Characterisation of the first gene promoters to be activated in zebrafish

By

Ledean Angel Cooper

A thesis submitted to
The University of Birmingham
For the degree of
DOCTOR OF PHILOSOPHY

Institute of Cancer and Genomic Sciences
College of Medical and Dental Sciences
University of Birmingham, UK
June 2020

UNIVERSITY OF
BIRMINGHAM

University of Birmingham Research Archive

e-theses repository

This unpublished thesis/dissertation is copyright of the author and/or third parties. The intellectual property rights of the author or third parties in respect of this work are as defined by The Copyright Designs and Patents Act 1988 or as modified by any successor legislation.

Any use made of information contained in this thesis/dissertation must be in accordance with that legislation and must be properly acknowledged. Further distribution or reproduction in any format is prohibited without the permission of the copyright holder.

Abstract

Prior to onset of zygotic genome activation, the genome is held in a state of strict transcriptional repression. Once a threshold nucleo-cytoplasmic (NC) ratio is reached repressive maternally deposited factors are suitably diluted, which in combination with degradation of maternal RNAs by miR-430, provides an environment amenable to global transcription which generally coincides with the mid-blastula transition (MBT) in most animals. Despite this, increasing reports suggest a group of genes are activated before the onset of global zygotic genome activation, with the earliest genes being activated through an as of yet unidentified mechanism.

In order to study the processes that drive expression of a specific subset of genes prior to genome activation, it was necessary to characterise the promoter structure of early expressed genes, of which miR-430 is the most highly and earliest expressed. Identification of characteristic profiles for *cis*-regulatory elements, such as core and proximal promoter, was done for the purpose of understanding how promoter structure influences the first transcriptional activation in the developing vertebrate embryo.

Here I demonstrate that early genes have a promoter structure distinct from other developmental classes, which suggests they are activated by a distinct mechanism. In regard to miR-430, I have provided evidence for a distinct site with high local concentration of *cis-regulatory* elements with distinctive epigenetic profile, even compared to other early expressed genes.

Dedication

I would like to dedicate this thesis to my daughter.

Acknowledgements

First and foremost I would like to thank my supervisor Professor Ferenc Mueller for providing me with the opportunity to undertake this PhD and all of the support and guidance he has provided since. I am also thankful to the Midlands Integrative Biosciences Training Partnership's (MIBTP) highly informative and comprehensive doctoral training programming that has supported my career development from the onset all the way through to the end of my doctoral studies.

I would also like to express my gratitude to Shawn Burgess and Andrew Beggs for providing PacBio and Nanopore long read sequencing data.

I will be forever grateful to Dr Yavor Hadzhiev for his support, advice and guidance.

Table of Contents

1 GENERAL INTRODUCTION 1

1.1 Promoter codes and expression dynamics 1

1.1.1 Core promoter elements 1

1.1.1.1 TATA-box, TBP and its related factors 4

1.1.1.2 Initiator, BRE and organism specific CPEs 6

1.1.2 Additional *cis-regulatory* elements 7

1.1.2.1 Enhancers 8

1.1.2.2 Silencers 9

1.1.2.3 Insulators 10

1.1.2.4 Proximal promoter 10

1.1.3 Promoter regulation in the third dimension 11

1.1.3.1 3D genome organisation 11

1.1.3.2 Chromatin architecture and epigenetic modifications 15

1.2 Identification of promoter architecture 22

1.2.1 Introduction to computational analysis of transcription codes and dynamics 22

1.2.2 Methods for precise determination of transcription start sites 22

1.2.3 Classification of promoter types 25

1.2.3.1 Nucleotide composition in core promoters 26

1.2.3.2 Initiation patterns and the TSS 28

1.3 Zebrafish as a model organism 30

1.3.1 Developmental stages of the zebrafish 30

1.4 Mechanism for transcription initiation in early zebrafish embryos 34

1.4.1 The Midblastula Transition 34

1.4.2 Nucleo-cytoplasmic (N/C) ratio 37

1.4.3 Role of chromatin competency in transcription initiation 38

1.4.4 Pioneer and stem cell factors role in transcription initiation in early zebrafish embryos 40

1.4.5 Maternal mRNA degradation 44

1.4.6 Structure and function of Micro RNAs 46

1.4.7 miR-430 is fundamental for early development 46

1.4.8 The first zygotic transcription occurs via an unknown mechanism 48

1.4.9 Transcription of active genes occurs in distinct sub-nuclear compartments 50

1.5 Analysis of dynamics of gene regulation by visualisation of transcription 52

1.5.1 Methods for visualising transcription 52

1.6 Aims and objectives 57

2 MATERIALS AND METHODS 58

2.1 Materials 58

2.1.1 Reagents, Kits and Equipment 58

2.1.2 Buffers and solutions 61

2.2 Methods 62

2.2.1 Processing and analyses of Capped Analyses of Gene Expression sequencing (CAGE-seq) data. 62

2.2.2 Processing of RNA sequencing (RNA-seq) data. 63

2.2.3 Processing of ChIP-seq and ATAC-seq data. 64

2.2.4 Selection of genes sets for epigenetic features comparisons 66

2.2.5 CpG Island Analysis 68

2.2.6 Transcription factor binding motif analyses 68

2.2.7 Annotation and SNP detection of miR430 promoter and design of morpholino oligonucleotides for transcription visualisation. 69

2.2.8 Zebrafish maintenance 70

2.2.9 Dechorionation of zebrafish eggs 71

2.2.10 Microinjection of zebrafish embryos 71

2.2.11 Isolation of zebrafish genomic DNA 72

2.2.12 Isolation of plasmid DNA 72

2.2.13 Polymerase Chain Reaction (PCR) 73

2.2.14 Restriction digest 73

2.2.15 Agarose Gel electrophoresis 73

2.2.16 Cloning of mir430 promoter region 74

2.2.17 Synthesis of antisense RNA probes for *in situ* hybridization 76

2.2.18 *In vitro* transcription of *PhiC31* mRNA 77

2.2.19 Whole mount in situ hybridisation 77

2.2.20 Generation of stable transgenic zebrafish reporter lines. 79

2.2.21 Live embryo imaging. 80

3 CHARACTERISATION OF A MODEL GENE FOR REAL-TIME VISUALISATION OF TRANSCRIPTION IN THE DEVELOPING EMBRYO 81

3.1 Introduction 81

3.2 Aims 83

3.3 Computational analysis of *mir-430* locus for model validation 84

3.3.1 Mapping of NGS data revealed 8 functional promoters 85

3.3.2 *miR-430* locus has a classical promoter architecture 90

3.3.3 Temporal dynamics of miR-430 expression during ZGA 93

3.4 A model for spatio-temporal visualization of transcription initiation during early development *in vivo* 97

3.5 Discussion 101

4 CHARACTERISATION OF THE CHROMATIN STATE AND MODIFICATION DYNAMICS OF THE MIR-430 CLUSTER 104

4.1 Temporal dynamics of chromatin states and transcription 104

4.2 Aims 107

4.3 Analysis of chromatin state on a revised miR-430 cluster 108

4.4 Discussion 133

5 CHARACTERISATION OF PROMOTER SEQUENCE FEATURES AND MOTIFS OF EARLY EXPRESSED GENES 137

5.1 Stem cell factors are vital for onset of ZGA 137

5.2 Aims 141

5.3 Early promoter classification through analysis of core promoter elements present in first wave genes 142

5.4 Discussion 158

6 EARLY GENE PROMOTERS: DOES SEQUENCE OR ENVIRONMENT FACILITATE FIRST WAVE GENE ACTIVITY? 163

6.1 Aims 163

6.2 Functional validation of miR-430 promoter sequence in transient transgenic assay. 164

6.2.1 Generation of stable transgenic lines with the miR-430 promoter reporter constructs utilising the PhiC31 targeted integration system. 172

6.2.2 miR-430 promoter does not show activity when integrated in single-copy at an alternative genomic location 175

6.3 Discussion 179

7 SUMMARY AND CONCLUSIONS 184

7.1 Identification and characterisation of the promoter of the first genes expressed in the zebrafish embryo 184

7.2 Promoters of first genes to be activated in the embryo are distinct from other developmental classes 188

7.3 Unique multicopy environment may facilitate miR-430 activation as first gene to be expressed in the zebrafish embryo 189

8 APPENDIX 192

8.1 Public NGS dataset used in the analyses 192

8.2 Morpholino oligonucleotides 197

8.3 Position weight matrices used (from JASPAR database) 198

8.4 First Wave gene list 199

8.5 Main Wave gene list 206

8.6 Constitutive gene list 210

8.7 Post gastrulation gene list 214

8.8 Plasmid maps 218

9 REFERENCES 222

10 LIST OF PUBLICATIONS 242

List of figures

Figure 1-1 Schematic showing model of core promoter with cis-regulatory elements and relevant binding factors 3

Figure 1-2 A summary of interactions between various promoter elements and regulatory signals. 8

Figure 1-3 Hierarchical organisation of the genome in the nucleus 13

Figure 1-4 Chromosomes represent the highest level of chromatin organisation in the nucleus 16

Figure 1-5 two principal methods for analysing chromatin accessibility 17

Figure 1-6 Representation of distribution of chromatin states and dynamics across the genome 19

Figure 1-7 Mapping CAGE tags to a reference genome 24

Figure 1-8 Three types of major core promoter defined in metazoan 26

Figure 1-9 A visual representation of Zebrafish developmental stages. 33

Figure 1-10 Factors affecting activation of the zygotic genome 36

Figure 1-11 Zygotic Genome Activation in metazoa 45

Figure 1-12 reporter-based visualisation of modified RNAs 54

Figure 1-13 in vivo imaging of transcription in the Drosophila embryo 55

Figure 3-1 miR-430 endogenous locus structure from assembly GRCz11 85

Figure 3-2 Demonstration of differences in CAGE-seq coverage between unique and multi-mapping approaches 87

Figure 3-3 Overview of miR-430 locus structure showing mapped CAGE tags demonstrating 8 predicted promoters 89

Figure 3-4 miR-430 is representative of a highly repetitive, gene dense, early expressed locus 91

Figure 3-5 Temporal dynamics of miR-430 expression during development 94

Figure 3-6 Evidence for mir-430 transcription pre-MBT 95

Figure 3-7 a model for visualisation of transcription in real-time during early genome activation 98

Figure 3-8 miR-430 MOs specifically bind to miR-430 transcript 99

Figure 3-9 miR-430 expression across the maternal to zygotic transition 100

Figure 4-1 Schematic of miR-430 cluster assembled from long read DNA-seq 109

Figure 4-2 Expression dynamics for first wave and main wave gene sets 113

Figure 4-3 Expression dynamics for constitutive and post gastrulation gene sets ... 114

Figure 4-4 Summary of dynamics of epigenomic data at the miR-430 assembly 115

Figure 4-5 Summary of dynamics of epigenomic data on example first wave gene. 116

Figure 4-6 Summary of dynamics of epigenomic data on example main wave gene 117

Figure 4-7 Summary of dynamics of epigenomic data on example post gastrulation gene 118

Figure 4-8 Summary of dynamics of epigenomic data on example constitutive gene 119

Figure 4-9 ATAC-seq dynamics and distribution relative to TSS 122

Figure 4-10 H3K27ac dynamics and distribution relative to TSS 124

Figure 4-11 H3K4me3 dynamics and distribution relative to TSS 128

Figure 4-12 first wave and constitutive genes have a higher occurrence of promoter associated CpG islands in comparison to other gene sets. 129

Figure 4-13 H3K27me3 dynamics and distribution relative to TSS 131

Figure 4-14 summary of epigenetic mark dynamics 132

Figure 5-1 Position motifs composition around the TSS in promoter sequences show enrichment in TATA-box for first wave genes 143

Figure 5-2 TBP binding site motif enrichment around the TSS in promoter sequences of first wave genes 146

Figure 5-3 Transcription initiation patterns characteristic for known promoter architectures 147

Figure 5-4 Distribution of promoter architectures reveals predominately sharp initiation for first wave genes 149

Figure 5-5 transcription factor binding site motif enrichment around the TSS in promoter sequences of first wave genes 152

Figure 5-6 transcription factor Klf4 binding site motif is positionally constrained at the TSS of first wave genes 155

Figure 5-7 Stem cell factor binding motif distribution on miR-430 cluster..... 157

Figure 6-1 Schematic demonstrating predicted promoter region from miR-430 locus used for functional validation...... 165

Figure 6-2 in transient reporter fluorescence detected post-MBT in vivo..... 166

Figure 6-3 Demonstration of early activation of miR-430 promoter in comparison to late activated gata2a..... 168

Figure 6-4 Lack of temporal control of transcription from episomal DNA before the main wave of zygotic genome activation..... 169

Figure 6-5 Schematic representation of PhiC31 integrase transgenesis system in the zebrafish model organism..... 174

Figure 6-6 transgenic PhiC31 UAS/Gal4 docking lines..... 175

Figure 6-7 lack of miR-430 promoter expression in established transgenic lines 177

List of tables

Table 6-1 Expression frequency of transient injected embryos by WISH 170

Table 6-2 Expression frequency of stable integrated transgenic embryos by WISH . 178

List of abbreviations

BAC	Bacterial artificial chromosome
bp	Base pair
ChIP-seq	Chromatin immunoprecipitation-sequencing
CPE	Core promoter element
DBD	DNA binding domain
DNA	deoxyribonucleic acid
Dpf	Days post fertilisation
GFP	Green fluorescent protein
GTF	General transcripttion factor
H3K27ac	Histone H3 lysine 27 acetylation
H3K27me3	Histone H3 lysine 27 trimethylation
H3K4me1	Histone H3 lysine 4 monomethylation
HEP	Highly expressed protein
hpf	Hours post fertilisation
inr	Initiator
kb	Kilo base pairs
Mb	Mega base pairs
MBT	Midblastula transition
NEB	New England Biolab
PBST	Phosphate buffered saline with Tween-20
PCR	Polymerase chain reaction
PIC	Pre-initiation complex

SDS	Sodium dodecyl sulfate
SHH	Sonic hedgehog
TBP	TATA-binding protein
TF	Transcription factor
TSS	Transcription start site
UCSC	University of California Santa Cruz
UTR	Untranslated region
ZGA	Zygotic genome activation

1 GENERAL INTRODUCTION

1.1 Promoter codes and expression dynamics

Correct spatial and temporal control of gene expression determines the identity of cells and tissues in a developing organism. Gene expression involves transcribing the DNA sequence of a gene into an RNA transcript via an RNA polymerase enzyme. In *Metazoa*, all protein-coding genes (and many non-coding genes) are transcribed by RNA polymerase II (Pol II). Integral to the regulation of transcription are promoters, which function locally, with more distal acting elements such as enhancers, silencers, and insulators: sequence-based DNA elements encoded in the genome, collectively classified as *cis*-regulatory elements (CRE). The reader is directed to chapter 5 of Alberts (2015) for a comprehensive review of DNA replication.

