase, nucleosomes between bidirectional promoters.

This suggests that the sequence features of this subset of chordate promoters are highly malleable and undergo concerted evolution, which is counterintuitive given that the transcriptional machinery is almost unchanged across this clade. These bidirectional promoters also share the same gene ontology enrichment as those in amphioxus, providing further evidence that these distinct sets of bidirectional promoters represent a single set of genes through evolution.

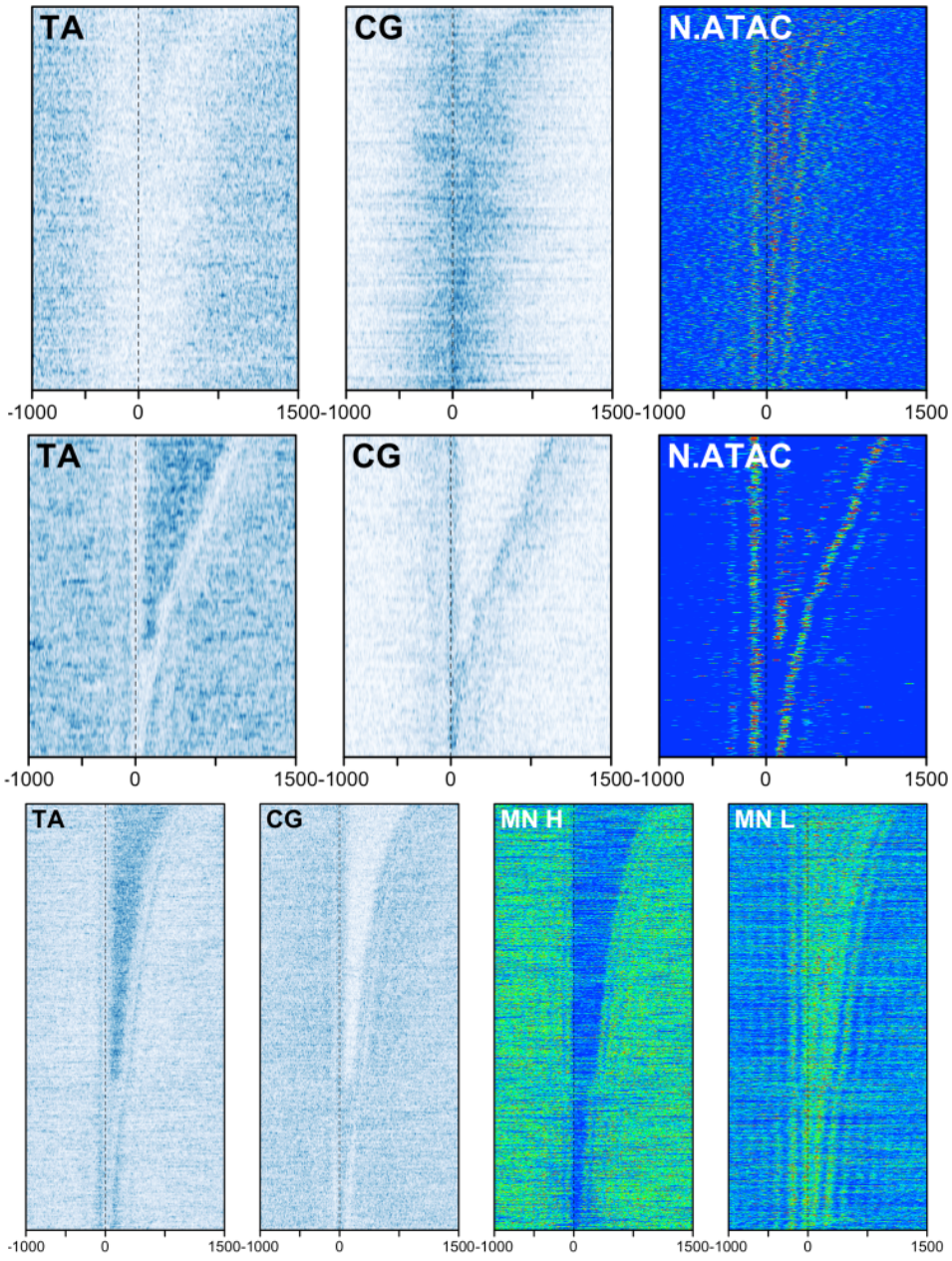

Figure 4.11: Bidirectional Promoters in Model Organisms a. Smoothed TA and CG dinucleotide densities and nucleosome positioning signal from NucleoATAC bidirectional promoters in mice. b. As a. but for zebrafish. c. Smoothed TA and CG dinucleotide densities and nucleosome position for bidirectional promoters in Drosophila melanogaster. MN H is high concentration MNase digestion, MN L is low concentration MNase digestion, data from (Gehrke et al., 2015)

4.4 Discussion

I used CAGE data to show that promoters in Amphioxus conform to two distinct structures, which show both lineage-specific innovation and pan-Metazoan features. Ubiquitously expressed, or "housekeeping", promoters have a divergent architecture, and are frequently found in back-to-back pairings, which appears to be the ancestral state. These bidirectional promoters also exhibit tight control of nucleosome positioning between each TSS, although the mechanism underlying this novel arrangement remains unclear. Developmental and tissue-specific promoters in amphioxus have highly conserved features, such as a narrow range of initiation at a subset of promoters and TATA-dependent transcription. These features are not particular to certain tissues; the breadth of expression matters more than the exact tissue in which a promoter is active. These results highlight the underappreciated role of promoters in gene regulation.

Manual inspection of the data shows that the CAGE signal at promoters is robust, despite the low number of uniquely aligned reads. This is likely due to the difficulty in adapting protocols to a novel organism, possible resulting in low RNA stability, and equally the relatively poor quality of the genome when compared to reference genomes, such as those of zebrafish and mouse.

This experiment demonstrates, for the first time using CAGE, the effect of YY1 on transcription initiation, and suggests an expanded role for YY1. Given its ubiquity in our sample, it could be considered a core promoter element with similar importance to other well-characterised motifs like the TATA box. Its unique distribution just downstream of the dominant peak of transcription initiation hints at a mechanistic function in restricting initiation to favourable regions, particularly as this associates with short first exons where precise initiation is required for correct RNA processing.

Nucleosome positioning at most promoters in amphioxus is stable relative to the dominant TSS, as expected from studies in other organisms. This happens both in the presence and absence of clear nucleosome positioning signals in the DNA, and the mechanisms by which clear +1 nucleosome positioning is established, in the absence of both nucleosome positioning signal or a TATA

box, is worth further investigation. Bands of **TA** enrichment upstream of ubiquitous promoters appear to function as a nucleosome positioning signal distinct from the familiar periodic enrichment bands.

At bidirectional promoters, a different pattern is observed. The 'phased' nucleosome positions at closely spaced back-to-back promoters are a novel observation. Bidirectional promoters in mouse, zebrafish and fly appear to tolerate either a nucleosome free region, or a region of disordered nucleosomes, quite happily. Promoter sequences in these organisms do not show such clear evidence of nucleosome positioning signals, so it is possible that precise control of nucleosome position is of greater importance to amphioxus, and therefore disordered nucleosome at promoters are not tolerated as well.

Bidirectional promoters also provide an interesting set of genes to study promoter evolution, since there is strong evidence from both GO ontology and whole genome duplication events that bidirectional promoters in all species come from a single set of ancestral promoters. Gene promoters are intrinsically bidirectional (Andersson et al., 2015a), and it is possible that transcription on the reverse strand encouraged evolutionary innovation by providing a ready-made promoter for any gene translocating to the correct position, or for genes to arise *de novo*. On the other hand, this hypothesis does not explain why more bidirectional promoters have not arisen since the base of the vertebrate lineage.

These results point to a high level of malleability in the structure of promoters in metazoa. Housekeeping genes, as shown elegantly by the conserved set of bidirectional promoters, underwent concerted evolution in several clades dispute well-conserved transcriptional machinery, and are quite different in all species examined. Further research is needed to elucidate the mechanisms underlying these changes, and their consequences (if any) for gene regulation.

4.5 Methods

4.5.1 CAGE-seq

RNA from 32-cell, 8hpf and 15hpf developmental stages and female gonads, muscle, neural tube and hepatic diverticulum from adult individuals were obtained as described above. CAGE was performed on these samples using the nAnT-iCAGE (non-amplifying non-tagging Illumina CAGE) protocol, as described in Murata et al. (2014) , with 7 micrograms of RNA per sample and a single replicate for each condition. Following library preparation, samples were sequenced in a single multiplex lane on a HiSeq 2500, with 50bp read length.

Mouse CAGE data was taken from FANTOM5 (FANTOM Consortium and the RIKEN PMI and CLST (DGT) et al., 2014). The samples used were whole body (E11) and heart, liver, thymus and uterus (E14). These were obtained as bed coordinates from the FANTOM data repository, through CAGEr (Haberle et al., 2015). Zebrafish CAGE was obtained from Nepal et al. (2013), comprising 12 stages of a developmental timecourse from unfertilised egg to Prim20. This was obtained through the 'ZebrafishDevelopmentalCAGE' R package available at http://promshift.genereg.net/CAGEr/. *Drosophila melanoagester* CAGE was obtained from ModEncode (Celniker et al., 2009), also through CAGEr.

4.5.2 CAGE alignment

CAGE tags were aligned to the Amphioxus genome using Bowtie v. 1.1.2 (Langmead et al., 2009) with a seed length of 25, allowing 2 mismatches in the seed region and discarding multi-mapping reads. The following mapping efficiencies were obtained:

The resulting alignments were processed to remove leading G nucleotides where this did not map to the reference. The tag counts at each nucleotide was normalised to follow a power-law distribution (Balwierz et al., 2009). All samples except for muscle passed quality control, showing the expected power-law distribution of tag counts and bimodal distribution of cluster

Table 4.4: Read count and alignment efficiency for CAGE samples

sample	#reads	#aligned	#!aligned	#multimappers
Fem Gonads	24,458,864	9,762,213 (39.91%)	11,133,520 (45.52%)	3,563,131 (14.57%)
Embryo @8 h	27,837,581	9,341,470 (33.56%)	9,932,848 (35.68%)	8,563,263 (30.76%)
Embryo @15 h	20,280,947	7,760,395 (38.26%)	6,862,480 (33.84%)	5,658,072 (27.90%)
Embryo @36 h	7,780,414	2,726,242 (35.04%)	2,394,206 (30.77%)	2,659,966 (34.19%)
Hepatic	17,684,847	5,447,194 (30.80%)	5,755,835 (32.55%)	6,481,818 (36.65%)
Muscle	1,348,544	278,466 (20.65%)	309,141 (22.92%)	760,937 (56.43%)
Neural Tube	5,087,538	2,027,998 (39.86%)	1,575,573 (30.97%)	1,483,967 (29.17%)

widths 4.1. The muscle sample was dominated by width 1 TCs, which are not indicative of biologically meaningful promoters, and so was excluded from further analysis.

4.5.3 CAGE Tag Clustering

Single base-pair CAGE transcription start sites (CTSSs) were clustered into tag clusters (TCs) using the distance-based clustering in CAGEr (Haberle et al., 2015), taking the 10th and 90th of percentile of tags to improve robustness. Each TC was also assigned a dominant CTSS, with the highest expression.

TCs were further clustered across cell types to produce comparable promoter regions. TCs without support greater than 1 TPM in at least one cell type were excluded from further analysis.

CAGE transcription start sites (CTSSs) map initiation at single base-pair resolution. CTSSs commonly occur in clusters which reflect the activity of a single gene promoter and give rise to functionally equivalent transcripts 4.1. Therefore, nearby individual CTSSs were combined using the distance-based

clustering method in CAGEr (Haberle et al., 2015) to produce tag clusters (TCs), which are summarise expression at individual promoters. To reduce sensitivity to outlying CTSSs and improve robustness, the width of each TC was calculated by discarding the first 10% of CAGE signal, and the last 10%. This is referred to as the Inter-quantile (IQ) range (Haberle et al., 2015). TCs were also assigned a dominant CTSS, which is the individual CTSS with the highest expression level. TCs with expression of less than 1 tag per million (TPM) were filtered out.

4.5.4 Expression clustering

The consensus tag clusters were further clustered by expression patterns using a self-organising map (SOM) (Wehrens et al., 2007), using a 5x5 arrangement. The SOM produced both defined clusters and a topographic relationship between clusters (Figure 2).

The topographic relationship of clusters in the SOM allows us to group similar clusters together to produce sets of consensus tag clusters (conceptually, sets of promoters) which are expressed in similar tissues during the course of development. The groups which emerged most clearly were "Embryonic", "Neural Tube", "Hepatic", "Female Gonads" and "Ubiquitous" (see Figure 4.2).

4.5.5 Feature enrichment and visualisation

We investigated the relative presence and enrichment of the following features: TATA box, YY1 motif, GC and AT content, SS and WW dinucleotides, first exons and nucleosome positioning signal. Heatmaps were plotted for visualisation by scanning either for exact dinucleotide matches or PWM matches at 80maximum score. PWMs for TATA and YY1 were taken from the JASPAR vertebrate collection (Mathelier et al., 2016). The binary matrix resulting from sequence matches (either PWM-based or exact) was then smoothed and down-sized using a binned Gaussian kernel density approach. Nucleosome positioning signal was winsorised to the 99th percentile, and smoothed using a Gaussian blur. Promoters were sorted by IQ

width, unless otherwise specified. All plots were made using R/Bioconductor, specifically with my heatmaps package, which is included in Bioconductor 3.5 (Huber et al., 2015).

4.5.6 ATAC-seq Data

The amphioxus ATAC using the ATAC- seq data from this paper, and ATAC-seq data for mouse and zebrafish were taken from (Gehrke et al., 2015), using the limb and whole-body 24hpf respectively. Additional datasets were aligned using Bowtie (25bp seed region allowing 2 mismatches). To calculate nucleosome positioning from aligned ATAC-seq data, we used NucleoATAC v0.3.2 (Schep et al., 2015) run using default parameters.

4.5.7 Gene Annotation and GO Enrichment

Gene names were assigned to CAGE peaks based on the transcriptome assembly. Genes were assigned to CAGE peaks if they were within 1kb of a TSS. Gene ontology enrichment was performed using the human gene annotations of the amphioxus transcriptome assembly, through the GOstats (Falcon and Gentleman, 2007) and human genome annotation packages in Bioconductor. Gene annotations for other species used the same criteria, using Ensembl 67 gene annotation (Aken et al., 2016). GO enrichment was calculated from species-specific annotation in the bioconductor annotation packages (Carlson, 2017).

Chapter 5

Discussion

In this thesis I have investigated the role of both promoters and enhancers in gene regulation: in particular, the differences in gene regulation between developmental genes and housekeeping genes. In Chapter 2, I showed that enhancers work together in large arrays to direct the expression of their targets, and that these targets are very strongly enriched in developmental and signalling genes. I also presented evidence that physical contacts important for enhancer regulation, but may be insufficient to explain the specificity of enhancer targeting alone. In Chapter 3, I presented an R/Bioconductor package 'heatmaps', written with the aim of visualising spatial associations between DNA sequence and experimental assays. In Chapter 4, I used this package, alongside other analyses, to show that CAGE-defined promoters in the European Amphioxus have at least two major architectures, the first corresponding to tissue-specific, or developmentally regulated genes, and the second to ubiquitously expressed housekeeping genes. I also identified a set of bidirectional promoters which appear to be conserved throughout vertebrate evolution.

5.1 Developmental Promoters in Amphioxus are Defined at the Sequence Level

I have identified major differences in promoter architecture between genes that are regulated specifically during development, and genes with more stable expression. These housekeeping promoters are defined by a TA-rich band upstream of the promoter, which marks the -1 nucleosome, and two periodic GC-rich bands which flank the +1 nucleosome. Both of these features may have consequences for nucleosome positioning, which is worth investigating further. It is also not clear what role, if any, this precise positioning of nucleosomes plays in gene regulation, although it appears that configurations of back-to-back genes where this nucleosome is absent are disfavoured.

Developmentally regulated promoters contain neither of these features, but have a band of CG enrichment at the dominant TSS and a significant minority of TATA-containing promoters. In addition, very few developmentally regulated promoters are in closely-spaced bidirectional arrangements. Intuitively, this relate to the constitutive expression (i.e. largely unregulated) of housekeeping genes, since there is little space for transcription factors to bind. Additionally, enhancers physically contacting such promoters could be assumed to contact both promoters in the bidirectional pair, which would hinder precise control of gene expression. Looking ahead, bioinformatics approaches are unlikely to resolve these questions entirely, and I believe that, were the European Amphioxus to one day become a major laboratory model organism, they would be very much worth further investigation.

5.2 Developmentally regulated genes in humans respond preferentially to enhancers

In humans, developmental genes show a much greater sensitivity to enhancers compared with housekeeping genes, based on a model of enhancers working together to produce complex expression patterns across multiple tissues. The genes identified are very strongly enriched for developmental processes, tran-

scription factors and cell-surface receptors involved in signalling. These genes are more likely to overlap CpG islands, and when they do overlap CpG islands, the islands are longer, indicating that the promoters of genes under long-range regulation are meaningfully different to those of other genes.

However, these results are far from conclusive. There are many unexplained aspects to this model, such as how gene regulation can work in TADs containing many target genes. The model is also quite susceptible to noise (as with any statistical model), and so there exist a large number of indeterminate cases where there is neither conclusive evidence for or against long-range regulation.

The role of DNA-DNA interactions or enhancer looping is also unclear. From a purely physical point of view, complex loci containing many genes present a challenge to looping-based models of promoter-enhancer interactions. Wit many TADs containing hundreds of enhancers, it would be topologically difficult to arrange each enhancer to contact a different promoter using discrete loops in every case. On the other, promoters do show increased contact with their enhancers at highly expressed genes, and it stands to reason that promoter-enhancer loops are *necessary* for enhancer regulation, if not entirely sufficient.

I would conjecture that a mechanistic model whereby housekeeping genes are, at some level, immune to the regulatory effects of enhancers resolves many of these problems. Enhancers within a TAD would be free to contact any genes, but would have an effect only at a subset of these genes. These other genes would be controlled by transcription factors binding directly at the promoter, or constitutively expressed. This is in line with experimental observations made using self-transcribing enhancer assays. Zabidi et al. (2015) showed a markedly different response of several *D. melanogaster* promoters to regulation by enhancers, and that core promoter architecture differed between the two groups.

5.3 Revisiting the GRB model of gene regulation

The differential response of developmental and housekeeping genes was proposed alongside the discovery of Genomic Regulatory Blocks (GRBs), based on evidence from conserved non-coding elements (CNEs) (Kikuta et al., 2007; Akalin et al., 2009; Harmston et al., 2017). I believe that, when taken together, the ideas presented in this thesis make up significant evidence in favour of this model of gene regulation. Differences in promoter architecture and differential response to enhancers by certain classes of genes could be linked, because it is differences in promoter sequence which govern the response of specific genes to enhancers. In turn, the effect of enhancers is limited by strong TAD boundaries.

Underlying these observations, there is a simple but fundamental rationale for why developmental genes require different regulation: their expression patterns are much more varied and complex. Many key developmental transcription factors and signalling pathways are expressed across diverse tissues, performing different roles within each. If the total expression patterns for each gene are determined by a mosaic of individual enhancers, which may be bound by many different factors, this would provide modular, flexible control of gene expression, as suggested by Lorberbaum et al. (2016), and this behaviour may shed light on the link between enhancer clusters and developmental genes (Whyte et al., 2013; Parker et al., 2013).

However, mechanistically there are still many gaps in this picture. Promoter architecture is, in general, not well understood at a quantitative level while many promoter features have been described and show significant associations with patterns of regulation, these cannot in general be translated into predictive tools. If these features can be understood at a deeper level, this might provide a starting point for the elucidation of the mechanisms driving this behaviour. There is even less understanding of the biology of enhancers, and the level of conservation at enhancers remains completely unexplained (Harmston et al., 2013).