1.1.1 Core promoter elements

The core promoter is a DNA sequence extending approximately 40 base pairs up and downstream of the start site of transcription (TSS) (see **Figure 1-1** A). It encodes the information necessary for organised binding of gene specific regulatory factors and general transcription factors (GTFs) required for initiation of transcription (Weil et al. 1979, Goodrich et al. 2010). Six GTF complexes: TFIIA, B, D, E, F and H assemble at the core promoter in combination with RNA polymerase 2 (Pol II) (see **Figure 1-1** B-C) (Weil et al. 1979) and several other

factors, including chromatin remodellers and large coactivators (Med/ARC) to form the pre-initiation complex (PIC) (Roeder 1996, Cramer 2002, Goodrich et al. 2010). Many of the sequence motifs, the so-called core promoter elements (CPEs) required for the binding of these factors are constrained in their position, yet remain interchangeable within the promoter (Gershenzon et al. 2005).

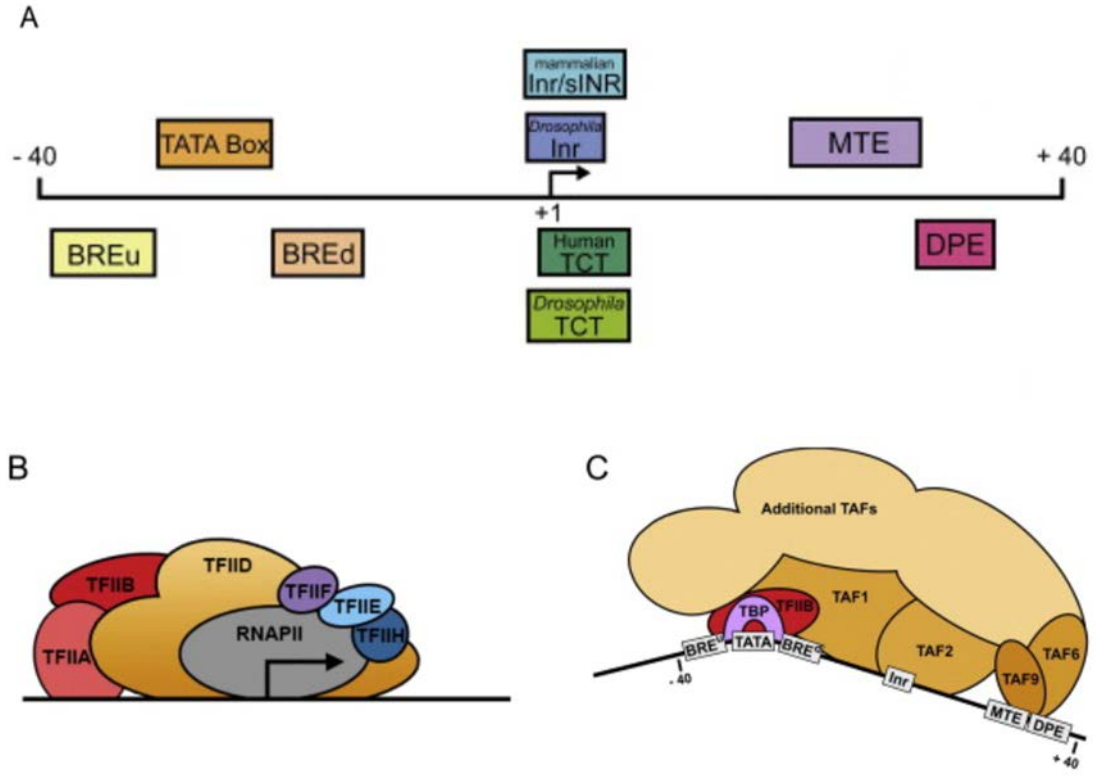


Figure 1-1 Schematic showing model of core promoter with cis-regulatory elements and relevant binding factors

(A) *Core promoter region with position of various core promoter elements relative to +1 TSS. BRE refers to B recognition element with 'u' and 'd' representing up and downstream variants respectively, Inr refers to the initiator i.e. the motif containing the start site of transcription intitation. MTE refers to motif ten element and DPE to the downstream promoter element.* ***(B)*** *Assembly and composition of PIC at Pol II promoter* ***(C)*** *Zoom in from* ***(B)*** *of proposed assembly of factors binding core promoter element from two PIC components (TFIIB and TFIID). Adapted from Danino et al. (2015).*

1.1.1.1 TATA-box, TBP and its related factors

One of the most well studied core promoter elements is the TATA-box. First discovered by Goldberg (1979), it is one of the three canonical or core CPEs, located approximately 30 base pairs (bp) upstream of the TSS (Ponjavic et al. 2006) and named for the 'TATAWAW' motif present in the consensus sequence (Hahn et al. 1989, Hahn et al. 1989), which is recognised and bound by both TFIIB and TFIID complexes **(Figure 1-1)** (Wang et al. 1996). The TATA-box bound 'TATA-binding protein' (TBP) associates with initiator interacting TATA-associated factors (TAFs) to form the TFIID subunit; first successfully characterised, in detail, in 1988 in the yeast model organism (Buratowski et al. 1988, Cavallini et al. 1988). TFIID acts as a scaffold for other GTFs to assemble at the core promoter (reviewed in (Albright et al. 2000)).

It was once accepted that TBP is a universal regulator of transcription, as it was found to be associated with transcription originating from all three RNA polymerases, despite each being dedicated to the transcription of different classes of genes (Hernandez 1993). These observations were later clarified by the discovery of 2 distinct subtypes of TBP-like factors. TBP-like factor (TLF, also named TRF2, TLP or TBPL1), which is broadly conserved across Metazoa (Dantonel et al. 1999, Maldonado 1999, Moore et al. 1999, Ohbayashi et al. 1999, Rabenstein et al. 1999) and TATA-binding protein 2 (TBP2, also named TRF3 or TBPL2) which is specific to vertebrates (Persengiev et al. 2003, Bartfai et al. 2004, Jallow et al. 2004). A third subtype, TBP-related factor (TRF1), localises to the central nervous system and male gonads and was found to associate with

the TATA-box (*in vitro)* (Holmes et al. 2000), although so far it has only been identified in Insects (Crowley et al. 1993, Hansen et al. 1997).

In regard to TBP2, residues involved in binding DNA and other GTFs are highly conserved between both TBP and TRF1, likely allowing direct association with the TATA-box (Zhao et al. 2002). As with all TBP-family members, the C- terminal domain is well conserved while the N- terminal domain is variable between members of the TBP2 homologues (Persengiev et al. 2003, Bartfai et al. 2004, Jallow et al. 2004). TLF on the other hand, differs in physical shape to TBP and as a result is the only known member of the TBP family not to associate with the TATA-box, instead is implicated in novel functional interactions with non-canonical promoter classes (Dantonel et al. 1999). For instance, TLF-mediated inhibition of transcription was detected at TATA-containing promoters, (Moore et al. 1999) yet in chicken, TLF had an additional role in activating TATA-independent gene promoters (Ohbayashi et al. 2003). TLF binding to non-canonical promoters is achieved through association with other proteins, such as TFIIA and TFIIB (Hansen et al. 1997, Ohbayashi et al. 2003), by which TLF is able to seed recruitment of other components of the PIC to the core promoter.

However not all transcriptional regulatory mechanisms rely on the TBP protein family. A TPB-free TAF-containing complex (TATC) is able to drive assembly of the PIC at TATA-containing promoters independent of either TBP-family factors or TFIID on naked DNA *in vitro* (Wieczorek et al. 1998). More recently, expression of a subset of TBP-family insensitive (TFI), genes functionally related to

mesoderm and organizer expression and function were observed in absence of TBP and its related factors during *Xenopus* development, suggesting that expression must be driven by an alternative mechanism (Gazdag et al. 2016). These findings demonstrate how a machinery once believed to be invariant, actually has a high degree of divergence, implicit in regulation of transcription across the organism throughout its lifetime.

1.1.1.2 Initiator, BRE and organism specific CPEs

Despite being one of the three canonical CPEs, the TATA-box is only found in roughly 10-20% of eukaryotic genes (Ohler et al. 2002, Basehoar et al. 2004, Jin et al. 2006). The additional two canonical CPEs: the initiator (INR) and downstream B recognition element (BRE – bound by TFIIB) (Deng et al. 2005) are located at, and upstream (Lagrange et al. 1998) of the TSS respectively, reviewed in (Kadonaga 2002). There are many additional CPEs specific to different organisms, some of which include: the *Drosophila*-specific downstream promoter element (DPE) (Burke et al. 1996), motif ten element (MTE) (Lim et al. 2004), DNA recognition element (Hirose et al. 1993) and a variation of the pyrimidine-purine INR 'C-A' DNA sequence to a 'T-C-T' pyrimidine based one (Parry et al. 2010) (Positions of various aforementioned CPEs within the core promoter and binding of listed GTFs are illustrated in **Figure 1-1**).

1.1.2 Additional *cis-regulatory* elements

Spatiotemporal gene regulation is influenced by non-coding stretches of DNA termed collectively *cis-regulatory* sequences. The core promoter itself is often not sufficient to efficiently initiate transcription and additional interaction with proximal *cis*-regulatory modules and/or long-range enhancers is required (Roeder 1991, Wray et al. 2003). Four key *cis*-regulatory elements: enhancers, silencers, insulators and the proximal promoter are fundamental for regulation of genes (**Figure 1-2**) and will be discussed in the following section.

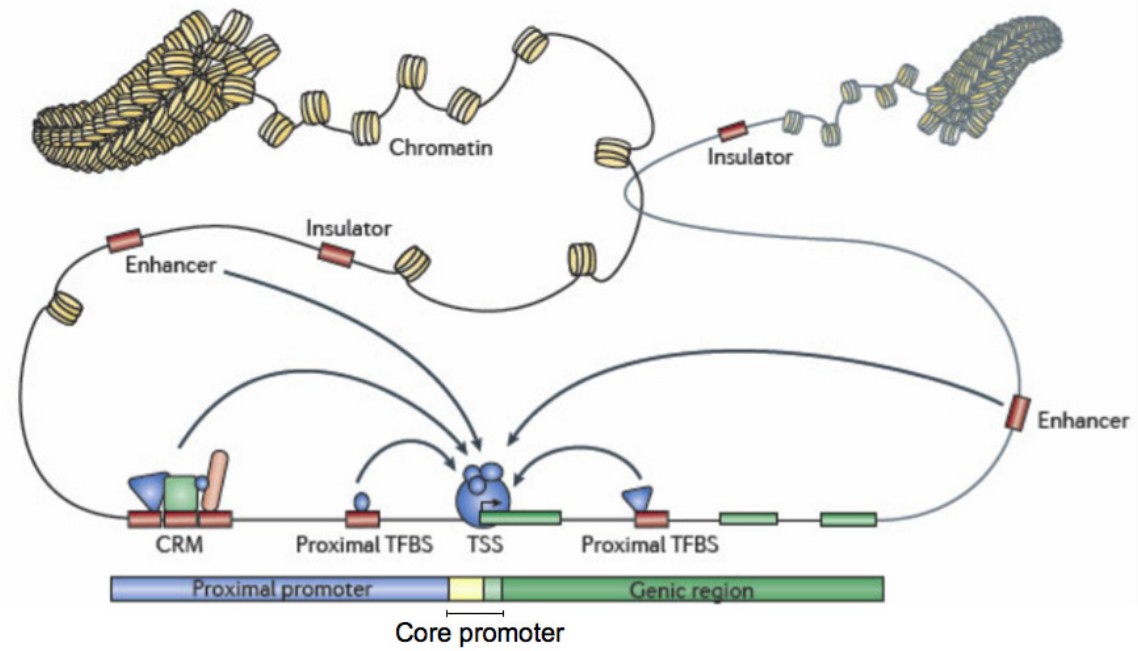


Figure 1-2 A summary of interactions between various promoter elements and regulatory signals.

Chromatin is the condensed form of DNA wrapped around histones to form nucleosomes. When decondensed, the DNA becomes accessible to binding factors. The topology of the region facilitates interactions between these local and distal binding factors to either initiate or inhibit transcription. CRM refers to cis-regulatory module, TFBS to transcription factor binding site and TSS to transcription start site. Adapted from Lenhard et al. (2012).

1.1.2.1 Enhancers

Enhancers are *cis*-regulatory sequences, which are only able to activate or upregulate transcription when in combination with a promoter. Like promoters, enhancers are sites where various transcription factors and chromatin remodellers bind (Andersson 2015), but unlike CPEs which have positional

constraints relative to the TSS, they are often positioned tens of kilobases away from the genes they regulate in higher organisms and have even been found distributed more than a megabase away (Lettice et al. 2003, Nobrega et al. 2003, Qin et al. 2004, Symmons et al. 2014).

A genome wide study of the interactions between the TSS and distal elements found that less than 10% of interactions were with distal elements associated with the closest gene (Sanyal et al. 2012), demonstrating that proximity of distal regulatory element to a gene is not always predictive of interactions. Promoter-enhancer interactions facilitated by looping of the chromatin (Tolhuis et al. 2002), physically bringing the two distal regions together in space thus enabling long-distance interactions, have since been mechanistically explored (Dekker 2008) with the aid of chromatin conformation capture (3C) technology (Dekker et al. 2002), and further exploited through a process of induced or 'forced chromatin looping' (Deng et al. 2012). This process has been used as a means to control gene expression with the potential to one day be used for therapeutic purposes through a 'rewiring' of the genome (Deng et al. 2014).

1.1.2.2 Silencers

Silencers on the other hand, function in direct contrast to enhancers. First observed more than 3 decades ago in yeast (Brand et al. 1985), many sub-types have since been identified and broadly grouped together as two distinct functional classes. Silencer elements are regulatory elements within DNA that actively repress transcription in an orientation- and position-independent manner by

driving a repressing mechanism through binding to transcription factors, of which most commonly interfere with PIC formation at the core promoter. Another smaller functional group of silencers, termed neuron restrictive silencers (NRE), disrupt transcription passively by binding repressor proteins that commonly result in interference with upstream elements (reviewed in Ogbourne et al. (1998))

1.1.2.3 Insulators

Boundaries between nearby chromatin domains, known as insulators (Sun et al. 1999), like silencers also function to block interactions between distal enhancers and promoters (Geyer et al. 1992, Kellum et al. 1992). Insulators are able to inhibit distant enhancer effects on core promoter regions or block the spread of the non-transcribed heterochromatin (Burgess-Beusse et al. 2002).

1.1.2.4 Proximal promoter

The proximal promoter regions surround the core promoter and consist of single and/or clustered transcription factor binding sites (TFBS). Such clusters of TFBSs are known as *cis*-regulatory modules (CRM) and are believed to act as tethering sites for distant enhancers facilitating interaction with the core promoter (Riethoven 2010, Lenhard et al. 2012).

1.1.3 Promoter regulation in the third dimension

Thus far, regulation of transcription has been discussed in a strictly linear manner in regard to DNA sequence and recruitment of transcription factors. One must consider other principal factors determining the fate of cells through regulation of gene expression. Chromatin architecture dictates accessibility of DNA to regulatory factors and epigenetic modifications, in addition to post-translational histone modifications and methylation of the DNA strand - potentially heritable modifications accumulated throughout an organism's lifespan. When considered in a three-dimensional space, this demonstrates how genome organisation represents an even more complex level of transcriptional regulation.

1.1.3.1 3D genome organisation

Our understanding of the regulatory landscape has shifted dramatically in the last two decades due in part to our ever increasing insight into the 3-dimensional (3D) organisation of the genome at the DNA level via 3C technology and its derivatives: 4C, 5C and Hi-C (Dekker et al. 2002, Dostie et al. 2006, Zhao et al. 2006, Lieberman-Aiden et al. 2009). In combination with molecular methods like fluorescence in situ hybridization (FISH) (Levsky et al. 2003) and super-resolution microscopy (Sigal et al. 2018) these methods give an insight in to how chromosomes are folded and the interactions that occur within the three-dimensional space of the nucleus.

Generation of maps from Hi-C data revealed that the nucleus is separated into two distinct A and B compartments (**Figure 1-3** B). One compartment is enriched for transcriptionally active gene rich open chromatin (A) and the other largely consists of inactive heterochromatic gene deserts (B), with interactions occurring preferentially within compartments as opposed to between them (Lieberman-Aiden et al. 2009).

Transcription within these compartments is influenced by spatial proximity of enhancers to their target promoters. One way in which interaction between these elements can occur in vertebrate genomes, where *cis-regulatory* elements are often separated by long distances, is through looping of the chromatin (**Figure 1-3** C), allowing contact between distant elements in the three-dimensional nuclear space (Bonev et al. 2016).

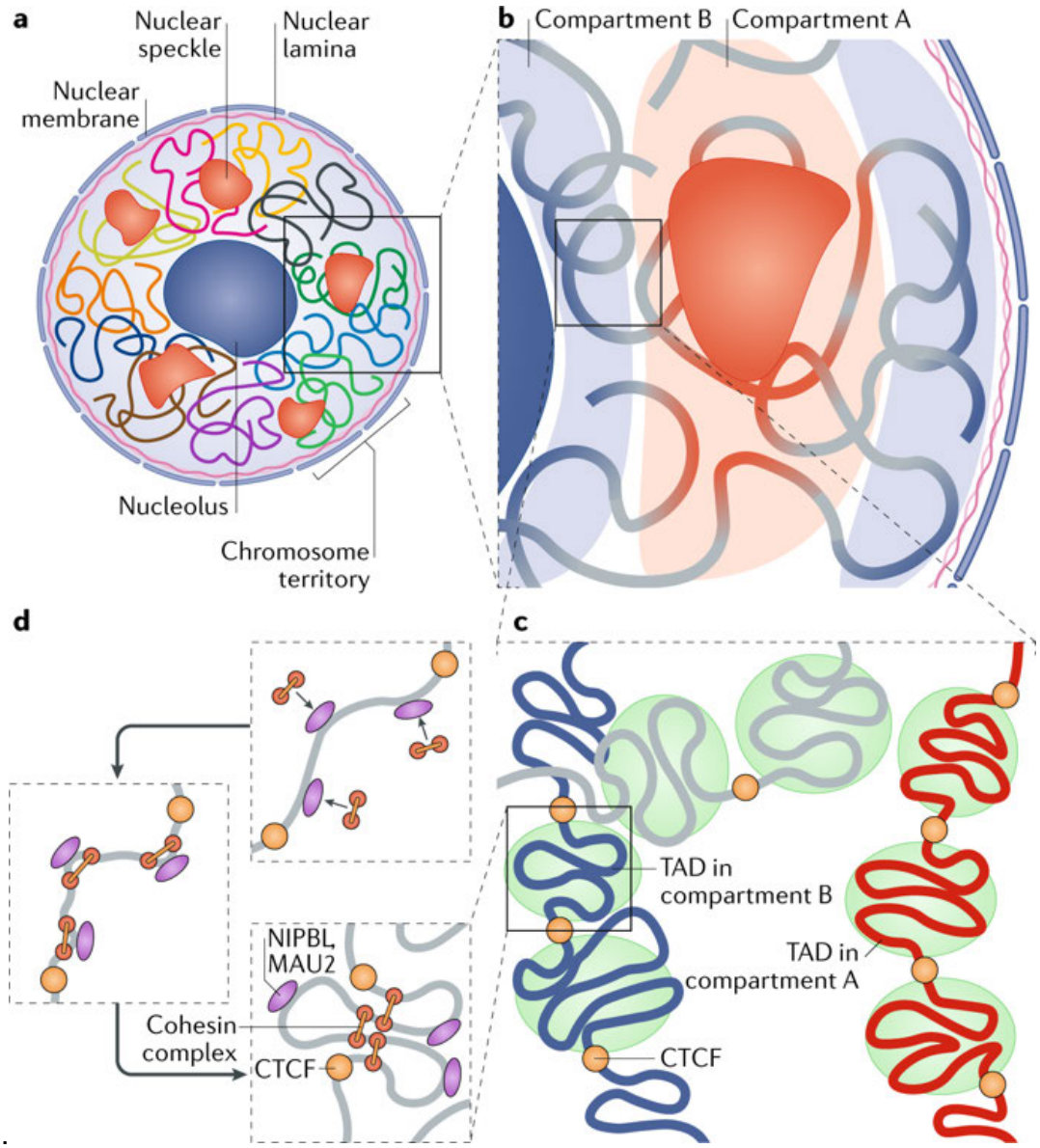


Figure 1-3 Hierarchical organisation of the genome in the nucleus

(A) *Graphical representation of the interphase nucleus with different coloured chromosomes occupying separate territories.* ***(B)*** *the nucleus is organised into two distinct compartments. One gene-rich transcriptionally active 'A' compartment enriched in euchromatin which is preferentially associated with nuclear speckles. A second gene-poor inactive 'B' compartment which is preferentially heterochromatin interacts mainly with the nuclear periphery.* ***(C)*** *Genomic domains with strong self-interaction are arranged into topologically associated domains (TADs) that are insulated from one*

another by CTCF binding at TAD boundaries. ***(D)*** *Within TADs looping of chromatin is mediated by cohesion complex that is loaded onto chromatin by complex containing* nipped-B-like protein (NIPBL) and MAU2. Extrusion of chromatin loops *function to bring distant enhancer and promoter elements into close proximity to enable interactions regulating expression of target genes. Repurposed from (Zheng et al. 2019)*

Such interacting domains are usually demarcated by both insulating CCCTC-binding factor (CTCF) and cohesion chromatin architectural proteins to prevent communications between non-interacting promoters and enhancers (**Figure 1-3** C). Through analysis on genome-wide contact maps it was found that specific regions of chromatin self-interact more frequently (reviewed in Pombo et al. (2015)). This effective isolation of whole sections of linear DNA, that although adjacent in the first dimension are sequestered into distinct topological compartments in the 3D genome, influences expression of genes in both time and space and will likely become a defining factor in the characterisation of 'long-range' regulatory interactions, for example between promoter and enhancer elements, with particular significance during gametogenesis, embryonic development, lineage commitment and stem cell differentiation (reviewed in Zheng et al. (2019)). Although it is clear that our understanding of the regulation of transcription in the 3D genome is still not complete, in a recent study Williamson et al. (2019) demonstrated experimentally that despite perturbation of the *shh* locus, resulting in disruption of the 3D architecture while maintaining integrating of the cis-regulatory elements, spatio-temporal gene regulation during development was suitably robust, raising the question as to whether TADs are a

required intermediator of gene expression through developmentally active enhancers (see review by Beagan et al. (2020)).

1.1.3.2 Chromatin architecture and epigenetic modifications

Chromatin is arranged in a series of nucleosomes, structures consisting of a DNA helix wound around densely packed octamers of core histone proteins H2A, H2B, H3 and H4 as well as linker proteins H1. Dense compaction of this fibre is neccessary for the genomic material to be accomodated into the limited space of the cellular nucleus where it is positioned into distinct chromosomal territories (reviewed in (Kadonaga 2002), Woodcock (2006), Cremer et al. (2014) Bonev et al. (2016)) with gene rich chromosomes preferentially located close to the nuclear core and gene-poor chromosmes situated closer to the periphery (**Figure 1-4** B).

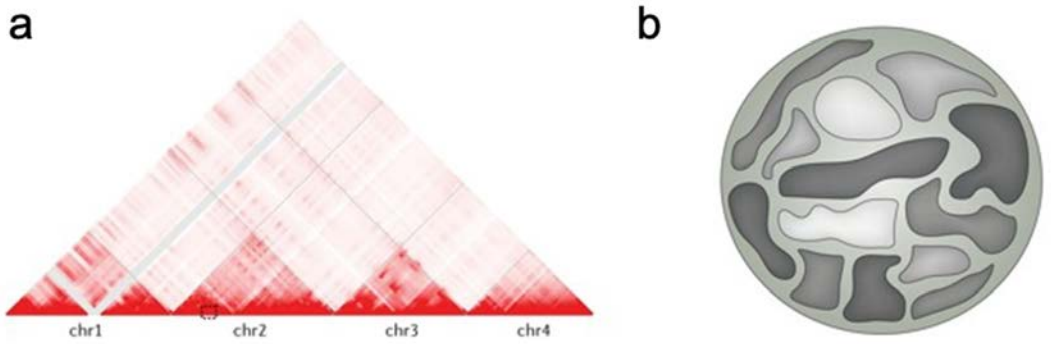


Figure 1-4 Chromosomes represent the highest level of chromatin organisation in the nucleus

(A) *Topologically associating domains (TADs) occur primarily as intra-chromosomal interactions, shown from Hi-C data. (**B**) Chromosomes occupy distinct territories within the nucleus with gene rich chromosomes preferentially located closer to the nuclear core and gene-poor chromosomes localised more toward the nuclear membrane. Figure modified from (Bonev et al. 2016).*

Compaction of chromatin affects accessibility of *cis-regulatory* elements to transcription machinery and relevant transcription factors. Several methods have been developed to inform on accessibility of chromatin genome wide, based on site specific cleavage of unbound DNA by endonucleases, followed by ligation-mediated PCR.

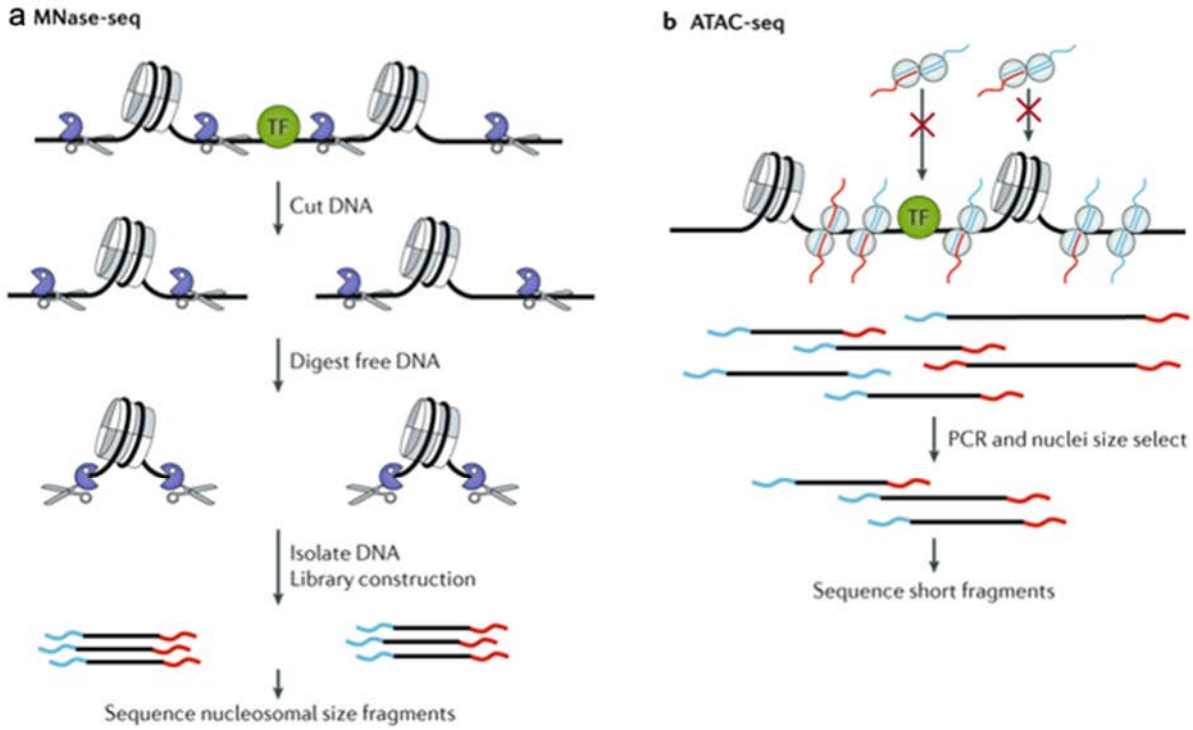


Figure 1-5 two principal methods for analysing chromatin accessibility

(A) *Micrococcal nuclease sequencing (MNase-seq) utilises endo-/exonuclease activity of MNase to cleave and digest accessible DNA.* ***(B)*** *Assay for transposase-accessible chromatin using sequencing (ATAC-seq) utilises hyperactive Tn5 transposase to cleave accessible DNA while simultaneously ligating sequencing adapters.*

Micrococcal nuclease digestion with sequencing (MNase-seq) involves cleavage of formaldehyde crosslinked accessible chromatin via activity of a non-specific micrococcal nuclease. The intact DNA, protected from digestion by bound histones or other DNA binding factors (such as TFs), is purified from the proteins and deep-sequenced with increased accessibility inferred from decreased presence in the sequencing library (Henikoff et al. 2011, Cui et al. 2012) (**Figure 1-5** a).

A contrasting method, assay for transposase-accessible chromatin using sequencing (ATAC-seq) utilises a hyperactive mutant Tn5 transposase that cleaves at specific recognition sites in open chromatin, while simultaneously ligating sequencing adapters, a process known as tagmentation. Following transposition tagged fragments can then be purified, PCR amplified and sequenced using next-generation sequencing (NGS) technologies (Buenrostro et al. 2013) (**Figure 1-5** b). In contrast to MNase-seq, enrichment in the sequencing library infers an increase in chromatin accessibility. These techniques provide an insight into nucleosome occupancy and position as well as transcription factor binding, which is fundamental in regard to the study of gene regulation (Klein et al. 2020).

The reader is directed to a review by (Klemm et al. 2019) for a comparison between four primary techniques for probing open chromatin: ATAC-seq, MNase-seq and two alternative methods not described in this thesis: Nucleosome occupancy and methylome sequencing (NOMe-seq) and DNase I hypersensitive site sequencing (DNase-seq), and also for an overview of the key mechanisms mediating remodelling of chromatin in the eukaryotic cell. One way in which accessibility of *cis-regulatory* elements to transcription machinery and relevant transcription factors (**Figure 1-6**) is modulated is through epigenetic modification of chromatin.