5.4 Promoters have characteristic features which change over time

The bidirectional promoters I identified in Chapter provide a unique opportunity to study the evolution of a single set of promoters through the vertebrate lineage. While it is possible to trace the evolution of an individual gene promoter by comparing coding sequences using BLAST or a similar program, it is hard to derive promoter features from a single example: they follow general patterns rather than specific rules in most cases. The co-ordinated changes in nucleotide frequencies throughout this set of promoters appear to represent concerted evolution at multiple loci. However, despite these changes, the core promoter machinery, and the fundamental mechanisms of gene regulation, have not undergone significant divergence across the vertebrate lineage, which raises an interesting question as to why such a divergence has occurred, and how such changes are tolerated by the core promoter machinery.

Intuitively, any changes at the core promoters of housekeeping genes would be deleterious, since mis-expression of these genes would likely cause dis-regulation of vital cellular process. Secondly, how is function conserved at gene promoters when the sequence features are not? It is possible that corresponding changes in promoter machinery buffer these changes slowly over time. For example, the nucleosome positioning which appears to be very important at Amphioxus promoters may be more relaxed in mammals, or the function carried out by the sequence to position nucleosomes may have been taken over by DNA-binding proteins. Alternatively, the function could be conserved in the sequence in a very similar manner, but not one that is understandable with current tools. This is analogous to conservation of small RNAs: the secondary structure can be conserved between elements even when the sequences diverge significantly (Gruber et al., 2008).

5.5 Understanding biological sequence

Differences in promoter structure are better captured by visualising nucleotide and dinucleotide frequencies than by analysing over-represented motifs. This may be caused in part due to the difficulty of *de novo* motif discovery; on the other hand, few motifs have been discovered which function as core promoter elements in vertebrates, and none of them are essential for all functional gene promoters. This represents a significant gap in our understanding of promoters, and possibly our understanding of biological sequences in general, since our current tools are not capable of elucidating the minimal requirements to form a promoter. Alternatively, there may be no unique factors determining the capability of DNA sequence to act as a promoter, and that transcription at most promoters is driven by a range of cell-type specific transcription factors.

Recent advances in machine learning have opened up new possibilities for biological sequence analysis. Tools such as artificial neural networks are able to model complex spatial patterns, such as those found through the plotting of heatmaps in Chapter 4. This allows for the automatic detection of patterns that were previously difficult to quantify, which currently require visual inspection to detect. The ability to model more complex, and potentially more biologically relevant, sequence patterns may lead to major advances with the next-generation of sequence analysis tools.

Bibliography

A. Akalin, D. Fredman, E. Arner, X. Dong, J. C. Bryne, H. Suzuki, C. O. Daub, Y. Hayashizaki, and B. Lenhard. Transcriptional features of genomic regulatory blocks. *Genome Biol.*, 10(4):R38, Apr. 2009.

A. Akalin, V. Franke, K. Vlahoviček, C. E. Mason, and D. Schübeler. Genomation: a toolkit to summarize, annotate and visualize genomic intervals. *Bioinformatics*, 31(7):1127–1129, Apr. 2015.

B. L. Aken, S. Ayling, D. Barrell, L. Clarke, V. Curwen, S. Fairley, J. Fernandez Banet, K. Billis, C. García Girón, T. Hourlier, K. Howe, A. Kähäri, F. Kokocinski, F. J. Martin, D. N. Murphy, R. Nag, M. Ruffier, M. Schuster, Y. A. Tang, J.-H. Vogel, S. White, A. Zadissa, P. Flicek, and S. M. J. Searle. The ensembl gene annotation system. *Database*, 2016, June 2016.

R. Andersson, Y. Chen, L. Core, J. T. Lis, A. Sandelin, and T. H. Jensen. Human gene promoters are intrinsically bidirectional. *Mol. Cell*, 60(3):346–347, Nov. 2015a.

R. Andersson, A. Sandelin, and C. G. Danko. A unified architecture of transcriptional regulatory elements. *Trends Genet.*, 31(8):426–433, Aug. 2015b.

F. J. Anscombe. Graphs in statistical analysis. *Am. Stat.*, 27(1):17–21, 1973.

C. D. Arnold, D. Gerlach, C. Stelzer, Ł. M. Boryń, M. Rath, and A. Stark. Genome-wide quantitative enhancer activity maps identified by STARR-seq. *Science*, 339(6123):1074–1077, Mar. 2013.

J. N. Athanikar, R. M. Badge, and J. V. Moran. A YY1-binding site is required for accurate human LINE-1 transcription initiation. *Nucleic Acids Res.*, 32(13): 3846–3855, July 2004.

S. D. Bailey, X. Zhang, K. Desai, M. Aid, O. Corradin, R. Cowper-Sal Lari, B. Akhtar-Zaidi, P. C. Scacheri, B. Haibe-Kains, and M. Lupien. ZNF143 provides sequence specificity to secure chromatin interactions at gene promoters. *Nat. Commun.*, 2:6186, Feb. 2015.

T. L. Bailey, M. Boden, F. A. Buske, M. Frith, C. E. Grant, L. Clementi, J. Ren, W. W. Li, and W. S. Noble. MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.*, 37(Web Server issue):W202–8, July 2009.

P. J. Balwierz, P. Carninci, C. O. Daub, J. Kawai, Y. Hayashizaki, W. Van Belle, C. Beisel, and E. van Nimwegen. Methods for analyzing deep sequencing expression data: constructing the human and mouse promoterome with deepCAGE data. *Genome Biol.*, 10(7):R79, July 2009.

D. P. Bartel. MicroRNAs: Genomics, biogenesis, mechanism, and function. *Cell*, 116(2):281–297, Jan. 2004.

A. Bassett, S. Cooper, C. Wu, and A. Travers. The folding and unfolding of eukaryotic chromatin. *Curr. Opin. Genet. Dev.*, 19(2):159–165, Apr. 2009.

N. Battulin, V. S. Fishman, A. M. Mazur, M. Pomaznoy, A. A. Khabarova, D. A. Afonnikov, E. B. Prokhortchouk, and O. L. Serov. Comparison of the three-dimensional organization of sperm and fibroblast genomes using the Hi-C approach. *Genome Biol.*, 16:77, Apr. 2015.

G. Bejerano. Ultraconserved elements in the human genome. *Science*, 304(5675): 1321–1325, May 2004.

S. Benko, J. A. Fantes, J. Amiel, D.-J. Kleinjan, S. Thomas, J. Ramsay, N. Jamshidi, A. Essafi, S. Heaney, C. T. Gordon, D. McBride, C. Golzio, M. Fisher, P. Perry, V. Abadie, C. Ayuso, M. Holder-Espinasse, N. Kilpatrick, M. M. Lees, A. Picard, I. K. Temple, P. Thomas, M.-P. Vazquez, M. Vekemans, H. Roest Crollius, N. D. Hastie, A. Munnich, H. C. Etchevers, A. Pelet, P. G. Farlie, D. R. Fitzpatrick, and S. Lyonnet. Highly conserved non-coding elements on either side of SOX9 associated with pierre robin sequence. *Nat. Genet.*, 41 (3):359–364, Mar. 2009.

Bioconductor Core. Bioconductor - packages: Guidelines. `https://www.bioconductor.org/developers/package-guidelines/`, Oct. 2017. Accessed: 2017-10-21.

A. P. Bird. CpG-rich islands and the function of DNA methylation. *Nature*, 321 (6067):209–213, 1986.

E. M. Blackwood and J. T. Kadonaga. Going the distance: a current view of enhancer action. *Science*, 281(5373):60–63, July 1998.

J. D. Buenrostro, B. Wu, H. Y. Chang, and W. J. Greenleaf. ATAC-seq: A method for assaying chromatin accessibility Genome-Wide. *Curr. Protoc. Mol. Biol.*, 109:21.29.1–9, Jan. 2015.

T. W. Burke and J. T. Kadonaga. The downstream core promoter element, DPE, is conserved from drosophila to humans and is recognized by TAFII60 of drosophila. *Genes Dev.*, 11(22):3020–3031, Nov. 1997.

J. E. Butler and J. T. Kadonaga. Enhancer-promoter specificity mediated by DPE or TATA core promoter motifs. *Genes Dev.*, 15(19):2515–2519, Oct. 2001.

E. Calo and J. Wysocka. Modification of enhancer chromatin: what, how, and why? *Mol. Cell*, 49(5):825–837, 2013.

E. Cannavò, P. Khoueiry, D. A. Garfield, P. Geeleher, T. Zichner, E. H. Gustafson, L. Ciglar, J. O. Korbel, and E. E. M. Furlong. Shadow enhancers are pervasive features of developmental regulatory networks. *Curr. Biol.*, 26(1):38–51, Jan. 2016.

M. Carlson. GO.db: A set of annotation maps describing the entire gene ontology, 2017.

P. Carninci, C. Kvam, A. Kitamura, T. Ohsumi, Y. Okazaki, M. Itoh, M. Kamiya, K. Shibata, N. Sasaki, M. Izawa, M. Muramatsu, Y. Hayashizaki, and C. Schneider. High-efficiency full-length cDNA cloning by biotinylated CAP trapper. *Genomics*, 37(3):327–336, Nov. 1996.

P. Carninci, A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. M. Semple, M. S. Taylor, P. G. Engström, M. C. Frith, A. R. R. Forrest, W. B. Alkema, S. L. Tan, C. Plessy, R. Kodzius, T. Ravasi, T. Kasukawa, S. Fukuda, M. Kanamori-Katayama, Y. Kitazume, H. Kawaji, C. Kai, M. Nakamura, H. Konno, K. Nakano, S. Mottagui-Tabar, P. Arner, A. Chesi, S. Gustincich, F. Persichetti, H. Suzuki, S. M. Grimmond, C. A. Wells, V. Orlando, C. Wahlestedt, E. T. Liu, M. Harbers, J. Kawai, V. B. Bajic, D. A. Hume, and Y. Hayashizaki. Genome-wide analysis of mammalian promoter architecture and evolution. *Nat. Genet.*, 38(6):626–635, June 2006.