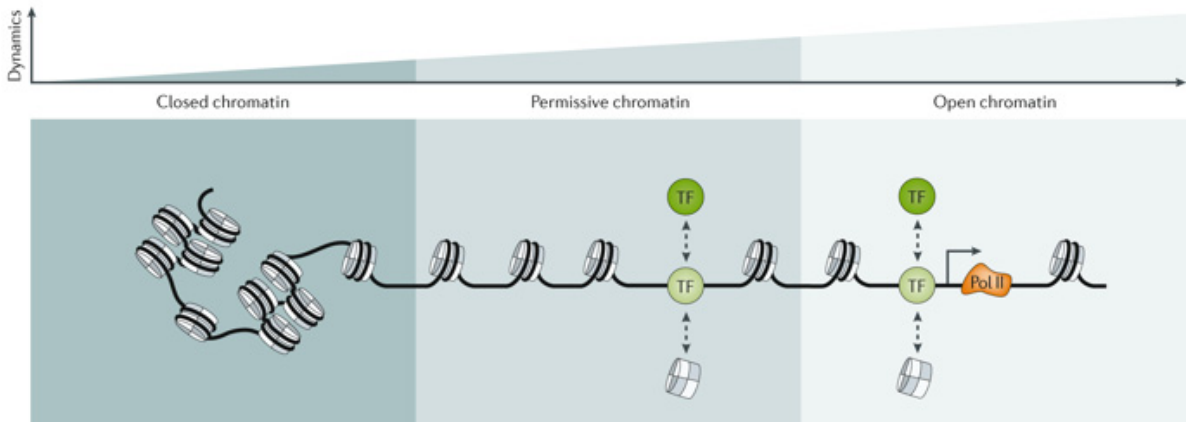


Figure 1-6 Representation of distribution of chromatin states and dynamics across the genome

Closed chromatin is not accessible to trans-acting factors, while permissible chromatin is amenable to remodelling into open chromatin regions competent in establishing active transcription from gene promoters through binding of RNA Pol II and transcription factors (TF). Figure repurposed from Klemm et al. (2019).

Epigenetic modifications refer to changes within chromosomes that function to modulate expression of genes without changing the underlying genetic code. Such modifications exist in two forms: DNA methylation and post-translational histone modification. Both will be discussed here in brief, in relation to regulation of transcription and how such modifications can be detected by Chromatin Immunoprecipitation analysis (ChIP).

Observation of DNA-protein interactions has allowed insight into how histone modifications and transcription factors are able to modulate gene expression. In its simplest form, ChIP techniques involve fragmentation of crosslinked DNA bound to DNA. An antibody specific for the factor of interest is used to pulldown protein- or nucleosome-bound DNA, crosslinking is then reversed through

heating and subsequent DNA analysed and sequenced via next generation sequencing technologies in ChIP-seq protocols (Solomon et al. 1988, Park 2009).

The core proteins that constitute the nucleosome are subject to more than 8 distinct types of enzymatically catalysed post-translational modifications, primarily at the unstructured N-terminus of the histone tails, and includes, but is not limited to, acetylation, methylation, ubiquitination and phosphorylation (Bernstein et al. 2007, Kouzarides 2007).

Unlike methylation, which can have variable effects depending which residue is modified, acetylation of lysine is almost always synonymous with transcriptional activity and accessible (eu)chromatin via neutralisation of the basic charge which is believed to weaken DNA-histone interactions (Hong et al. 1993), with enrichment at *cis*-regulatory elements of active gene promoters common (Litt et al. 2001). Histone modifications on euchromatin can vary widely, but modifications common to the sites of transcription, or those that modulate it, for instance H3K4 and K36 methylation, are associated with transcribed chromatin (Liang et al. 2004, Morris et al. 2007). Specifically, H3K4me mono methylation enrichment is predictive of active enhancers, whereas tri-methylation is suitably enriched at promoters (Heintzman et al. 2007). Conversely, mono-,di- and tri-methylated H3K36 is found all across the transcribed region (Bannister et al. 2005). Contrariwise to the effects of other methylated lysine residues described, methylation of the histone tail at lysines 9, 27 and 20 are all generally associated with transcriptional repression (Bernstein et al. 2007).

Although a process not conserved in all eukaryotes (see review Suzuki et al. (2008)), direct methylation of DNA functions to regulate transcription, while occurring almost exclusively at CpG dinucleotides in vertebrates, established via addition of a methyl group the 5th cytosine on the DNA backbone (5mC) by DNA methyl-transferases (DNMTs), with the majority of CpGs in the genome existing in this methylated state (Bird 2002, Goll et al. 2005, Suzuki et al. 2008). Short DNA sequences across the genome rich in both GC content and predominately unmethylated CpG dinucleotides are termed Vertebrate CpG islands (CGIs). These regions are associated with the majority of mammalian core promoters, thus indicating highly probable sites of transcription initiation (Gardiner-Garden et al. 1987, Carninci et al. 2006, Deaton et al. 2011).

1.2 Identification of promoter architecture

1.2.1 Introduction to computational analysis of transcription codes and dynamics

Development of technologies engineered for precise identification of transcription start sites at single nucleotide resolution, as well as precise nucleosome positioning on a genome wide scale contributed to better understanding of the core promoter structure and its various sub-types. Classification of distinct promoter classes was achieved largely through computational analysis of 5' ends of capped RNAs - used to locate TSSs (Carninci et al. 2006). In order to accurately assess transcriptional dynamics during development it is necessary to have an ample depth of read coverage (Carninci et al. 2006, Lenhard et al. 2012, Haberle et al. 2016).

1.2.2 Methods for precise determination of transcription start sites

There are many approaches to study 5' ends of transcripts. One such technique: Cap Analysis of Gene Expression or CAGE has been used to study promoter architecture through genome-wide high depth of coverage identification of start sites of transcription (Shiraki et al. 2003). CAGE is a technique that can reveal promoter TSSs and their utilisation dynamics at single nucleotide resolution through quantification of 5' ends of capped eukaryotic RNA (**Figure 1-7**),

providing spatio-temporal and promoter type characteristic profile (Carninci et al. 2006, Takahashi et al. 2012). In brief, this method involves selection of capped RNAs by specifically attaching a biotin tag to the 7-methylguanylate cap, sequencing their 5' ends and mapping the resulting sequence (CAGE tag) to the genome, thus identifying the exact nucleotide and position from which the transcription of these RNAs was initiated (**Figure 1-7**) (Takahashi et al. 2012). In recent years, several iterations of the CAGE protocol have been developed, such as no-amplification no-tagging CAGE (nAnT-iCAGE), a modified methodology developed to reduce potential bias from PCR amplification and restriction enzyme cleavage (Murata et al. 2014). Despite this improvement, one of the main disadvantages of CAGE remained, that it requires large amounts of material, making it inappropriate for low-input applications such as to assess promoter dynamics at the level of single cell populations. Recently two novel methods: NanoCAGE (Salimullah et al. 2011, Poulain et al. 2017) and Super-Low Input Carrier CAGE (SLIC-CAGE) - an improved method with less bias that utilises a selectively degradable carrier RNA were designed to rectify this issue with both methods purporting an increase in sensitivity to an extent that 1000-fold less material is required than the standard CAGE protocol, making it more amenable for low-input applications (Cvetesic et al. 2018).

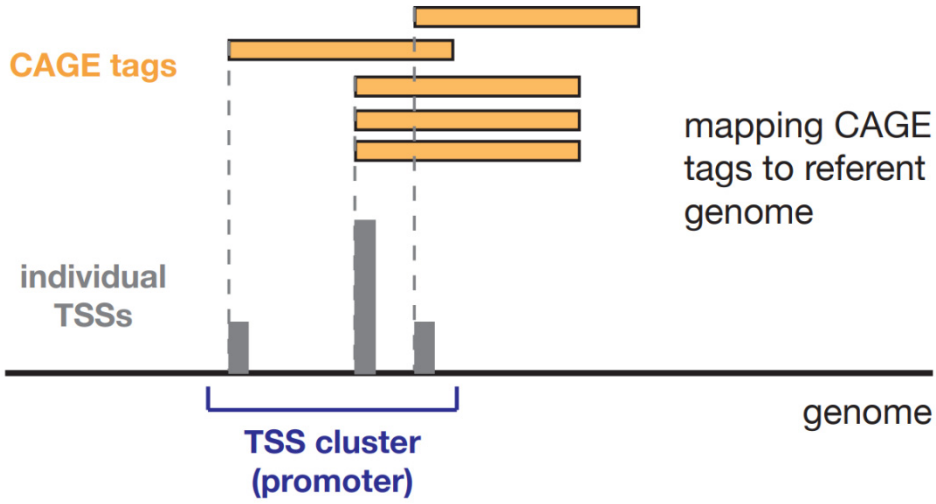


Figure 1-7 Mapping CAGE tags to a reference genome

Schematic demonstrating how computational mapping of 5' ends of CAGE tags to a reference genome allows high resolution determination of start sites of transcription (TSS cluster). The 5' position of the CAGE tags informs of the first base of the resulting transcribed RNA molecule with the number of tags informing on the quantity of RNA produced, thus serves as readout for steady state RNA levels. (Haberle et al. 2016).

Another variation of 5' end analysis is CapSeq, which differs to CAGE by purification method, designed to capture longer RNAs transcribed by Pol II while reducing background signal from detection of structural RNAs (Gu et al. 2012). Additional methods include Rapid amplification of cDNA ends (RACE) coupled with a next generation sequencing (NGS) technology – NGS-RACE (Leenen et al. 2016) and GRO-cap (Kruesi et al. 2013). NGS-RACE has a greater sequencing depth when compared to CAGE, but due to inability to quantitatively assess transcript abundance it functions more as a complement to CAGE (Leenen et al. 2016). Useful, for example for viewing a specific TSS region in more detail.

To get an insight into real-time indication of TSS usage, a technique to isolate newly transcribed or nascent RNA (GRO-seq) has been coupled with the 5' end capture (GRO-cap) (Core et al. 2008, Kruesi et al. 2013). In conventional CAGE, transcripts analysed are of total mature RNA, enriched for long-lived transcripts (Haberle et al. 2016) thus providing a measure of the transcript accumulation up until the point of isolation. GRO-cap therefore provides a novel opportunity for real-time indication of promoter dynamics.

1.2.3 Classification of promoter types

Classification of promoters through study of characteristic CAGE profiles, made possible due to the stereotypical shape of CAGE TSS (CTSS) arrays has led to the classification of 3 main core promoter classes (a.k.a architectures): adult tissue-specific, ubiquitously expressed and a developmentally regulated class involved in control of both embryo pattern formation and differentiation (Akalin et al. 2009) (shown in **Figure 1-8** with their specific class-defining elements). Although various other sub-types, such as a translational machinery specific 5′ terminal oligopyrimidine (5'TOP/TCT) promoters that encode ribosomal proteins and other translation-related proteins (Parry et al. 2010), have been defined (Carninci et al. 2006, Lenhard et al. 2012). Promoters are classified through computational analysis and divided into their respective classes based on having an either sharp (single base) or broad (multiple start sites) initiation sequence, by the arrangement of nucleosomes in either an ordered or disordered fashion and also on how they are regulated. For example, developmental promoters receive regulatory input from distal enhancers so have long-distance interactions

(Lenhard et al. 2012, Haberle et al. 2016, Haberle et al. 2018). This classification of promoter architectures is informative as it appears to distinguish biologically distinct gene categories and also appears to indicate differential transcription initiation dynamics in different ontogeny and tissue specific contexts (Nepal et al. 2020).

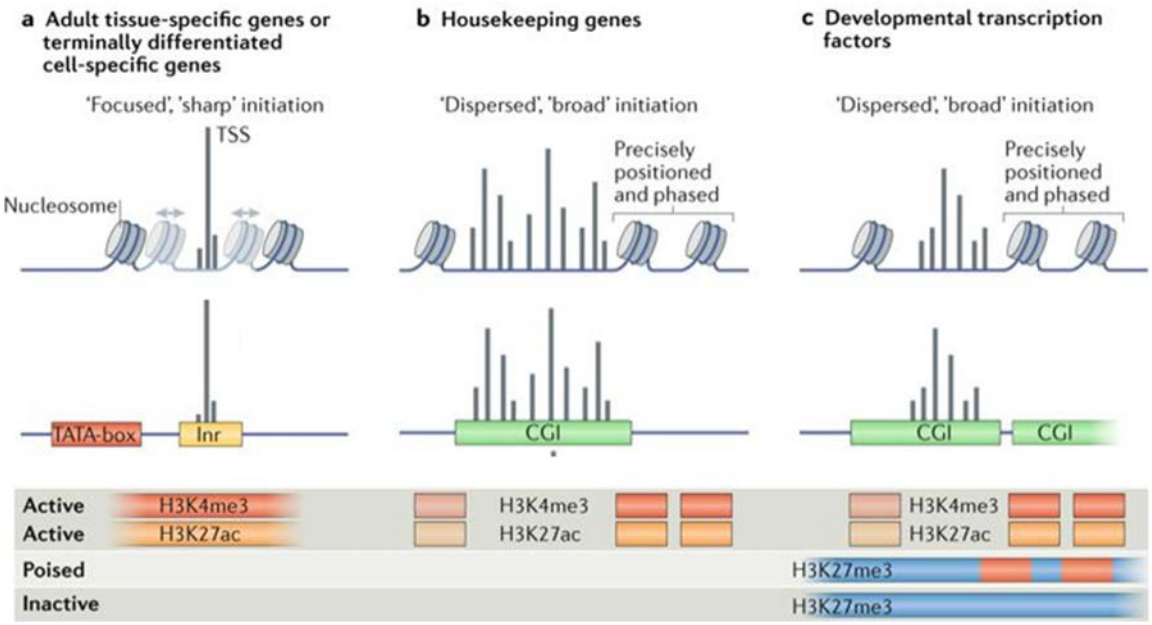


Figure 1-8 Three types of major core promoter defined in metazoan

Schematic illustrating the three major types of core promoter identified by their respective defining attributes: nucleosome periodicity, initiation pattern, motifs present and characteristic chromatin architecture. Adapted from Haberle et al. (2018).

1.2.3.1 Nucleotide composition in core promoters

Adult tissue specific core promoters possess indistinctly positioned nucleosomes (Rach et al. 2011), chromatin modifications concomitant with active transcription, sharp initiation patterns associated with canonical pyrimidine / purine dinucleotide

initiation and TATA-box motifs (**Figure 1-8** a). In contrast to the classical initiation pattern described, many core promoters have been found to possess a non-canonical polypyrimidine initiation motif termed 5'TOP (terminal oligo-polypyrimidine) in mammals (Perry 2005, Roepcke et al. 2006) and TCT initiator in Drosophila (Parry et al. 2010) with TCT initiation-dependant genes also found to be conserved in the zebrafish (Nepal et al. 2013). Recent work by Nepal et al. (2020) demonstrated intertwining of these polypyrimidine motifs with canonical motifs resulting in identification of a new class of promoter termed dual-initiation promoters, found to be divergently active during zygotic genome activation (ZGA), demonstrating the potential for close interaction between distinct promoter classes.

A feature, common to many vertebrate core promoters of housekeeping genes (**Figure 1-8** b) and developmental transcription factor specific core promoter classes (**Figure 1-8** c) is overlapping Vertebrate CpG islands (CGIs) (Yamashita et al. 2005, Carninci et al. 2006, Saxonov et al. 2006, Akalin et al. 2009). Regions enriched in CpG dinucleotide repeats are commonly associated with bidirectional transcription, defined as transcription initiating at the same TSS in both directions along the sense and anti-sense strands (Trinklein et al. 2004). Promoters can be further classified as either low- or high- CpG, a feature which is intrinsically linked with their underlying initiation patterns (Saxonov et al. 2006, Lenhard et al. 2012), despite this it has been shown that some promoters possess both a TATA-box and high CpG Island overlapping the core promoter and as a result are proficient in initiating transcription in a both a TATA-dependant or independent manner

(Ponjavic et al. 2006), thus dividing them into functional classes depending on initiation pattern, ‘sharp’ (focused) or ‘broad’ (dispersed), has been suggested as a more accurate means for classification (Rach et al. 2011).

1.2.3.2 Initiation patterns and the TSS

The majority of genes do not possess classical TATA-box motif promoter architecture (Carninci et al. 2006), so unlike the sharp transcription initiation seen in TATA-box promoters, transcription from high-CG classes is classified as broad with heterogeneous initiation sites (Reynolds et al. 1984) in regions lacking in nucleosomes (Rach et al. 2011). Nucleosomes positioned either side of these regions are precisely placed and marked by chromatin modifications indicative of active transcription with the exception of the bivalently marked genes of the developmental transcription factor (TF) class. In mouse embryonic stem (ES) cells Bernstein et al. (2006) found that the promoters of TF class genes remain poised, with promoters additionally marked by repressive chromatin modification H3K27me3. Bivalent marking is understood to be a means to silence developmental genes with subsequent poising ensuring activation only occurs in the appropriate cell lineages.

Unlike high-CG promoters, which in mammals and many vertebrates lack defined DNA motifs (Carninci et al. 2006), promoters of many low-CG genes, characteristic to adult tissue specific genes, are reliant on TATA-box motif for recruitment of the PIC which is both directed and guided by bound TFs. The

specificity in landing-site in combination with the fixed distance of the initiator sequence results in a lack of flexibility in start site of transcription (Ponjavic et al. 2006). Thus this promoter is commonly referred to as having a sharp initiation pattern, due to its single initiation site, positioned at the +1 position on the initiator (Inr) motif (Carninci et al. 2006, Lenhard et al. 2012).

Work from our lab (Haberle et al. 2014), relating to promoter architecture during activation of the zygotic genome, has shown that many maternal genes lack a canonical TATA-box and instead possess a W-box motif. W-box motifs are similar to the TATA box in that they are both A-T rich sequences, with equivalent positioning upstream of the initiator, yet W-boxes differ in their information content, being shorter in length than the canonical TATA box. Zygotic genes on the other hand, rely neither on a W- or TATA box for transcriptional activation, thus unlike maternal transcription, have broad initiation patterns precisely aligned with the first downstream nucleosome. Interestingly, these two promoter grammars have been found intertwined on the same core promoter of genes which are expressed in both oocytes and zygotically in the developing embryo, highlighting how two independent initiation mechanisms can be interwoven at the same promoter origin, with such promoters dubbed ‘shifting promoters’. This difference in initiation was seen regardless of whether the promoters were found to be shifting promoters or whether they were exclusively either maternally or zygotically expressed, demonstrating that a different mechanism governs TSS selection between the oocyte and embryo.

1.3 Zebrafish as a model organism

The freshwater fish *Danio rerio*, or simply zebrafish, named as such due to its distinctive horizontal stripes, is a member of the Cyprinadae family and native to East Asia. Like worms and fruit fly, the zebrafish embryo is transparent, in addition to being relatively large in comparison to the adult fish, externally developing and inexpensive to maintain, making it a desirable model for early developmental research, particularly for studies involving imaging.

1.3.1 Developmental stages of the zebrafish

Upon fertilisation, the extra embryonic chorion separates from the zygote and cytoplasm moves to the animal pole to form the blastodisc. After fertilisation and prior to the Midblastula Transition (MBT) cells cleave at a rate of once every 15 minutes, with such cleavage being meroblastic i.e., occurring at the animal pole of the embryo.

The initial cleavages occur synchronously i.e., cells divide at the same time. It takes 2 and ¾ hours to produce 512 cells and form a blastula, with the first cleavages occurring synchronously. At this stage MBT begins, characterised by activation of the zygotic genome, loss of cell synchrony, lengthening of the cell cycle and emergence of cell motility (Kane et al. 1993). Epiboly follows, with the migration of cells around the yolk sac marking the start of gastrulation: the specification of the three primary germ layers: mesoderm, endoderm and ectoderm which are the precursors to all embryonic tissues. After 6 hpf involution

occurs, by which the cells of the future dorsal side begin to migrate underneath the surrounding cells, forming the embryonic shield. Bud stage at 10 hpf marks the end of gastrulation and the start of segmentation and the formation of the somites: mesodermal sections of tissue that will eventually give rise to the spleen, muscle and blood.

The final developmental stages prior to hatching involve subdivision of the ectodermal neural plate into a regionalised neural tube, a process known as neurulation. At the end of the first day of development segmentation is complete by which time most of the primordium are specified. After which, the embryo enters the pharyngula period and proceeds to hatch at 2 dpf, an overview of the key stages mentioned are summarised in **Figure 1-9**.

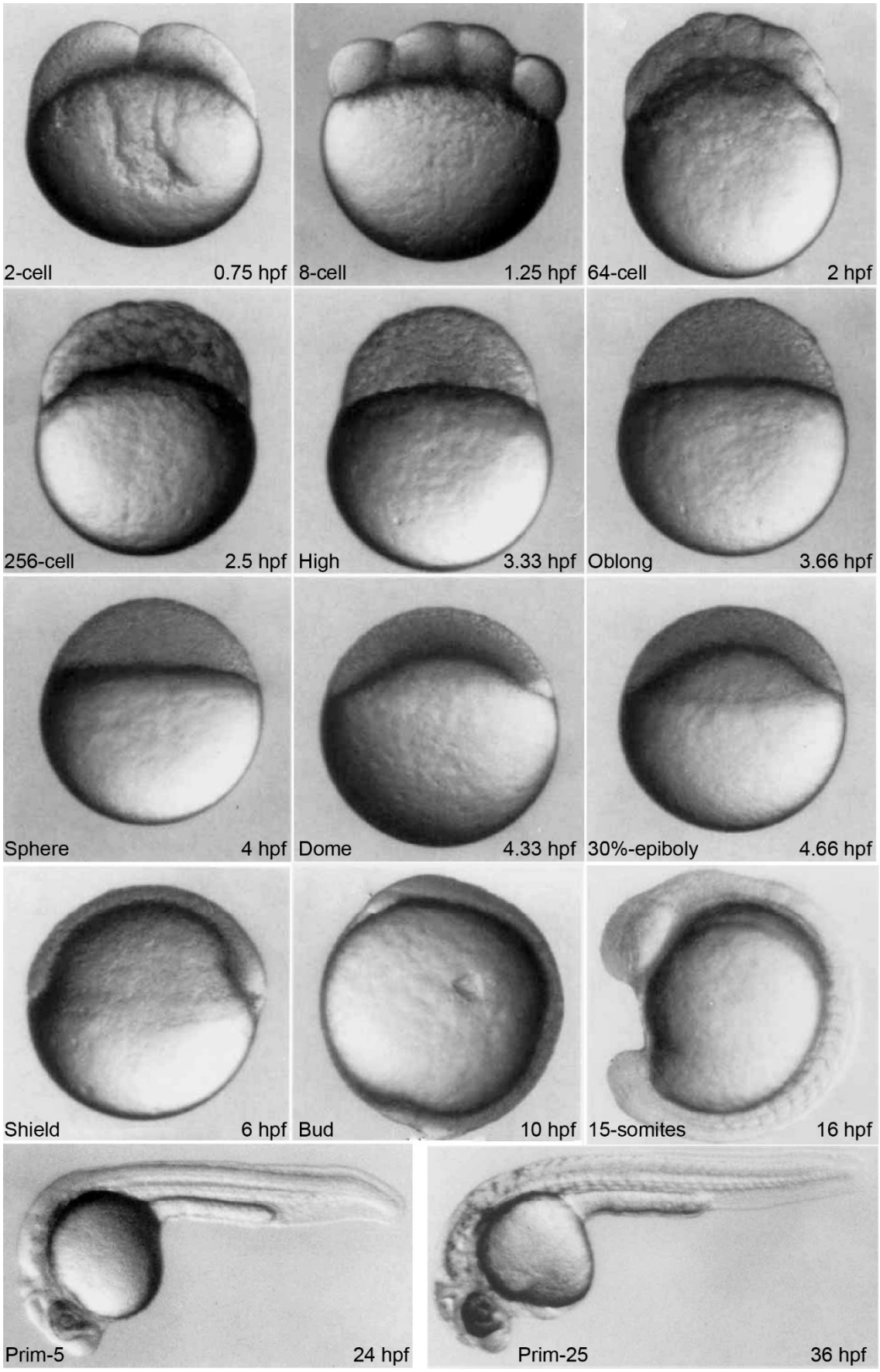

2-cell 0.75 hpf
8-cell 1.25 hpf
64-cell 2 hpf
256-cell 2.5 hpf
High 3.33 hpf
Oblong 3.66 hpf
Sphere 4 hpf
Dome 4.33 hpf
30%-epiboly 4.66 hpf
Shield 6 hpf
Bud 10 hpf
15-somites 16 hpf
Prim-5 24 hpf
Prim-25 36 hpf

Figure 1-9 A visual representation of Zebrafish developmental stages.

Visual representation of the developmental stages of the zebrafish from the initial cleavage at 0.75 hpf up to prim-25 at 36 hpf imaged using a light microscope. For each stage, the accompanying developmental stage name is included on the bottom left-hand side of the image and the time in hours post fertilisation (hpf) is included on the bottom right-hand side. Modified from Kimmel et al. (1995).

1.4 Mechanism for transcription initiation in early zebrafish embryos

The initially transcriptionally quiescent embryo relies on deposited maternal factors to drive the first cellular cleavages, with the transition from transcriptionally inactive to global genome activation providing a good model for study of dynamics of transcription initiation *in vivo* (reviewed in (Lee et al. 2014, Vastenhouw et al. 2019)).

1.4.1 The Midblastula Transition

Studies into the mechanisms behind the switch from maternal to zygotic control over embryo development, span over five decades. Associated with activities such as: lengthening of the cell cycle, asynchronous division, cell motility and the start of transcription (Signoret et al. 1971, Hara 1977, Satoh 1977, Gerhart 1980, Newport et al. 1982, Edgar et al. 1986, Kane et al. 1993) the mid-blastula transition (MBT) is the developmental stage at which the zygotic genome becomes active and nuclear control of transcription begins in both fish and frog.

Occurring only after a specified number of cellular divisions: 10 in *D. rerio (Kane et al. 1993)*, 12 in *X. laevis (Newport et al. 1982)* and 14 in *D. melanogaster* (Foe et al. 1983), MBT occurs during the blastula stage of embryonic development and correlates with an increasing cell cycle length and loss of synchrony. It is commonly misused interchangeably with activation of the zygotic genome (ZGA) because it frequently coincides with onset of zygotic transcription in externally

developing embryos, despite the fact that in mammals onset of ZGA occurs even before the first cellular cleavage (O'Farrell et al. 2004, Tadros et al. 2009).

ZGA is a process regulated by a balance between repressive factors such as histones and replication factors with transcription factors that compete for DNA binding. Successive cellular divisions without cell growth leads to increase in the nucleo-cytoplasmic (N/C) ratio resulting in dilution of repressive factors, such as histones, (**Figure 1-10**) allowing zygotic transcription to occur. Although it is known that a small population of genes are able to circumvent this global block and be transcribed before the global onset of transcription that occurs at the MBT in many organisms, although the mechanism by which this occurs is not well understood.

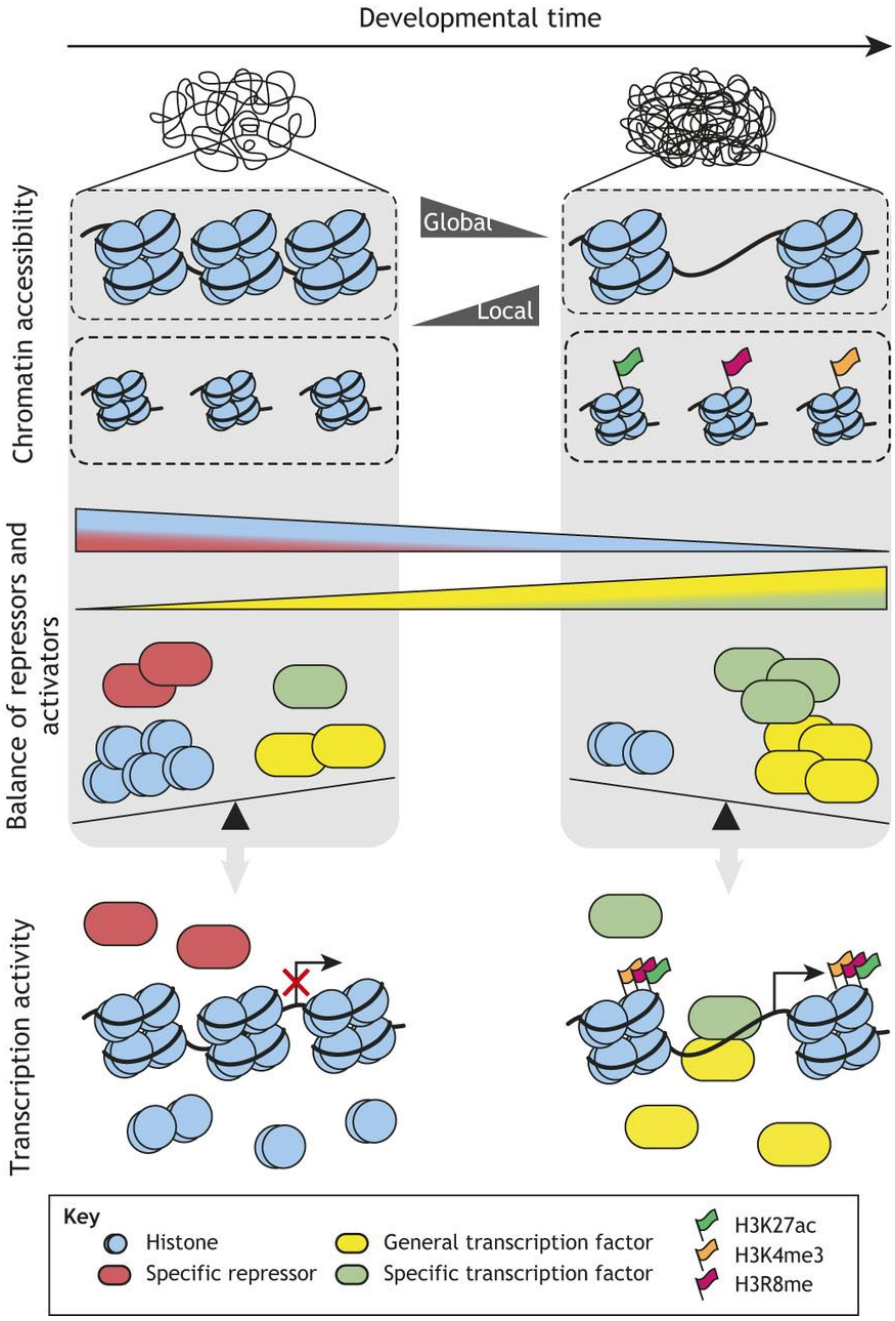


Figure 1-10 Factors affecting activation of the zygotic genome

Global compaction of chromatin with increased local accessibility, dilution of repressive maternally deposited factors and repressive histones when combined with an increase in both specific and non-specific transcription factors results in an environment more amenable to transcriptional activation. Repurposed from Vastenhouw et al. (2019).