S. E. Celniker, L. A. L. Dillon, M. B. Gerstein, K. C. Gunsalus, S. Henikoff, G. H. Karpen, M. Kellis, E. C. Lai, J. D. Lieb, D. M. MacAlpine, G. Micklem, F. Piano, M. Snyder, L. Stein, K. P. White, R. H. Waterston, and modENCODE Consortium. Unlocking the secrets of the genome. *Nature*, 459(7249):927–930, June 2009.

S.-S. Chae, J.-H. Paik, M. L. Allende, R. L. Proia, and T. Hla. Regulation of limb development by the sphingosine 1-phosphate receptor S1p1/EDG-1 occurs via the hypoxia/VEGF axis. *Dev. Biol.*, 268(2):441–447, Apr. 2004.

C. S. Chao, K. D. McKnight, K. L. Cox, A. L. Chang, S. K. Kim, and B. J. Feldman. Novel GATA6 mutations in patients with pancreatic agenesis and congenital heart malformations. *PLoS One*, 10(2):e0118449, Feb. 2015.

R. V. Chereji, T.-W. Kan, M. K. Grudniewska, A. V. Romashchenko, E. Berezikov, I. F. Zhimulev, V. Guryev, A. V. Morozov, and Y. M. Moshkin. Genome-wide profiling of nucleosome sensitivity and chromatin accessibility in drosophila melanogaster. *Nucleic Acids Res.*, 44(3):1036–1051, Feb. 2016.

S. L. Clarke, J. E. VanderMeer, A. M. Wenger, B. T. Schaar, N. Ahituv, and G. Bejerano. Human developmental enhancers conserved between deuterostomes and protostomes, Aug. 2012.

J. M. Claverie and S. Audic. The statistical significance of nucleotide position-weight matrix matches. *Comput. Appl. Biosci.*, 12(5):431–439, Oct. 1996.

H. A. Cole, B. H. Howard, and D. J. Clark. Genome-wide mapping of nucleosomes in yeast using paired-end sequencing. *Methods Enzymol.*, 513:145–168, 2012.

E. de Wit and W. de Laat. A decade of 3C technologies: insights into nuclear organization. *Genes Dev.*, 26(1):11–24, Jan. 2012.

J. Dekker, K. Rippe, M. Dekker, and N. Kleckner. Capturing chromosome conformation. *Science*, 295(5558):1306–1311, Feb. 2002.

F. Delsuc, H. Brinkmann, D. Chourrout, and H. Philippe. Tunicates and not cephalochordates are the closest living relatives of vertebrates. *Nature*, 439 (7079):965–968, Feb. 2006.

J. R. Dixon, S. Selvaraj, F. Yue, A. Kim, Y. Li, Y. Shen, M. Hu, J. S. Liu, and B. Ren. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature*, 485(7398):376–380, May 2012.

J. R. Dixon, I. Jung, S. Selvaraj, Y. Shen, J. E. Antosiewicz-Bourget, A. Y. Lee, Z. Ye, A. Kim, N. Rajagopal, W. Xie, Y. Diao, J. Liang, H. Zhao, V. V. Lobanenkov, J. R. Ecker, J. A. Thomson, and B. Ren. Chromatin architecture reorganization during stem cell differentiation. *Nature*, 518(7539):331–336, Feb. 2015.

J. Dostie, T. A. Richmond, R. A. Arnaout, R. R. Selzer, W. L. Lee, T. A. Honan, E. D. Rubio, A. Krumm, J. Lamb, C. Nusbaum, R. D. Green, and J. Dekker. Chromosome conformation capture carbon copy (5c): a massively parallel solution for mapping interactions between genomic elements. *Genome Res.*, 16(10): 1299–1309, Oct. 2006.

S. H. C. Duttke, S. A. Lacadie, M. M. Ibrahim, C. K. Glass, D. L. Corcoran, C. Benner, S. Heinz, J. T. Kadonaga, and U. Ohler. Human promoters are intrinsically directional. *Mol. Cell*, 57(4):674–684, Feb. 2015.

ENCODE Project Consortium, E. Birney, J. A. Stamatoyannopoulos, , et al. Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. *Nature*, 447(7146):799–816, June 2007.

P. G. Engström, H. Suzuki, N. Ninomiya, A. Akalin, L. Sessa, G. Lavorgna, A. Brozzi, L. Luzi, S. L. Tan, L. Yang, G. Kunarso, E. L.-C. Ng, S. Batalov, C. Wahlestedt, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, C. Wells, V. B. Bajic, V. Orlando, J. F. Reid, B. Lenhard, and L. Lipovich. Complex loci in human and mouse genomes. *PLoS Genet.*, 2(4):e47, Apr. 2006.

P. G. Engström, S. J. Ho Sui, O. Drivenes, T. S. Becker, and B. Lenhard. Genomic regulatory blocks underlie extensive microsynteny conservation in insects. *Genome Res.*, 17(12):1898–1908, Dec. 2007.

P. G. Engström, D. Fredman, and B. Lenhard. Ancora: a web resource for exploring highly conserved noncoding elements and their association with developmental regulatory genes. *Genome Biol.*, 9(2):R34, Feb. 2008.

C. Ernst and C. C. Morton. Identification and function of long non-coding RNA. *Front. Cell. Neurosci.*, 7:168, Oct. 2013.

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

J. Ernst, P. Kheradpour, T. S. Mikkelsen, N. Shoresh, L. D. Ward, C. B. Epstein, X. Zhang, L. Wang, R. Issner, M. Coyne, M. Ku, T. Durham, M. Kellis, and B. E. Bernstein. Mapping and analysis of chromatin state dynamics in nine human cell types. *Nature*, 473(7345):43–49, May 2011.

S. Falcon and R. Gentleman. Using GOstats to test gene lists for GO term association. *Bioinformatics*, 23(2):257–258, Jan. 2007.

FANTOM Consortium and the RIKEN PMI and CLST (DGT), A. R. R. Forrest, et al. A promoter-level mammalian expression atlas. *Nature*, 507(7493):462–470, Mar. 2014.

P. W. Finch, X. He, M. J. Kelley, A. Uren, R. P. Schaudies, N. C. Popescu, S. Rudikoff, S. A. Aaronson, H. E. Varmus, and J. S. Rubin. Purification and molecular cloning of a secreted, frizzled-related antagonist of wnt action. *Proc. Natl. Acad. Sci. U. S. A.*, 94(13):6770–6775, June 1997.

J. Fraser, I. Williamson, W. A. Bickmore, and J. Dostie. An overview of genome organization and how we got there: from FISH to Hi-C. *Microbiol. Mol. Biol. Rev.*, 79(3):347–372, Sept. 2015.

T. M. Frayling, N. J. Timpson, M. N. Weedon, E. Zeggini, R. M. Freathy, C. M. Lindgren, J. R. B. Perry, K. S. Elliott, H. Lango, N. W. Rayner, B. Shields, L. W. Harries, J. C. Barrett, S. Ellard, C. J. Groves, B. Knight, A.-M. Patch, A. R. Ness, S. Ebrahim, D. A. Lawlor, S. M. Ring, Y. Ben-Shlomo, M.-R. Jarvelin, U. Sovio, A. J. Bennett, D. Melzer, L. Ferrucci, R. J. F. Loos, I. Barroso, N. J. Wareham, F. Karpe, K. R. Owen, L. R. Cardon, M. Walker, G. A. Hitman, C. N. A. Palmer, A. S. F. Doney, A. D. Morris, G. D. Smith, A. T. Hattersley, and M. I. McCarthy. A common variant in the FTO gene is associated with body mass index and predisposes to childhood and adult obesity. *Science*, 316 (5826):889–894, May 2007.

M. Gardiner-Garden and M. Frommer. CpG islands in vertebrate genomes. *J. Mol. Biol.*, 196(2):261–282, July 1987.

K. Gaston and P.-S. Jayaraman. Transcriptional repression in eukaryotes: repressors and repression mechanisms. *Cell. Mol. Life Sci.*, 60(4):721–741, Apr. 2003.

M. Gaszner and G. Felsenfeld. Insulators: exploiting transcriptional and epigenetic mechanisms. *Nat. Rev. Genet.*, 7(9):703–713, Sept. 2006.

A. R. Gehrke, I. Schneider, E. de la Calle-Mustienes, J. J. Tena, C. Gomez-Marin, M. Chandran, T. Nakamura, I. Braasch, J. H. Postlethwait, J. L. Gómez-Skarmeta, and N. H. Shubin. Deep conservation of wrist and digit enhancers in fish. *Proc. Natl. Acad. Sci. U. S. A.*, 112(3):803–808, Jan. 2015.

R. C. Gentleman, V. J. Carey, D. M. Bates, B. Bolstad, M. Dettling, S. Dudoit, B. Ellis, L. Gautier, Y. Ge, J. Gentry, K. Hornik, T. Hothorn, W. Huber, S. Iacus, R. Irizarry, F. Leisch, C. Li, M. Maechler, A. J. Rossini, G. Sawitzki, C. Smith, G. Smyth, L. Tierney, J. Y. H. Yang, and J. Zhang. Bioconductor: open software development for computational biology and bioinformatics. *Genome Biol.*, 5(10):R80, Sept. 2004.

M. Ghandi, M. Mohammad-Noori, N. Ghareghani, D. Lee, L. Garraway, and M. A. Beer. gkmSVM: an R package for gapped-kmer SVM. *Bioinformatics*, 32(14): 2205–2207, July 2016.

Y. Ghavi-Helm, F. A. Klein, T. Pakozdi, L. Ciglar, D. Noordermeer, W. Huber, and E. E. M. Furlong. Enhancer loops appear stable during development and are associated with paused polymerase. *Nature*, 512(7512):96–100, Aug. 2014.

J. Goecks, A. Nekrutenko, J. Taylor, and Galaxy Team. Galaxy: a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol.*, 11(8):R86, Aug. 2010.

A. Goloborodko, M. V. Imakaev, J. F. Marko, and L. Mirny. Compaction and segregation of sister chromatids via active loop extrusion. *Elife*, 5, May 2016.

V. Gotea, H. M. Petrykowska, and L. Elnitski. Bidirectional promoters as important drivers for the emergence of Species-Specific transcripts. *PLoS One*, 8(2): e57323, 2013.

A. R. Gruber, S. H. Bernhart, I. L. Hofacker, and S. Washietl. Strategies for measuring evolutionary conservation of RNA secondary structures. *BMC Bioinformatics*, 9:122, Feb. 2008.

Z. Gu. EnrichedHeatmap: Making enriched heatmaps, 2017.

Y. Guo, K. Monahan, H. Wu, J. Gertz, K. E. Varley, W. Li, R. M. Myers, T. Maniatis, and Q. Wu. CTCF/cohesin-mediated DNA looping is required for protocadherin α promoter choice. *Proc. Natl. Acad. Sci. U. S. A.*, 109(51):21081–21086, Dec. 2012.

Y. Guo, Q. Xu, D. Canzio, J. Shou, J. Li, D. U. Gorkin, I. Jung, H. Wu, Y. Zhai, Y. Tang, Y. Lu, Y. Wu, Z. Jia, W. Li, M. Q. Zhang, B. Ren, A. R. Krainer, T. Maniatis, and Q. Wu. CRISPR inversion of CTCF sites alters genome topology and Enhancer/Promoter function. *Cell*, 162(4):900–910, Aug. 2015.

J. B. Gurdon. The generation of diversity and pattern in animal development. *Cell*, 68(2):185–199, Jan. 1992.

V. Haberle. seqpattern: Visualising oligonucleotide patterns and motif occurrences across a set of sorted sequences, 2015.

V. Haberle, N. Li, Y. Hadzhiev, C. Plessy, C. Previti, C. Nepal, J. Gehrig, X. Dong, A. Akalin, A. M. Suzuki, W. F. J. van IJcken, O. Armant, M. Ferg, U. Strähle, P. Carninci, F. Müller, and B. Lenhard. Two independent transcription initiation codes overlap on vertebrate core promoters. *Nature*, 507(7492):381–385, Feb. 2014.

V. Haberle, A. R. R. Forrest, Y. Hayashizaki, P. Carninci, and B. Lenhard. CAGEr: precise TSS data retrieval and high-resolution promoterome mining for integrative analyses. *Nucleic Acids Res.*, 43(8):e51, Apr. 2015.

E. Hafen, A. Kuroiwa, and W. J. Gehring. Spatial distribution of transcripts from the segmentation gene fushi tarazu during drosophila embryonic development. *Cell*, 37(3):833–841, July 1984.

N. Harmston, A. Baresic, and B. Lenhard. The mystery of extreme non-coding conservation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, 368(1632):20130021, Dec. 2013.

N. Harmston, E. Ing-Simmons, G. Tan, M. Perry, M. Merkenschlager, and B. Lenhard. Topologically associating domains are ancient features that coincide with metazoan clusters of extreme noncoding conservation. *Nat. Commun.*, 8(1):441, Sept. 2017.

B. He, C. Chen, L. Teng, and K. Tan. Global view of enhancer-promoter interactome in human cells. *Proc. Natl. Acad. Sci. U. S. A.*, 111(21):E2191–9, May 2014.

D. Hnisz, B. J. Abraham, T. I. Lee, A. Lau, V. Saint-André, A. A. Sigova, H. A. Hoke, and R. A. Young. Super-enhancers in the control of cell identity and disease. *Cell*, 155(4):934–947, Nov. 2013.

J.-W. Hong, D. A. Hendrix, and M. S. Levine. Shadow enhancers as a source of evolutionary novelty. *Science*, 321(5894):1314, Sept. 2008.

R. A. Hoskins, J. M. Landolin, J. B. Brown, J. E. Sandler, H. Takahashi, T. Lassmann, C. Yu, B. W. Booth, D. Zhang, K. H. Wan, L. Yang, N. Boley, J. Andrews, T. C. Kaufman, B. R. Graveley, P. J. Bickel, P. Carninci, J. W. Carlson, and S. E. Celniker. Genome-wide analysis of promoter architecture in drosophila melanogaster. *Genome Res.*, 21(2):182–192, Feb. 2011.

W. Huber, V. J. Carey, R. Gentleman, S. Anders, M. Carlson, B. S. Carvalho, H. C. Bravo, S. Davis, L. Gatto, T. Girke, R. Gottardo, F. Hahne, K. D. Hansen, R. A. Irizarry, M. Lawrence, M. I. Love, J. MacDonald, V. Obenchain, A. K. Oleś, H. Pagès, A. Reyes, P. Shannon, G. K. Smyth, D. Tenenbaum, L. Waldron, and M. Morgan. Orchestrating high-throughput genomic analysis with bioconductor. *Nat. Methods*, 12(2):115–121, Feb. 2015.

C. B. Hug, A. G. Grimaldi, K. Kruse, and J. M. Vaquerizas. Chromatin architecture emerges during zygotic genome activation independent of transcription. *Cell*, 169(2):216–228.e19, Apr. 2017.

R. Ihaka and R. Gentleman. R: A language for data analysis and graphics. *J. Comput. Graph. Stat.*, 5(3):299–314, Sept. 1996.

G. R. Ilsley, J. Fisher, R. Apweiler, A. H. De Pace, and N. M. Luscombe. Cellular resolution models for even skipped regulation in the entire drosophila embryo. *Elife*, 2:e00522, Aug. 2013.

K. Jabbari and G. Bernardi. Cytosine methylation and CpG, TpG (CpA) and TpA frequencies. *Gene*, 333:143–149, May 2004.

R. Jäger, G. Migliorini, M. Henrion, R. Kandaswamy, H. E. Speedy, A. Heindl, N. Whiffin, M. J. Carnicer, L. Broome, N. Dryden, T. Nagano, S. Schoenfelder, M. Enge, Y. Yuan, J. Taipale, P. Fraser, O. Fletcher, and R. S. Houlston.

Capture Hi-C identifies the chromatin interactome of colorectal cancer risk loci. *Nat. Commun.*, 6:6178, Feb. 2015.

M. Jeong, D. Sun, M. Luo, Y. Huang, G. A. Challen, B. Rodriguez, X. Zhang, L. Chavez, H. Wang, R. Hannah, S.-B. Kim, L. Yang, M. Ko, R. Chen, B. Göttgens, J.-S. Lee, P. Gunaratne, L. A. Godley, G. J. Darlington, A. Rao, W. Li, and M. A. Goodell. Large conserved domains of low DNA methylation maintained by dnmt3a. *Nat. Genet.*, 46(1):17–23, Jan. 2014.

A. Jolma, J. Yan, T. Whitington, J. Toivonen, K. R. Nitta, P. Rastas, E. Morgunova, M. Enge, M. Taipale, G. Wei, K. Palin, J. M. Vaquerizas, R. Vincentelli, N. M. Luscombe, T. R. Hughes, P. Lemaire, E. Ukkonen, T. Kivioja, and J. Taipale. DNA-binding specificities of human transcription factors. *Cell*, 152(1-2):327–339, Jan. 2013.

J. T. Kadonaga. Regulation of RNA polymerase II transcription by Sequence-Specific DNA binding factors. *Cell*, 116(2):247–257, Jan. 2004.

N. Kaplan, I. K. Moore, Y. Fondufe-Mittendorf, A. J. Gossett, D. Tillo, Y. Field, E. M. LeProust, T. R. Hughes, J. D. Lieb, J. Widom, and E. Segal. The DNA-encoded nucleosome organization of a eukaryotic genome. *Nature*, 458(7236): 362–366, Mar. 2009.

P. Kerner, A. Ikmi, D. Coen, and M. Vervoort. Evolutionary history of the iroquois/irx genes in metazoans. *BMC Evol. Biol.*, 9:74, Apr. 2009.

H. Kikuta, M. Laplante, P. Navratilova, A. Z. Komisarczuk, P. G. Engström, D. Fredman, A. Akalin, M. Caccamo, I. Sealy, K. Howe, J. Ghislain, G. Pezeron, P. Mourrain, S. Ellingsen, A. C. Oates, C. Thisse, B. Thisse, I. Foucher, B. Adolf, A. Geling, B. Lenhard, and T. S. Becker. Genomic regulatory blocks encompass multiple neighboring genes and maintain conserved synteny in vertebrates. *Genome Res.*, 17(5):545–555, May 2007.

E. Kleiman, H. Jia, S. Loguercio, A. I. Su, and A. J. Feeney. YY1 plays an essential role at all stages of b-cell differentiation. *Proc. Natl. Acad. Sci. U. S. A.*, 113 (27):E3911–20, July 2016.

R. Kodzius, M. Kojima, H. Nishiyori, M. Nakamura, S. Fukuda, M. Tagami, D. Sasaki, K. Imamura, C. Kai, M. Harbers, and Others. CAGE: cap analysis of gene expression. *Nat. Methods*, 3(3):211–222, 2006.

A. K. Kutach and J. T. Kadonaga. The downstream promoter element DPE appears to be as widely used as the TATA box in drosophila core promoters. *Mol. Cell. Biol.*, 20(13):4754–4764, July 2000.

B. Langmead, C. Trapnell, M. Pop, and S. L. Salzberg. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.*, 10(3):R25, Mar. 2009.

D. Lee, R. Karchin, and M. A. Beer. Discriminative prediction of mammalian enhancers from DNA sequence. *Genome Res.*, 21(12):2167–2180, Dec. 2011.

K.-S. Lee, H.-J. Kim, Q.-L. Li, X.-Z. Chi, C. Ueta, T. Komori, J. M. Wozney, E.-G. Kim, J.-Y. Choi, H.-M. Ryoo, and Others. Runx2 is a common target of transforming growth factor β1 and bone morphogenetic protein 2, and cooperation between runx2 and smad5 induces osteoblast-specific gene expression in the pluripotent mesenchymal precursor cell line C2C12. *Mol. Cell. Biol.*, 20 (23):8783–8792, 2000.