1.4.2 Nucleo-cytoplasmic (N/C) ratio

Initial observations by Newport and Kirschner (Newport et al. 1982) established the first evidence for the theory that maternally derived cytoplasmic factor(s) present in the early embryo, such as maternally deposited histones, were actively functioning to inhibit transcription. In addition, it has been shown that nucleosomes are able to block access to transcriptional start sites by the replication machinery (Lorch et al. 1987). Short cell cycles resulting in DNA replication and cellular cleavage in the absence of growth, likely result in an enviroment where the nuclear content of the cell becomes more abundant in relation to cytoplasm, eventually creating a state amenable to transcription (Newport et al. 1982). The theory that these repressive factors are being gradually depleted over time is the core component of what is known as the nucleocytoplasmic ratio model. A recent study in *Xenopus* embryos by Chen et al. (2019) gives further support for this model with the finding that the onset of ZGA strongly correlates with cell size but not with elapsed time or number of cell divisions and is spatially non-uniform.

In addition to maternally deposited transcriptional repressors, it is believed an overabundance of maternally deposited histones present in the early embryo (Adamson et al. 1974, Anderson et al. 1980), are able to bind with high affinity to DNA despite low sequence specificity (Campos et al. 2009), resulting in competition with regulatory proteins for DNA-binding (Amodeo et al. 2015, Joseph et al. 2017, Shindo et al. 2019). In accordance with the nuclear cytoplasmic model, injection of histone cocktails have been shown to delay onset

of ZGA in both frog (Almouzni et al. 1995, Amodeo et al. 2015) and fish (Joseph et al. 2017). Interestingly, most histone H3 are free within the nucleus (unbound to DNA) and are titrated away with successive cell cycles, a reduction that corresponds to a global decrease in H3 chromatin occupancy and subsequent compensation with an increase in modified histone H3.3, which the authors suggest may contribute to slowing of the cell cycle (Shindo et al. 2019). Many of the factors known to affect the N/C ratio are represented graphically in **Figure 1-10**.

1.4.3 Role of chromatin competency in transcription initiation

Competency of the early embryo to perform transcription pre-MBT was demonstrated by Newport et al. (1982) through injection of a plasmid containing a tRNA gene that was transcribed before global onset of ZGA , findings that were later substantiated in both mouse (Wiekowski et al. 1993) and zebrafish (Harvey et al. 2013). Although they observed that transcription was only transient, thus they hypothesised that DNA was being held in a repressed state due to intergration into chromatin.

Assembly of chromatin is a major mechanim by which basal transcription is inhibited in pre-MBT embryos (Prioleau et al. 1994). Through placement of nucleosomes the basal transcription factors, such as TBP, are blocked from associating with the core promoter and other *cis-regulatory* elements. Consequently if chromatin assembly is disrupted or basal transcription factors are

preloaded onto DNA templates, such as TBP onto an exogenous plasmid (Prioleau et al. 1994), transcriptional silencing is temporarily circumvented. Addition of non-specific DNA was found to prevent transcriptional silencing of preloaded plasmids via depletion of maternally deposited histones (Almouzni et al. 1995). Synthesis of core histones during oogenesis (Adamson et al. 1974) are sufficient to build tens of thousands of nuclei, as was shown in embryos of *X. laevis* (Woodland et al. 1977) and sustain assembly of chromatin past the MBT. Structure of chromatin was found to also repress transcription when recapitulated *in vitro* (Almouzni et al. 1991, Amodeo et al. 2015). These results collectively demonstrate the role of the nucleosome in establishing an environment inhibitory to transcription pre-MBT through occlusion of the TSS and associated binding sites and thus how chromatin assembly functions to suppress basal transcription before the MBT.

Through study of earliest genes to be transcribed in zebrafish, such as *mir-430,* Heyn et al. (2014) found that the mitochondrial genome is transcriptionally active from fertilisation onwards, thus its regulation is independent from the nuclear genome. This poses an interesting point because mitochondrial DNA (mtDNA) is not packaged into chromatin, instead it is packaged into histone free nucleoids of which the composition of binding proteins is still not clear (Bogenhagen 2012). Recently Chan and colleagues (Chan et al. 2019) reported that the enhancer regulator proteins P300 and Brd4 are sufficient to trigger ZGA by regulating chromatin acetylation in zebrafish embryos. With overexpression of these factors resulting in premature genome activation and conversely, inhibition of histone

acetylation blocking genome activation, this strongly suggests that closed chromatin plays a key role in the global transcription block of the nuclear genome in the early embryo.

Thus it is believed that a specific group of transcription factors known as pioneering factors are required for making chromatin accessible to the transcriptional machinery driving subsequent activation of transcription (Zaret et al. 2011).

1.4.4 Pioneer and stem cell factors role in transcription initiation in early zebrafish embryos

Pluripotency can broadly be defined as the ability of cells to develop into any of the three primary germ layers. All cells in a developing embryo pass through this period of undefined cellular identity, where position in the embryo dictates the specific developmental programmes it will acquire. Transcription factors that promote or maintain this indeterminate state are purported as either ‘stem cell’ or ‘pluripotency’ factors. Co-localisation of pluripotency factors onto embryonic stem (ES) cell chromatin is known to regulate transcription, remodel chromatin and direct cell identity (Boyer et al. 2005, Loh et al. 2006, Orkin et al. 2011). The possibility for reversal of this process of differentiation was demonstrated by Takahashi et al. (2006), who by exposing both mouse and human somatic cells to Sox2 and Oct4, in addition to stem cell factors Klf4 and c-Myc (now known collectively as the ‘Yamanaka factors’), were able to reprogramme these cells to a more primitive and undefined state, known as induced pluripotent stem cells

(iPSC). Together these observations demonstrate the influence these factors have on the transcriptional programme and their ability to direct transcriptional architecture, cellular fate and identity.

Able to bind closed chromatin, pioneer factors are the first factors to bind and function to locally facilitate both opening of the chromatin and recruitment of other DNA binding factors and transcriptional machinery to the core promoter. FoxA and GATA family of TFs are example of pioneer factors able to access their binding sites even in condensed chromatin, effectively functioning to increase accessibility by localised remodelling, achieved through facilitating localised nucleosome repositioning (Zaret et al. 2011, Chen et al. 2014).

A well-characterised example of a pioneer factor is *Drosophila* zinc finger protein Zelda (Zld). Both a transcription factor (Harrison et al. 2011) and chromatin remodeller (Foo et al. 2014, Li et al. 2014, Sun et al. 2015), ZLD was shown to be essential for opening hundreds of chromatin domains associated with the earliest genes to be expressed during development, while also facilitating recruitment of other transcription factors in the early embryo (Schulz et al. 2015). More recently in *Drosophila,* ZLD was shown to be required continuously throughout zygotic genome activation, as opposed to only being present transiently while *cis*-regulatory regions are established (McDaniel et al. 2019). This study demonstrates the necessity for continued activity of a pioneer factor

for embryo reprogramming and suggests a possibility for a similar requirement for equivalent pioneer factors, such as Pou5f3 and Oct4, in other organisms.

In zebrafish, Nanog, SoxB1 (Sox2 homolog) and Pou5f3 (Oct4 homolog) are all transcription factors ubiquitously expressed in the early embryo (Belting et al. 2001, Okuda et al. 2006, Okuda et al. 2010, Xu et al. 2012). Promoters of pluripotency and highly enriched prior to onset of zygotic genome activation, they are capable both of binding chromatin and eliciting gene expression in the early embryo (Lee et al. 2013, Leichsenring et al. 2013), therefore making them likely candidates for pioneer factors in this organism. Pou5f3 in particular has been found to be involved in temporal control of transcriptional networks, with global loss resulting in a severe disturbance of developmental progression immediately after the 10^{th} cell cycle, the stage at which the genome becomes activated in zebrafish (Lunde et al. 2004, Onichtchouk et al. 2010, Belting et al. 2011). Recently, Veil et al. (2019) have provided a mechanism demonstrating how Pou5f3, SoxB1, and Nanog may facilitate genome activation by binding to and destabilising the so-called high nucleosome affinity regions (HNAR) genome-wide, prior to ZGA, thus maintaining an open chromatin state and allowing the onset of zygotic genome activation.

Despite these findings, the function of Nanog is still highly contested. Gagnon et al. (2018) demonstrated, through use of a Nanog mutant, that Nanog is required for proper formation of yolk syncytial layer (YSL), a multinucleate extra-embryo tissue situated immediately beneath the cellular blastoderm, functioning to

anchor it to the yolk and generate both Nodal and BMP signals required for mesendoderm (the precursor tissue to both mesoderm and endoderm) differentiation. Ultimately, they were unable to identify any direct requirement for Nanog in differentiation of embryonic cells, while also highlighting the fact that the majority of zygotic genes are in fact activated in Nanog morphants (Xu et al. 2012, Lee et al. 2013).

Concordantly, in *Xenopus*, Gentsch et al. (2019) have provided experimental confirmation that the homolog maternal pluripotency factors Pou5f3 and Sox3 (Ventx genes are speculated to be the Xenopus equivalent of Nanog TFs, which thus far lacks a frog orthologue (Scerbo et al. 2012)) remodel chromatin via three methods, namely: co-recruitment of transcription factors, opening and marking of regulatory elements and through facilitating chromatin looping between putative *cis*-regulatory modules and promoters. Their work demonstrates not only the commonality between vertebrate and higher eukaryotes such as mice and humans, in regard to pioneering activity of these factors (Boyer, Lee et al. 2005, Loh, Wu et al. 2006, Orkin and Hochedlinger 2011), but also how they remodel the chromatin, establishing competency in pluripotent cells in responding to inductive signals, necessary for formation of the germ layers during gastrulation (Lunde et al. 2004).

1.4.5 Maternal mRNA degradation

During the first cleavages, prior to MBT, maternally provided RNAs and proteins drive early development and eventual activation of the zygotic genome. The maternal to zygotic transition (MZT) is the overarching event tying together two major events in early embryo development, namely activation of transcription and clearance of maternally deposited factors (**Figure 1-11**) (see reviews by Schulz et al. (2019) and Vastenhouw et al. (2019)). Control of gene expression in oocytes is orchestrated via strict regulation of translation of maternal transcripts, localisation and stability by both micro RNAs (miRNAs) and RNA-binding proteins (RBPs). Degradation of maternal transcripts and proteins occurs concurrently with activation of the zygotic genome. Initially degradation is orchestrated by maternally deposited factors, but as the zygotic genome begins to be transcribed, control of degradation switches from maternal to zygotic (Bashirullah et al. 1999). The link between clearance and genome activation is best demonstrated by the fact that in several species, stabilisation of maternal transcripts occurs as a result of failure to activate the zygotic genome (Bashirullah et al. 1999, Hamatani et al. 2004, De Renzis et al. 2007, Ferg et al. 2007, Lee et al. 2013).

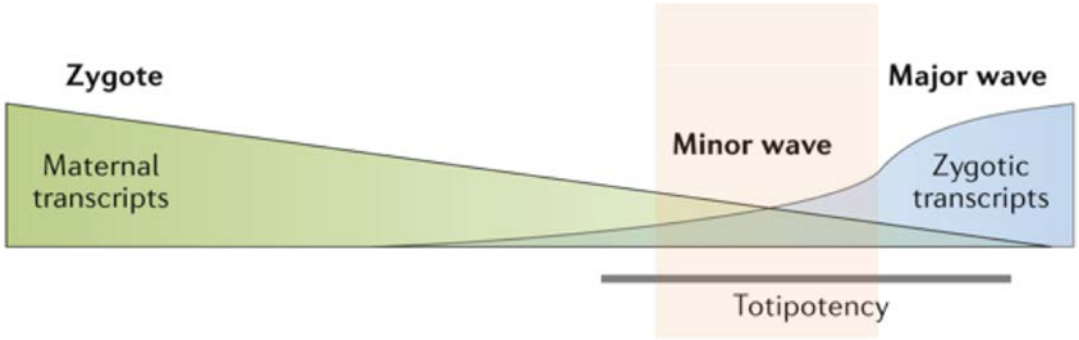


Figure 1-11 Zygotic Genome Activation in metazoa

During the early stages of embryogenesis control of embryo development shifts from maternal factors to zygotic, a transition dubbed Zygotic Genome Activation (ZGA) and begins with expression of the first 'early' genes, unique in that they are able to escape global genome repression by an as-of-yet unidentified mechanism. Adapted from Schulz et al. (2019)

An additional purpose for mRNA degradation may encompass repurposing of building blocks for DNA synthesis (Vastenhouw et al. 2019). In fly embryos, less than half of DNA precursors required to initiate ZGA are maternally deposited, with the remaining being *de novo* synthesised during cleavage cycles (Song et al. 2017).

Inversely, disruption of zygotic clearance, by interference with a major clearance mechanism – specifically the maturation of an integral microRNA (miR-430), results not only in failure to clear maternal transcripts but also of the zygotic genome to become activated (Giraldez et al. 2006).

1.4.6 Structure and function of Micro RNAs

MicroRNAs are small RNA molecules, approximately 20 nucleotides or more in size that do not code for proteins. Instead, they function as post-transcriptional regulators of gene expression by facilitating binding of RNA-induced silencing complex (RISC) to complementary messenger RNAs, leading to either silencing through cleavage, disruption in translation efficiency or destabilisation through deadenylation - thus promoting target mRNA degradation (Ambros 2004, Bartel 2004, Fabian et al. 2010).

The majority of microRNAs are self-contained in the genome with expression controlled by their own promoter. Others, such as *mir-430*, are found in clusters where expression may be co-regulated by other members of the cluster and multiple microRNAs produced as a single primary transcript (pri-miRNA) (Lau et al. 2001). Experimental evidence for this phenomenon has also been shown for *mir-427*, the functional equivalent to zebrafish miR-430 in frog (Lund et al. 2009) and suggested to occur in mammals, where embryonic stem cell-specific sequence-related miRNAs were found to be clustered in a 2.2 kb genomic locus (Houbaviy et al. 2003).

1.4.7 miR-430 is fundamental for early development

In addition to itself being one of the earliest transcribed genes, miR-430 (Chen et al. 2005, Lee et al. 2013, Heyn et al. 2014), and the *X. tropicalis* orthologue: miR-427 (Watanabe et al. 2005, Lund et al. 2009), are key regulators of initial

activation of transcription, with expression shown to be dependent on stem cell factors Nanog, Pou5f3 (Oct4) and SoxB1 (Sox2), implicated in global activation of the zygotic genome in vertebrates (Lee et al. 2013).

Functionally, this miRNA is indirectly responsible for degradation of many maternal factors necessary for activation of the zygotic genome to occur (Giraldez et al. 2006, Lund et al. 2009). The importance of miR-430 during early development was initially demonstrated in the zebrafish, by injection of mature miR-430 into embryos engineered to lack all endogenous functional miRNA transcripts, resulting in reinstatement of normal brain morphogenesis - demonstrating its importance in early development and cell lineage determination (Giraldez et al. 2005).

Zygotic expression of miR-430 is conserved in medaka, another teleost fish (Tani et al. 2010), in addition to *Xenopus* (Watanabe et al. 2005, Lund et al. 2009) and miRNAs of similar sequence to *miR-430a* have also been identified in *Fugu rubripes* and *Tetraodon nigroviridis*. Although there is no sequence similarity, the *mir-309* cluster of miRNA genes in *Drosophila* functions in a parallel pathway, mediating the deadenylation and subsequent clearance of a modest subset of maternal transcripts (Bushati et al. 2008), suggesting targeting of the maternal transcriptome may have evolved separately in vertebrate and invertebrate lineages (Vastenhouw et al. 2019). Clusters of miRNA genes also play a role in early development in higher eukaryotes such as mouse (Houbaviy et al. 2003, Tang et al. 2007) and human (Suh et al. 2004), suggesting a conserved

mammalian-specific developmental role functioning as key regulators in pluripotent cells, for example, during implantation (Houbaviy et al. 2003).

1.4.8 The first zygotic transcription occurs via an unknown mechanism

Advances in sensitivity and specificity for nascent transcript detection and *in vivo* imaging are continuously increasing the resolution and scope of what can be observed transcriptionally, in the early embryo, with the onset of the transcription being brought forward to ever earlier time points (reviewed by (Lee et al. 2014, Vastenhouw et al. 2019)). As such, study of the intricacies underlying these processes is still very much limited by current methods, particularly challenging is distinguishing transcripts from genes that have both a maternal and zygotic contribution. Nevertheless, current opinion is shifting and research by several labs in the field is driving the belief of an alternative understanding of the currently accepted model of ZGA. Moving away from the long accepted activation across 2 waves, with one minor wave driven by maternal factors and a second major wave orchestrated by zygotic factors largely coinciding with the MBT, to an alternative hypothesis that is suggests the process to be more gradual and continuous over time (Chan et al. 2019, Chen et al. 2019, Gentsch et al. 2019, Kwasnieski et al. 2019)

Evidence thus far has shown that a synchronous combination of all the variables discussed, relating to degradation of maternal factors, chromatin competency and remodeling, in addition to coordinated titration of repressors and

accumulation of transcriptional activators, ultimately must act in accordance with one another to strike a balance that will result in an environment amenable to global transcription initiation. Only then at this time, will the embryo be fully prepared to start its own independent life. (Reviewed in: (Baroux et al. 2008, Tadros et al. 2009, Lee et al. 2014, Vastenhouw et al. 2019)). Despite this, it is now known that a population of genes are transcribed prior to the MBT (Heyn et al. 2014, Gentsch et al. 2019, Kwasnieski et al. 2019). The mechanisms by which the first expressed genes can circumvent the global repression holding the rest of the genome in a repressed state are not well understood.

Recent experimental data in frog from Gentsch et al. (2019) shows that the first genes to expressed are related to histones, miRNAs involved in degradation of maternal factors, genes critical in Nodal signaling pathways as well as for formation of the Spemann-Mangold organiser (Spemann et al. 1924), with the equivalent being the embryonic shield in zebrafish *(Shih et al. 1996, Martinez Arias et al. 2018).*

There is evidence emerging that suggests transcripts of the first expressed genes may be short in length, lack introns (Heyn et al. 2014) and have a tendency to be prematurely terminated during their transcription (Kwasnieski et al. 2019); proposed to be a consequence of the quick successive cell cycles of non-mammalian embryos (Tadros et al. 2009, Vastenhouw et al. 2019). Furthermore, their encoded proteins are unique in that they are shorter than variants of those

deposited into the maternal pool and those that are transcribed at the onset of global activation of the zygotic genome (Gentsch et al. 2019).

1.4.9 Transcription of active genes occurs in distinct sub-nuclear compartments

Compartmentalisation of the nucleus has been discussed in relation to chromatin, yet additional membrane-less functional structures exist within the nucleus that lack DNA and instead contain specific RNAs and proteins. These structures are referred to as nuclear bodies and will be discussed here in brief. The largest of these known compartments is the nucleolus, recognised as the site of ribosomal subunit assembly, ribosomal RNA (rRNA) transcription and processing, reviewed in (Olson et al. 2002). Other well-known nuclear bodies are the domains present in the interchromatin regions enriched in pre-mRNA splicing factors known as nuclear speckles, Cajal bodies (CBs) involved in assembly of the splicing machinery and histone locus bodies (HLBs) that are known for their role in histone mRNA 3' end processing (reviewed in (Machyna et al. 2013)).

Membrane-less compartments are believed to be formed by macromolecules creating two distinct phases resulting in formation of different and diverse macro environments locally, the most common being termed liquid-liquid phase separation (LLPS, reviewed in Mir et al. (2019)). Formation of membrane-less compartments such nuclear bodies is believed to drive, at least in part, the compartmentalisation of the nucleus (Brangwynne et al. 2011).

Much evidence hints at formation of distinct hubs where transcription actively occurs. One such theory proposes that LLPS is key in regulating gene expression, believed to function through facilitating localized condensates of transcription factors with coactivators and Pol II (Boija et al. 2018, Chong et al. 2018) that may aid in expediting activation of transcription independently of the transcriptionally repressive environment within the cellular nucleus. This action has been speculated to occur specifically at super enhancers, although have yet been shown to form directly (Cho et al. 2018, Sabari et al. 2018). Furthermore, it has been predicted that these compartments can form even over long distances, facilitating promoter-enhancer interactions and global chromatin structure (Shrinivas et al. 2019).

1.5 Analysis of dynamics of gene regulation by visualisation of transcription

Visualisation of both spatial and temporal patterns of gene expression is of particular interest during early development when specific developmental programmes are adopted, resulting in generation of cell lineages. At early stages each cell exhibits diverse and dynamic patterns of expression and can be bound by various factors influencing their state. Detection of transcription spanning specific time periods has been achieved through various means, such as CAGE (Nepal et al. 2013) and RNA-seq (Pauli et al. 2012, White et al. 2017). The major limitation of these methods is that they are measuring steady state RNA, thus are not suitable for assaying transcriptional dynamics in real time. For this purpose, you need to visualise transcription as it occurs in a living cell.

1.5.1 Methods for visualising transcription

Visualisation of early embryo dynamics has primarily been carried out with the purpose to monitor mRNA localisation to study spatial dynamics of translation (Buxbaum et al. 2015). Over the last decade there has been a categorical explosion in single cell and single molecule imaging, with various techniques developed for visualisation of molecules in cells, a classic example being one of the many variations of *in situ* hybridisation techniques such as single-molecule FISH (smFISH). Although robust and effective, the major drawback of these techniques is that they do not allow imaging of live cells and are not quantitative.

One of the first techniques to track RNA/DNA hybridisation in living cells through visualisation of fluorescent molecules was with molecular beacons (Tyagi et al. 1996). Principally through use of an engineered oligodeoxynucleotide stem-loop structure with both fluorescent matched donor and acceptor chromophores on the 5' and 3' ends respectively, molecular beacons function to prevent emission of signal due to fluorescence resonance energy transfer. Upon hybridisation to a complementary sequence, the stem-loop structure opens and in doing so reduces the fluorescence resonance energy transfer between the donor and acceptor moieties permitting emission of signal upon excitation with the appropriate wavelength of light. This technique suffered from limitations, not only due to difficulties in designing of the stem-loop sequence, but also due to hybridisation occurring only after sufficient threshold density of target sequences was reached, limiting its use for monitoring or quantitating the real time dynamics transcription and RNA tracking.

A more suitable technique for real time tracking of transcription dynamics is the MCP/MS2 bacteriophage system (see **Figure 1-12**). The technique benefits from less delay in signal detection due to the very high affinity of the MCP protein to the MS2 stem loops, reducing the concentration threshold for binding of the target RNA. On the other hand, this technique requires introduction of modified transgenes containing the MS2 stem loop 'tail' and results in a lower signal to noise ratio due to the fluorescence of the MCP fusion protein even when unbound to the target RNA. Nevertheless, this system pioneered by Bertrand et al. (1998) to track *in vivo* mRNA localisation in yeast, was successfully repurposed by

Michael Levine's group to study transcription and transcriptional bursting in the developing *Drosophila* embryo. They demonstrated striking changes in expression patterns over different key developmental time points with clear compartmentalisation in expression from the even-skipped (eve) stripe 2 enhancer (Bothma et al. 2014). In addition to *Drosophila*, visualisation studies of transcriptional bursting *in vivo*, defined as switches between permissive and inhibitory states of transcription initiation, have also been studied using the MS2 system in amoeba (Corrigan et al. 2016).

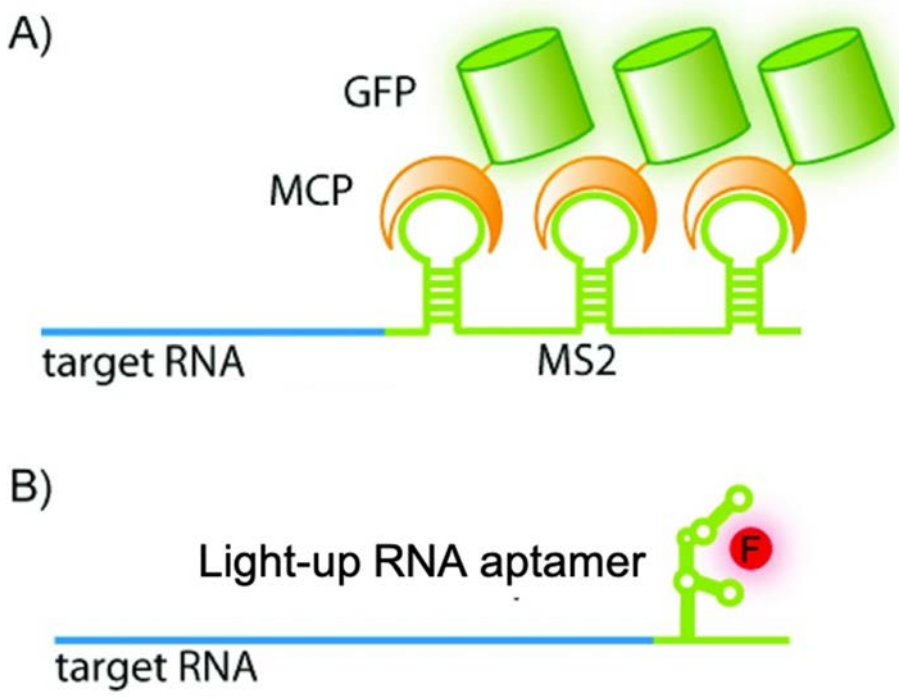


Figure 1-12 reporter-based visualisation of modified RNAs

Both MS2/MCP system rely on the addition of aptamer 'tails' to an RNA of interest which are bound to fluorescent proteins allowing spatio-temporal visualisation of the target RNAs (A) green fluorescent protein-fused-MS2 coat protein (GFP-MCP) binds to consensus sequence on stem loop RNA tail (B) Fluorogenic dye binds complementary fluorogenic aptamer tail. Adapted from (Mannack et al. 2016).

A drawback of this system, as well as other aptamer-based approaches that utilise binding of 'tags' to an RNA molecule, is that the often large increase in size may inhibit the RNAs normal expression, structure, function or localisation (Buxbaum et al. 2015, Garcia et al. 2015). Although also vulnerable to burdening bound RNA, Light-Up RNA aptamers are fluorogenic, thus, like molecular beacons, only fluoresce under permissive conditions, such as in the presence of a target molecule. Aptamer fluorogenicity allows high contrast imaging of RNA transcripts within the cell with fluorescence intensity increasing 1000's of times upon binding a fluorogenic dye, see review (Bouhedda et al. 2018), thus unlike the MCP/MS2 system they allow visualisation of transcription with much greater contrast.

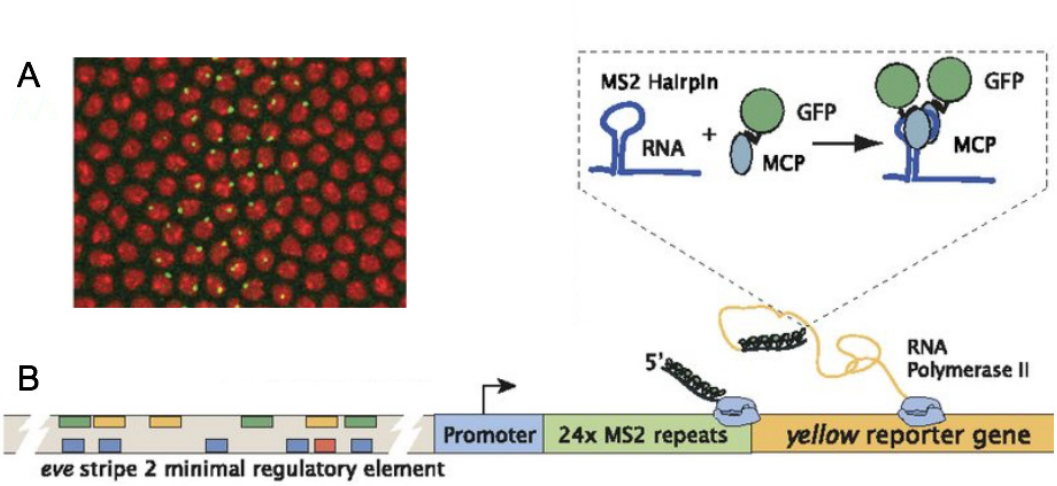


Figure 1-13 in vivo imaging of transcription in the Drosophila embryo

(A) *Projected confocal stack of live imaging of eve stripe 2 compartmentalised transcriptional activity in the Drosophila embryo - shown as a distinct stripe; histones (red), MCP-GFP (green).* ***(B)*** *Structure of the reporter construct: the eve stripe 2 DNA region was placed upstream of 24 repeats of the MS2 stem loops and a yellow reporter gene. The MCP-GFP protein was expressed maternally. Adapted from Bothma et al. (2014).*

Recently the CRISPR-cas system has become an interesting method for visualisation of RNA in live cells. Nelles et al. (2016) demonstrated that an RNA targeting Cas9 (RCas9), when targeted to the nucleus, allows tracking and visualisation of highly abundant mRNA in live cells. More recently, another single protein CRISPR-Cas effector, Cas13, was found to specifically target RNA (Shmakov et al. 2015)

Use of various modified subtypes of Cas13 that, like RCas9 (Nelles et al. 2016), are bound to a fluorescent protein and no longer able to cleave their target substrate (dCas13), allow tracking of RNA in live cells without need for time-consuming modifications to endogenous genes and complex designing of probes (Yang et al. 2019). Furthermore, the system is fairly flexible and has been shown to work reliably when used in conjunction with either the MS2/MCP or RCas9 system for dual colour labelling of RNAs (Yang et al. 2019). The variety of technologies discussed demonstrate how both single cell and single molecule imaging are driving advances within this dynamic field.