B. Lenhard, A. Sandelin, and P. Carninci. Metazoan promoters: emerging characteristics and insights into transcriptional regulation. *Nat. Rev. Genet.*, 13(4): 233–245, Mar. 2012.

L. A. Lettice, S. J. H. Heaney, L. A. Purdie, L. Li, P. de Beer, B. A. Oostra, D. Goode, G. Elgar, R. E. Hill, and E. de Graaff. A long-range shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly. *Hum. Mol. Genet.*, 12(14):1725–1735, July 2003.

P. P. Levings, J. Bungert The FEBS Journal, and 2002. The human βglobin locus control region. *Wiley Online Library*, 2002.

G. Li, M. J. Fullwood, H. Xu, F. H. Mulawadi, S. Velkov, V. Vega, P. N. Ariyaratne, Y. B. Mohamed, H.-S. Ooi, C. Tennakoon, C.-L. Wei, Y. Ruan, and W.-K. Sung. ChIA-PET tool for comprehensive chromatin interaction analysis with paired-end tag sequencing. *Genome Biol.*, 11(2):R22, Feb. 2010.

R. P. Lifton, M. L. Goldberg, R. W. Karp, and D. S. Hogness. The organization of the histone genes in drosophila melanogaster: functional and evolutionary implications. *Cold Spring Harb. Symp. Quant. Biol.*, 42 Pt 2:1047–1051, 1978.

G. Locke, D. Tolkunov, Z. Moqtaderi, K. Struhl, and A. V. Morozov. High-throughput sequencing reveals a simple model of nucleosome energetics. *Proceedings of the National Academy of Sciences*, 107(49):20998–21003, Dec. 2010.

H. K. Long, S. L. Prescott, and J. Wysocka. Ever-Changing landscapes: Transcriptional enhancers in development and evolution. *Cell*, 167(5):1170–1187, Nov. 2016.

D. S. Lorberbaum, A. I. Ramos, K. A. Peterson, B. S. Carpenter, D. S. Parker, S. De, L. E. Hillers, V. M. Blake, Y. Nishi, M. R. McFarlane, A. C. Chiang, J. A. Kassis, B. L. Allen, A. P. McMahon, and S. Barolo. An ancient yet flexible

cis-regulatory architecture allows localized hedgehog tuning by patched/ptch1. *Elife*, 5, May 2016.

K. M. Lower, J. R. Hughes, M. De Gobbi, S. Henderson, V. Viprakasit, C. Fisher, A. Goriely, H. Ayyub, J. Sloane-Stanley, D. Vernimmen, C. Langford, D. Garrick, R. J. Gibbons, and D. R. Higgs. Adventitious changes in long-range gene expression caused by polymorphic structural variation and promoter competition. *Proc. Natl. Acad. Sci. U. S. A.*, 106(51):21771–21776, Dec. 2009.

K. Luger, A. W. Mäder, R. K. Richmond, D. F. Sargent, and T. J. Richmond. Crystal structure of the nucleosome core particle at 2.8 a resolution. *Nature*, 389(6648):251–260, Sept. 1997.

D. G. Lupiáñez, K. Kraft, V. Heinrich, P. Krawitz, F. Brancati, E. Klopocki, D. Horn, H. Kayserili, J. M. Opitz, R. Laxova, F. Santos-Simarro, B. Gilbert-Dussardier, L. Wittler, M. Borschiwer, S. A. Haas, M. Osterwalder, M. Franke, B. Timmermann, J. Hecht, M. Spielmann, A. Visel, and S. Mundlos. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell*, 161(5):1012–1025, May 2015.

I. Maeso, M. Irimia, J. J. Tena, E. González-Pérez, D. Tran, V. Ravi, B. Venkatesh, S. Campuzano, J. L. Gómez-Skarmeta, and J. Garcia-Fernàndez. An ancient genomic regulatory block conserved across bilaterians and its dismantling in tetrapods by retrogene replacement. *Genome Res.*, 22(4):642–655, Apr. 2012.

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

A. Mathelier and W. W. Wasserman. The next generation of transcription factor binding site prediction. *PLoS Comput. Biol.*, 9(9):e1003214, Sept. 2013.

A. Mathelier, O. Fornes, D. J. Arenillas, C.-Y. Chen, G. Denay, J. Lee, W. Shi, C. Shyr, G. Tan, R. Worsley-Hunt, A. W. Zhang, F. Parcy, B. Lenhard, A. Sandelin, and W. W. Wasserman. JASPAR 2016: a major expansion and update of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.*, 44(D1):D110–5, Jan. 2016.

R. Matsuoka, M. C. Yoshida, Y. Furutani, S. Imamura, N. Kanda, M. Yanagisawa, T. Masaki, and A. Takao. Human smooth muscle myosin heavy chain gene mapped to chromosomal region 16q12. *Am. J. Med. Genet.*, 46(1):61–67, Apr. 1993.

V. Matys, E. Fricke, R. Geffers, E. Gößling, M. Haubrock, R. Hehl, K. Hornischer, D. Karas, A. E. Kel, O. V. Kel-Margoulis, and Others. TRANSFAC®: transcriptional regulation, from patterns to profiles. *Nucleic Acids Res.*, 31(1):374–378, 2003.

M. T. Maurano, R. Humbert, E. Rynes, R. E. Thurman, E. Haugen, H. Wang, A. P. Reynolds, R. Sandstrom, H. Qu, J. Brody, A. Shafer, F. Neri, K. Lee, T. Kutyavin, S. Stehling-Sun, A. K. Johnson, T. K. Canfield, E. Giste, M. Diegel, D. Bates, R. S. Hansen, S. Neph, P. J. Sabo, S. Heimfeld, A. Raubitschek, S. Ziegler, C. Cotsapas, N. Sotoodehnia, I. Glass, S. R. Sunyaev, R. Kaul, and J. A. Stamatoyannopoulos. Systematic localization of common disease-associated variation in regulatory DNA. *Science*, 337(6099):1190–1195, Sept. 2012.

M. Merkenschlager and E. P. Nora. CTCF and cohesin in genome folding and transcriptional gene regulation. *Annu. Rev. Genomics Hum. Genet.*, Apr. 2016.

B. Mifsud, F. Tavares-Cadete, A. N. Young, R. Sugar, S. Schoenfelder, L. Ferreira, S. W. Wingett, S. Andrews, W. Grey, P. A. Ewels, B. Herman, S. Happe, A. Higgs, E. LeProust, G. A. Follows, P. Fraser, N. M. Luscombe, and C. S. Osborne. Mapping long-range promoter contacts in human cells with high-resolution capture Hi-C. *Nat. Genet.*, 47(6):598–606, June 2015.

I. Miguel-Escalada, L. Pasquali, and J. Ferrer. Transcriptional enhancers: functional insights and role in human disease. *Curr. Opin. Genet. Dev.*, 33:71–76, Aug. 2015.

K. Monahan, N. D. Rudnick, P. D. Kehayova, F. Pauli, K. M. Newberry, R. M. Myers, and T. Maniatis. Role of CCCTC binding factor (CTCF) and cohesin in the generation of single-cell diversity of protocadherin-α gene expression. *Proc. Natl. Acad. Sci. U. S. A.*, 109(23):9125–9130, June 2012.

T. Montavon, N. Soshnikova, B. Mascrez, E. Joye, L. Thevenet, E. Splinter, W. de Laat, F. Spitz, and D. Duboule. A regulatory archipelago controls hox genes transcription in digits. *Cell*, 147(5):1132–1145, Nov. 2011.

M. Murata, H. Nishiyori-Sueki, M. Kojima-Ishiyama, P. Carninci, Y. Hayashizaki, and M. Itoh. Detecting expressed genes using CAGE. In E. Miyamoto-Sato, H. Ohashi, H. Sasaki, J.-I. Nishikawa, and H. Yanagawa, editors, *Transcription Factor Regulatory Networks: Methods and Protocols*, pages 67–85. Springer New York, New York, NY, 2014.

M. Murtha, Z. Tokcaer-Keskin, Z. Tang, F. Strino, X. Chen, Y. Wang, X. Xi, C. Basilico, S. Brown, R. Bonneau, Y. Kluger, and L. Dailey. FIREWACh: high-throughput functional detection of transcriptional regulatory modules in mammalian cells. *Nat. Methods*, 11(5):559–565, May 2014.

C. Nepal, Y. Hadzhiev, C. Previti, V. Haberle, N. Li, H. Takahashi, A. M. M. Suzuki, Y. Sheng, R. F. Abdelhamid, S. Anand, J. Gehrig, A. Akalin, C. E. M. Kockx, A. A. J. van der Sloot, W. F. J. van Ijcken, O. Armant, S. Rastegar, C. Watson, U. Strähle, E. Stupka, P. Carninci, B. Lenhard, and F. Müller.

Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis. *Genome Res.*, 23(11):1938–1950, Nov. 2013.

K. Nishida, M. C. Frith, and K. Nakai. Pseudocounts for transcription factor binding sites. *Nucleic Acids Res.*, 37(3):939–944, Feb. 2009.

P. Nissen, J. Hansen, N. Ban, P. B. Moore, and T. A. Steitz. The structural basis of ribosome activity in peptide bond synthesis. *Science*, 289(5481):920–930, Aug. 2000.

T. Nozaki, N. Yachie, R. Ogawa, A. Kratz, R. Saito, and M. Tomita. Tight associations between transcription promoter type and epigenetic variation in histone positioning and modification. *BMC Genomics*, 12:416, Aug. 2011.

T. O'Connor, M. Bodén, and T. L. Bailey. CisMapper: predicting regulatory interactions from transcription factor ChIP-seq data. *Nucleic Acids Res.*, Oct. 2016.

U. Ohler, G.-C. Liao, H. Niemann, and G. M. Rubin. Computational analysis of core promoters in the drosophila genome. *Genome Biol.*, 3(12): RESEARCH0087, Dec. 2002.

H. Pagès, P. Aboyoun, R. Gentleman, and S. DebRoy. Biostrings: String objects representing biological sequences, and matching algorithms, 2017.

S. C. J. Parker, M. L. Stitzel, D. L. Taylor, J. M. Orozco, M. R. Erdos, J. A. Akiyama, K. L. van Bueren, P. S. Chines, N. Narisu, NISC Comparative Sequencing Program, B. L. Black, A. Visel, L. A. Pennacchio, and F. S. Collins. Chromatin stretch enhancer states drive cell-specific gene regulation and harbor human disease risk variants. *Proceedings of the National Academy of Sciences*, 110(44):17921–17926, Oct. 2013.

T. J. Parry, J. W. M. Theisen, J.-Y. Hsu, Y.-L. Wang, D. L. Corcoran, M. Eustice, U. Ohler, and J. T. Kadonaga. The TCT motif, a key component of an RNA polymerase II transcription system for the translational machinery. *Genes Dev.*, 24(18):2013–2018, Sept. 2010.

G. Pau, F. Fuchs, O. Sklyar, M. Boutros, and W. Huber. EBImage—an R package for image processing with applications to cellular phenotypes. *Bioinformatics*, 26(7):979–981, Apr. 2010.

V. Perissi, K. Jepsen, C. K. Glass, and M. G. Rosenfeld. Deconstructing repression: evolving models of co-repressor action. *Nat. Rev. Genet.*, 11(2):109–123, Feb. 2010.

J. E. Phillips-Cremins, M. E. G. Sauria, A. Sanyal, T. I. Gerasimova, B. R. Lajoie, J. S. K. Bell, C.-T. Ong, T. A. Hookway, C. Guo, Y. Sun, M. J. Bland, W. Wagstaff, S. Dalton, T. C. McDevitt, R. Sen, J. Dekker, J. Taylor, and V. G. Corces. Architectural protein subclasses shape 3D organization of genomes during lineage commitment. *Cell*, 153(6):1281–1295, June 2013.

J. L. Platt, R. Salama, J. Smythies, H. Choudhry, J. O. J. Davies, J. R. Hughes, P. J. Ratcliffe, and D. R. Mole. Capture-C reveals preformed chromatin interactions between HIF-binding sites and distant promoters. *EMBO Rep.*, page e201642198, 2016.

C. Plessy, G. Pascarella, N. Bertin, A. Akalin, C. Carrieri, A. Vassalli, D. Lazarevic, J. Severin, C. Vlachouli, R. Simone, G. J. Faulkner, J. Kawai, C. O. Daub, S. Zucchelli, Y. Hayashizaki, P. Mombaerts, B. Lenhard, S. Gustincich, and P. Carninci. Promoter architecture of mouse olfactory receptor genes. *Genome Res.*, 22(3):486–497, Mar. 2012.

J. Ponjavic, B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki, and A. Sandelin. Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters. *Genome Biol.*, 7(8):1–18, 2006.

S. Pott and J. D. Lieb. What are super-enhancers? *Nat. Genet.*, 47(1):ng.3167, Dec. 2014.

J. Pratap, A. Javed, L. R. Languino, A. J. van Wijnen, J. L. Stein, G. S. Stein, and J. B. Lian. The runx2 osteogenic transcription factor regulates matrix metalloproteinase 9 in bone metastatic cancer cells and controls cell invasion. *Mol. Cell. Biol.*, 25(19):8581–8591, Oct. 2005.

E. A. Rach, D. R. Winter, A. M. Benjamin, D. L. Corcoran, T. Ni, J. Zhu, and U. Ohler. Transcription initiation patterns indicate divergent strategies for gene regulation at the chromatin level. *PLoS Genet.*, 7(1):e1001274, Jan. 2011.

A. Ragvin, E. Moro, D. Fredman, P. Navratilova, Ø. Drivenes, P. G. Engström, M. E. Alonso, E. de la Calle Mustienes, J. L. Gómez Skarmeta, M. J. Tavares, F. Casares, M. Manzanares, V. van Heyningen, A. Molven, P. R. Njølstad, F. Argenton, B. Lenhard, and T. S. Becker. Long-range gene regulation links genomic type 2 diabetes and obesity risk regions to HHEX, SOX4, and IRX3. *Proc. Natl. Acad. Sci. U. S. A.*, 107(2):775–780, Jan. 2010.

F. Ramírez, D. P. Ryan, B. Grüning, V. Bhardwaj, F. Kilpert, A. S. Richter, S. Heyne, F. Dündar, and T. Manke. deeptools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.*, 44(W1):W160–5, July 2016.

S. S. P. Rao, M. H. Huntley, N. C. Durand, E. K. Stamenova, I. D. Bochkov, J. T. Robinson, A. L. Sanborn, I. Machol, A. D. Omer, E. S. Lander, and E. L. Aiden. A 3D map of the human genome at kilobase resolution reveals principles of chromatin looping. *Cell*, 159(7):1665–1680, Dec. 2014.

Roadmap Epigenomics Consortium, A. Kundaje, W. Meuleman, J. Ernst, et al. Integrative analysis of 111 reference human epigenomes. *Nature*, 518(7539): 317–330, Feb. 2015.

C. S. Ross-Innes, R. Stark, A. E. Teschendorff, K. A. Holmes, H. R. Ali, M. J. Dunning, G. D. Brown, O. Gojis, I. O. Ellis, A. R. Green, S. Ali, S.-F. Chin, C. Palmieri, C. Caldas, and J. S. Carroll. Differential oestrogen receptor binding is associated with clinical outcome in breast cancer. *Nature*, 481(7381):389–393, Jan. 2012.

S. Roy, A. F. Siahpirani, D. Chasman, S. Knaack, F. Ay, R. Stewart, M. Wilson, and R. Sridharan. A predictive modeling approach for cell line-specific long-range regulatory interactions. *Nucleic Acids Res.*, 43(18):8694–8712, Oct. 2015.

H. T. Rube and J. S. Song. Quantifying the role of steric constraints in nucleosome positioning. *Nucleic Acids Res.*, 42(4):2147–2158, Feb. 2014.

A. L. Sanborn, S. S. P. Rao, S.-C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, K. P. Geeting, A. Gnirke, A. Melnikov, D. McKenna, E. K. Stamenova, E. S. Lander, and E. L. Aiden. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47):E6456–E6465, Nov. 2015a.

A. L. Sanborn, S. S. P. Rao, S.-C. Huang, N. C. Durand, M. H. Huntley, A. I. Jewett, I. D. Bochkov, D. Chinnappan, A. Cutkosky, J. Li, and Others. Chromatin extrusion explains key features of loop and domain formation in wild-type and engineered genomes. *Proceedings of the National Academy of Sciences*, 112(47): E6456–E6465, 2015b.

A. Sandelin, W. Alkema, P. Engström, W. W. Wasserman, and B. Lenhard. JASPAR: an openaccess database for eukaryotic transcription factor binding profiles. *Nucleic Acids Res.*, 32(suppl_1):D91–D94, Jan. 2004a.

A. Sandelin, P. Bailey, S. Bruce, P. G. Engström, J. M. Klos, W. W. Wasserman, J. Ericson, and B. Lenhard. Arrays of ultraconserved non-coding regions span the loci of key developmental genes in vertebrate genomes. *BMC Genomics*, 5 (1):99, Dec. 2004b.

D. P. Satijn, K. M. Hamer, J. den Blaauwen, and A. P. Otte. The polycomb group protein EED interacts with YY1, and both proteins induce neural tissue in xenopus embryos. *Mol. Cell. Biol.*, 21(4):1360–1369, Feb. 2001.

A. N. Schep, J. D. Buenrostro, S. K. Denny, K. Schwartz, G. Sherlock, and W. J. Greenleaf. Structured nucleosome fingerprints enable high-resolution mapping of chromatin architecture within regulatory regions. *Genome Res.*, Aug. 2015.

T. D. Schneider and R. M. Stephens. Sequence logos: a new way to display consensus sequences. *Nucleic Acids Res.*, 18(20):6097–6100, Oct. 1990.

S. Schoenfelder, M. Furlan-Magaril, B. Mifsud, F. Tavares-Cadete, R. Sugar, B.-M. Javierre, T. Nagano, Y. Katsman, M. Sakthidevi, S. W. Wingett, E. Dimitrova, A. Dimond, L. B. Edelman, S. Elderkin, K. Tabbada, E. Darbo, S. Andrews, B. Herman, A. Higgs, E. LeProust, C. S. Osborne, J. A. Mitchell, N. M. Luscombe, and P. Fraser. The pluripotent regulatory circuitry connecting promoters to their long-range interacting elements. *Genome Res.*, 25(4):582–597, Apr. 2015.

A. Sebé-Pedrós, B. M. Degnan, and I. Ruiz-Trillo. The origin of metazoa: a unicellular perspective. *Nat. Rev. Genet.*, 18(8):nrg.2017.21, May 2017.

E. Segal, Y. Fondufe-Mittendorf, L. Chen, A. Thåström, Y. Field, I. K. Moore, J.-P. Z. Wang, and J. Widom. A genomic code for nucleosome positioning. *Nature*, 442(7104):772–778, July 2006.

L. Shen, N. Shao, X. Liu, and E. Nestler. ngs.plot: Quick mining and visualization of next-generation sequencing data by integrating genomic databases. *BMC Genomics*, 15:284, Apr. 2014.

Y. Shi, J. S. Lee, and K. M. Galvin. Everything you have ever wanted to know about yin yang 1.. *Biochim. Biophys. Acta*, 1332(2):F49–66, Apr. 1997.

T. Shiraki, S. Kondo, S. Katayama, K. Waki, T. Kasukawa, H. Kawaji, R. Kodzius, A. Watahiki, M. Nakamura, T. Arakawa, S. Fukuda, D. Sasaki, A. Podhajska, M. Harbers, J. Kawai, P. Carninci, and Y. Hayashizaki. Cap analysis gene expression for high-throughput analysis of transcriptional starting point and identification of promoter usage. *Proc. Natl. Acad. Sci. U. S. A.*, 100(26):15776–15781, Dec. 2003.

L. E. Sidney, M. J. Branch, S. E. Dunphy, H. S. Dua, and A. Hopkinson. Concise review: evidence for CD34 as a common marker for diverse progenitors. *Stem Cells*, 32(6):1380–1389, June 2014.

M. Siebert and J. Söding. Bayesian markov models consistently outperform PWMs at predicting motifs in nucleotide sequences. *Nucleic Acids Res.*, June 2016.

A. Siepel, G. Bejerano, J. S. Pedersen, A. S. Hinrichs, M. Hou, K. Rosenbloom, H. Clawson, J. Spieth, L. W. Hillier, S. Richards, G. M. Weinstock, R. K. Wilson, R. A. Gibbs, W. J. Kent, W. Miller, and D. Haussler. Evolutionarily

conserved elements in vertebrate, insect, worm, and yeast genomes. *Genome Res.*, 15(8):1034–1050, Aug. 2005.

D. Simcha, N. D. Price, and D. Geman. The limits of de novo DNA motif discovery. *PLoS One*, 7(11):e47836, Nov. 2012.

S. Smemo, L. C. Campos, I. P. Moskowitz, J. E. Krieger, A. C. Pereira, and M. A. Nobrega. Regulatory variation in a TBX5 enhancer leads to isolated congenital heart disease. *Hum. Mol. Genet.*, 21(14):3255–3263, July 2012.

S. Smemo, J. J. Tena, K.-H. Kim, E. R. Gamazon, N. J. Sakabe, C. Gómez-Marín, I. Aneas, F. L. Credidio, D. R. Sobreira, N. F. Wasserman, J. H. Lee, V. Puviindran, D. Tam, M. Shen, J. E. Son, N. A. Vakili, H.-K. Sung, S. Naranjo, R. D. Acemel, M. Manzanares, A. Nagy, N. J. Cox, C.-C. Hui, J. L. Gomez-Skarmeta, and M. A. Nóbrega. Obesity-associated variants within FTO form long-range functional connections with IRX3. *Nature*, 507(7492):371–375, Mar. 2014.

F. Spitz and E. E. M. Furlong. Transcription factors: from enhancer binding to developmental control. *Nat. Rev. Genet.*, 13(9):613–626, Sept. 2012.

G. D. Stormo. DNA binding sites: representation and discovery. *Bioinformatics*, 16(1):16–23, Jan. 2000.

G. D. Stormo, T. D. Schneider, L. Gold, and A. Ehrenfeucht. Use of the perceptron algorithm to distinguish translational initiation sites in e. coli. *Nucleic Acids Res.*, 10(9):2997–3011, May 1982.

K. Struhl and E. Segal. Determinants of nucleosome positioning. *Nat. Struct. Mol. Biol.*, 20(3):267–273, Mar. 2013.

G. C. A. Taylor, R. Eskeland, B. Hekimoglu-Balkan, M. M. Pradeepa, and W. A. Bickmore. H4K16 acetylation marks active genes and enhancers of embryonic stem cells, but does not alter chromatin compaction. *Genome Res.*, 23(12): 2053–2065, Dec. 2013.

M. C. Thomas and C.-M. Chiang. The general transcription machinery and general cofactors. *Crit. Rev. Biochem. Mol. Biol.*, 41(3):105–178, May 2006.

R. Tibshirani. Regression shrinkage and selection via the lasso. *J. R. Stat. Soc. Series B Stat. Methodol.*, 58(1):267–288, 1996.

N. D. Trinklein, S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar, and R. M. Myers. An abundance of bidirectional promoters in the human genome. *Genome Res.*, 14(1):62–66, Jan. 2004.

N. L. van Berkum, E. Lieberman-Aiden, L. Williams, M. Imakaev, A. Gnirke, L. A. Mirny, J. Dekker, and E. S. Lander. Hi-C: a method to study the three-dimensional architecture of genomes. *J. Vis. Exp.*, (39), May 2010.

D. Villar, C. Berthelot, S. Aldridge, T. F. Rayner, M. Lukk, M. Pignatelli, T. J. Park, R. Deaville, J. T. Erichsen, A. J. Jasinska, J. M. A. Turner, M. F. Bertelsen, E. P. Murchison, P. Flicek, and D. T. Odom. Enhancer evolution across 20 mammalian species. *Cell*, 160(3):554–566, Jan. 2015.

A. Visel, S. Minovitsky, I. Dubchak, and L. A. Pennacchio. VISTA enhancer browser–a database of tissue-specific human enhancers. *Nucleic Acids Res.*, 35 (Database issue):D88–92, Jan. 2007.

G. P. Wagner, K. Kin, and V. J. Lynch. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci.*, 131(4):281–285, Dec. 2012.

M. Wand. KernSmooth: Functions for kernel smoothing supporting wand & jones (1995), 2015.

M. P. Wand. Fast computation of multivariate kernel estimators. *J. Comput. Graph. Stat.*, 3(4):433–445, 1994.

M. N. Weedon, I. Cebola, A.-M. Patch, S. E. Flanagan, E. De Franco, R. Caswell, S. A. Rodríguez-Seguí, C. Shaw-Smith, C. H.-H. Cho, H. L. Allen, J. A. Houghton, C. L. Roth, R. Chen, K. Hussain, P. Marsh, L. Vallier, A. Murray, International Pancreatic Agenesis Consortium, S. Ellard, J. Ferrer, and A. T. Hattersley. Recessive mutations in a distal PTF1A enhancer cause isolated pancreatic agenesis. *Nat. Genet.*, 46(1):61–64, Jan. 2014.

R. Wehrens, L. M. C. Buydens, and Others. Self-and super-organizing maps in r: the kohonen package. *J. Stat. Softw.*, 21(5):1–19, 2007.

W. Wei, V. Pelechano, A. I. Järvelin, and L. M. Steinmetz. Functional consequences of bidirectional promoters. *Trends Genet.*, 27(7):267–276, July 2011.

W. A. Whyte, D. A. Orlando, D. Hnisz, B. J. Abraham, C. Y. Lin, M. H. Kagey, P. B. Rahl, T. I. Lee, and R. A. Young. Master transcription factors and mediator establish super-enhancers at key cell identity genes. *Cell*, 153(2):307–319, Apr. 2013.

J. Widom. Toward a unified model of chromatin folding. *Annu. Rev. Biophys. Biophys. Chem.*, 18:365–395, 1989.

F. H. Wilkinson, K. Park, and M. L. Atchison. Polycomb recruitment to DNA in vivo by the YY1 REPO domain. *Proc. Natl. Acad. Sci. U. S. A.*, 103(51): 19296–19301, Dec. 2006.

A. Woolfe, M. Goodson, D. K. Goode, P. Snell, G. K. McEwen, T. Vavouri, S. F. Smith, P. North, H. Callaway, K. Kelly, K. Walter, I. Abnizova, W. Gilks, Y. J. K. Edwards, J. E. Cooke, and G. Elgar. Highly conserved non-coding sequences are associated with vertebrate development. *PLoS Biol.*, 3(1):e7, Jan. 2005.

H. Wu, B. Caffo, H. A. Jaffee, R. A. Irizarry, and A. P. Feinberg. Redefining CpG islands using hidden markov models. *Biostatistics*, 11(3):499–514, July 2010.

Q. Wu and T. Maniatis. A striking organization of a large family of human neural cadherin-like cell adhesion genes. *Cell*, 97(6):779–790, June 1999.

H. Xi, Y. Yu, Y. Fu, J. Foley, A. Halees, and Z. Weng. Analysis of overrepresented motifs in human core promoters reveals dual regulatory roles of YY1. *Genome Res.*, 17(6):798–806, June 2007.

Z. Xu, W. Wei, J. Gagneur, F. Perocchi, S. Clauder-Münster, J. Camblong, E. Guffanti, F. Stutz, W. Huber, and L. M. Steinmetz. Bidirectional promoters generate pervasive transcription in yeast. *Nature*, 457(7232):1033–1037, Feb. 2009.

R. Yamashita, Y. Suzuki, S. Sugano, and K. Nakai. Genome-wide analysis reveals strong correlation between CpG islands with nearby transcription start sites of genes and their tissue specificity. *Gene*, 350(2):129–136, May 2005.

C. Yang, E. Bolotin, T. Jiang, F. M. Sladek, and E. Martinez. Prevalence of the initiator over the TATA box in human and yeast genes and identification of DNA motifs enriched in human TATA-less core promoters. *Gene*, 389(1):52–65, Mar. 2007.

R. S. Young, Y. Kumar, W. A. Bickmore, and M. S. Taylor. Bidirectional transcription marks accessible chromatin and is not specific to enhancers. Jan. 2016.

M. A. Zabidi, C. D. Arnold, K. Schernhuber, M. Pagani, M. Rath, O. Frank, and A. Stark. Enhancer-core-promoter specificity separates developmental and housekeeping gene regulation. *Nature*, 518(7540):556–559, Feb. 2015.

Y. Zhang, T. Liu, C. A. Meyer, J. Eeckhoute, D. S. Johnson, B. E. Bernstein, C. Nusbaum, R. M. Myers, M. Brown, W. Li, and X. S. Liu. Model-based analysis of ChIP-Seq (MACS). *Genome Biol.*, 9(9):R137, Sept. 2008.

Z. Zhao, G. Tavoosidana, M. Sjölinder, A. Göndör, P. Mariano, S. Wang, C. Kanduri, M. Lezcano, K. S. Sandhu, U. Singh, V. Pant, V. Tiwari, S. Kurukuti, and R. Ohlsson. Circular chromosome conformation capture (4c) uncovers extensive networks of epigenetically regulated intra- and interchromosomal interactions. *Nat. Genet.*, 38(11):1341–1347, Nov. 2006.

Y. Zhu, Z. Chen, K. Zhang, M. Wang, D. Medovoy, J. W. Whitaker, B. Ding, N. Li, L. Zheng, and W. Wang. Constructing 3D interaction maps from 1D epigenomes. *Nat. Commun.*, 7:10812, Mar. 2016.