1.6 Aims and objectives

In order to explore the question of how transcription of the first genes in the embryo is accommodated within rapid 15 minute cell cycles with lack of gap phases the following objectives were set:

- To develop a system to detect transcription as it occurs in real time:
- To utilise miR-430: the best studied and most highly expressed early gene as model to study first genes to be expressed prior to ZGA
- To characterise miR-430 to understand how it is regulated

To provide further insight into mechanisms underlying the zygotic genome activation, in particular, how the first wave genes can overcome global genome repression and become active before the MBT the following aims were investigated in this thesis:

1. Characterisation of epigenetic and sequence features associated with miR-430 and other early gene promoters.
2. Sequence versus environment: to address whether a single copy of the miR-430 promoter will be sufficient to drive pre-MBT expression outside of its native environment.

2 MATERIALS AND METHODS

2.1 Materials

2.1.1 Reagents, Kits and Equipment

Ampicillin sodium salt	A9518	Sigma-Aldrich
Kanamycin sulfate	K1377	SigmaAldrich
Stellar™ Competent Cells	636763	Clontech
Agarose	BIO-41025	Bioline Ltd
MyTaq™ Mix	BIO-25041	Bioline Ltd
CloneAmp HiFi PCR Premix	639298	TaKaRa
Albumin from bovine serum (BSA)	A3059	Sigma-Aldrich
5-Bromo 4-chloro 3 indolyl phosphate (BCIP)	11383221001	Roche Diagnostics
20x SSC	J60839	Alfa Aesar
MESAB/E-222/Tricaine	E10521	Sigma-Aldrich
Formamide (deionized)	AM9342	Ambion
Gel Loading Dye, Blue 6X	B7021S	NEB
Heparin sodium salt	H3393	Sigma-Aldrich
LB (Luria Bertani) Agar	L2897	Sigma-Aldrich
LB (Luria Bertani)Broth	L3022	Sigma-Aldrich
Nitro blue tetrazolium (NBT)	11383213001	Roche Diagnostics

Paraformaldehyde (PFA)	P6148	Sigma-Aldrich
tRNA from wheat germ	R7876	Sigma-Aldrich
Tween-20	P1379	Sigma-Aldrich
1 kb DNA Ladder	N3232S	NEB
100 bp DNA Ladder	N3231S	NEB
Borosilicate glass capillaries (OD 1mm, ID 0.78mm)	BF100-7810	Warner Instruments
MicroSpin G25 Columns	27532501	GE Life Sciences
Phenol Red	P0290	Sigma-Aldrich
In-Fusion® HD Cloning Kit	639643	Clontech
Pronase from Streptomyces griseus	P5147	Sigma-Aldrich
Restriction enzymes	Varisous	NEB
QIAGEN-tip 20 Plasmid preparation kit	10023	Qiagen
DIG RNA Labelling Kit SP6/T7	11175025910	Roche Diagnostics
mMESSAGE mMACHINE™ SP6 Transcription Kit	AM1340	Invitrogen
NucleoSpin Gel and PCR Clean-up Kit	NZ74060950	Macherey-Nagel
Monarch RNA Cleanup Kit	T2050S	NEB
Digital microINJECTOR System	MINJ-1	Tritech Research
US Flaming Brown micropipette puller	P-97	Sutter Instruments

Heraeus Function line Incubator	B6	Thermo Scientific
NanoDrop ND-1000 Spectophotometer	ND-1000	NanoDrop Technologies
Sensoquest Thermocycler	613-3908	GeneFlow
Stereoscopic Zoom microscope	SMZ1500	Nikon
Zeiss Stereomicroscope	AxioZoom.V16	Carl Zeiss
LightSheet microscope	Z1	Carl Zeiss
Refrigerated Microfuge	Sigma 1-14	SciQuip
Refrigerated Centrifuge	Sigma 2-16KL	SciQuip
Heraeus Microcentrifuges	Pico 17	Thermo Scientific

2.1.2 Buffers and solutions

E3 Medium: 0.5 mM NaCl, 0.17 mM KCl, 0.33 mM $CaCl_2$, 0.33 mM $MgSO_4$.

PBST: PBS, 0.1% Tween-20.

Blocking Solution: 1x PBST, 2% sheep serum (vol/vol), 2 mg/ml BSA

Hybridization Buffer (+): 50% deionized formamide, 5X SSC, 0.1% Tween-20, 50 µg/ml heparin, 500 µg/ml tRNA, the pH was adjusted to 6.0 with 1M Citric acid.

Hybridization Buffer (-): 50% deionized formamide, 5X SSC, 0.1% Tween-20

Staining Buffer: 1 M Tris-HCl, adjusted to pH9.5, 50 mM $MgCl_2$, 100 mM NaCl, 0.1% Tween-20.

Staining Solution: 225 µl of 50 mg/ml NBT and 175µl of 50 mg/ml BCIP in 50 ml Staining Buffer.

Stop solution: PBS pH 5.5, 1 mM EDTA, 0.1% Tween-20.

2.2 Methods

2.2.1 Processing and analyses of Capped Analyses of Gene Expression sequencing (CAGE-seq) data.

CAGE-seq data set from 12 developmental stages Nepal et al. (2013) and nuclear RNA CAGE-seq from 3 early developmental stages 64-Cell, 128-Cells and 256 Cells (previously generated in our lab, data unpublished) were mapped with Bowtie v1.1.2 (Langmead et al. 2009) in default n-mode allowing 2 mismatches in the 27 bp seed region and reporting alignments of reads with up to 1 or 1000 valid alignments for unique and multimapping respectively. The command line parameters used are listed below:

```
Unique mapping:`-S -n 2 -l 27 -a --best -m 1`
Multi mapping: `-S -n 2 -l 27 -a --best -m 1000`
```

The output in Sequence Alignment Map (SAM) format was converted to Binary Alignment Map format (BAM) and sorted by coordinate using SAMtools (Li et al. 2009). The BAM files were provided as input for the CAGEr package (Haberle et al. 2015) which was used for downstream processing of the data, i.e. CAGE Transcription Start Site (CTSS) calling, data normalisation and generation of bigWig/bedGraph formats for data visualisation in the UCSC genome browser as well as generation of CTSS consensus (common across all samples) clusters, which define the core promoter region. The consensus clusters were subsequently assigned to a gene if they were within 500 bp upstream and 150 downstream of Ensembl (Ver. 95) annotated transcript start.

2.2.2 Processing of RNA sequencing (RNA-seq) data.

RNA-seq developmental time course of 18 zebrafish embryonic and larval stages from White et al. (2017) (ERP014517), were mapped to the zebrafish reference genome GRCz10 and Ensembl v87 transcript annotations using STAR (Dobin et al. 2013), allowing 5% mismatches over the total read pair length and reporting alignments of reads with up to 1000 valid alignments. The reads from the 5 replicates for each stage were merged prior mapping. The command line parameters used are listed below:

```
--outFilterMultimapNmax 1000
--outSAMmultNmax 1000
--outFilterMismatchNoverReadLmax 0.05
--alignSJoverhangMin 8
--alignSJDBoverhangMin 1
--outFilterType BySJout
--alignIntronMin 20
--alignIntronMax 500000
--alignMatesGapMax 500000
--outSAMtype BAM SortedByCoordinate
```

STAR was also used to generate genome coverage files in bedGraph format for visualising the data in the UCSC genome browser. By default STAR outputs two coverage (bedGraph) files, one generated from reads which mapped to single location (unique) and the other from all aligned reads in the BAM file (unique and multiple). The command line parameters used are listed below:

```
--runMode inputAlignmentsFromBAM
--outWigType bedGraph
--outWigStrand Stranded
--outWigNorm RPM
```

2.2.3 Processing of ChIP-seq and ATAC-seq data.

Before mapping, ChIP-seq and ATAC-seq reads were quality filtered and adapters were trimmed using TrimGalore (https://www.bioinformatics.babraham.ac.uk/projects/trim_galore/) with automatic adaptor detection option and quality threshold of 20. Only reads where both mates survived the filtering steps were kept in case of paired end data.

The data were mapped to a reference genome using segemehl aligner (https://www.bioinf.uni-leipzig.de/Software/segemehl) allowing up to 2000 alignments per read and alignments accuracy of 90 %:

`segemehl.x -M 2000 -A 0.9`

Genome coverage was generated using STAR :

```
--runMode inputAlignmentsFromBAM
--outWigType bedGraph
--outWigStrand Untranded
--outWigNorm RPM
```

Which as described above, outputs two coverage (bedGraph) files one generated from reads which mapped to single location (unique) and the other from all aligned reads (unique and multiple). The bedGraph files were converted to bigWig format for downstream analyses using bedGraphToBigWig, part of the UCSC genome browser utilities (http://hgdownload.soe.ucsc.edu/admin/exe/).

For analysis data was visualised with the Bioconductor SeqPlots package (https://www.bioconductor.org/packages/release/bioc/html/seqplots.html) providing the coverage files (bigWig format) and genomic regions of features of interest (BED format) as input with 20 bp bin size. From the provided input the getPlotSetArray function calculates the mean coverage signal from all genomic regions, for each 20 bp bin. A code example is shown below:

```
getPlotSetArray(
 bigWig, #genome coverage file in bigwig format
 BED, #genomic regions/features file in BED format
 bin = 20L, #bin size
 type = "pf", #type of the calculation, in this case point
              feature (pf), like TSS.
 xmin = 1500L, #Upstream calculation distance (bp)
 xmax = 1500L, #Downstream calculation distance (bp)
)
```

For visualisation purposes, before plotting, a smoothing spline with smoothing parameter 0.5 was fitted to the aggregation plot data points:

`smooth.spline(x, y, spar = 0.5)`, where the x and y are the values for the x- and y-axis respectively.

2.2.4 Selection of genes sets for epigenetic features comparisons

First wave genes were selected from Supplementary tables S1 and S2 in Heyn et al. (2014). Initially only genes with more than 2-fold upregulation in expression between 128- and 512-Cell stage were kept, and the miR-430 genes were also removed. From the remaining gene only those which were expressed zygotically (without maternal contribution) were kept. To select for zygotic only genes CAGE data was used. A gene was defined as early zygotic if the associated CAGE consensus cluster has less than 1 tpm expression in Unfertilized Egg, Fertilized Egg (1-Cell) and 64-Cell stages and more than 3 tpm at Sphere stage. After applying the above filtering criteria total of 80 genes remained. This was the gene number limiting set. The dominant TSS at Sphere stage was used as reference point in downstream analyses.

Main Wave genes were selected based on CAGE consensus clusters expression. The filtering criteria were as follows:

1. Less than 1 tpm expression in all early stages until and including Oblong stage
2. More than 5 tpm at Sphere stage and more than 10 tpm at 30% Epiboly
3. Not present in the First wave gene set.

Applying the above criteria resulted in 156 genes form which 80 were randomly selected for downstream analyses. The dominant TSS at 30% Epiboly stage was used as reference point in downstream analyses.

Post gastrulation genes were selected based on the following CAGE consensus cluster expression criteria:

1. Less than 1 tpm expression in all stages until and including Shield stage
2. More than 5 tpm at 14-Somites stage and more than 10 at Prim5 stage

Appling the above criteria resulted in 590 genes form which 80 were randomly selected for downstream analyses. The dominant TSS at Prim 5 stage was used as reference point in downstream analyses.

Constitutive gene set was selected form Supplementary table S1 in Haberle et al. (2014) initially prefiltered for those with shifting score greater than 0.7. Subsequently genes which showed relatively constant expression levels across all 12 stages of CAGE data were selected. As criteria for constant expression a coefficient of variation (CV) of less than 0.5 in consensus cluster expression across all 12 stages, was used.

The filtering resulted in 105 genes from which 80 were randomly selected. for downstream analyses. The dominant TSS at Prim 5 stage was used as reference point in downstream analyses.

2.2.5 CpG Island Analysis

Promoter associated CpG island enrichment analyses for each gene set was performed using the CpG islands annotation track form the UCSC genome browser for danRer11 genome assembly: http://hgdownload.soe.ucsc.edu/goldenPath/danRer11/database/cpgIslandExt.txt.gz

A gene promoter was considered associated with CpG island if the promoter region defined as 250 bp up- and downstream from the dominant CTSS overlapped with annotated CpG island longer than 300 bp.

2.2.6 Transcription factor binding motif analyses

Motif enrichment analyses was performed with the AME and CentriMo tools from the MEME (http://meme-suite.org/) suite which was downloaded and installed locally. All the analyses were done with the default parameters providing the relevant sequences and position weight matrixes as input.

Enrichment for clusters of TF binding sited was done using ClusterBuster (http://cagt.bu.edu/page/ClusterBuster_about). The optimal motif weights as abundances of the motifs (occurrences per kb) and the average distance between neighbouring motifs to use with was ClusterBuster calculated using Cluster Trainer (ran with default parameters), a supplementary tool for use with Cluster Buster.

2.2.7 Annotation and SNP detection of miR430 promoter and design of morpholino oligonucleotides for transcription visualisation.

Annotation of promoter regions was defined as 400 bases upstream and 100 bases downstream of the detected CTSS form CAGE-seq data with pre-miR430 genes determined from Ensembl v87 transcript annotations with addition of 2 pre-miR430 c-subtype sequences found by BLAT (Kent 2002) search, completing the a-c-b triplet structure.

Multiple sequence alignment with ClustalOmega (Sievers et al. 2014) was used to detect unique Single Nucleotide Polymorphisms (SNPs) in miR430 promoter sequence, confirmed by unique aligning reads of nascent RNA (Heyn et al. 2014) to two individual promoters in the miR430 cluster.

A region immediately downstream of the dominant CTSS up to the first pre-miR430 gene of the triplet was chosen for design of targeting morpholino oligonucleotides (MOs) for transcription visualisation. Three 25 bp antisense (targeting) and three sense (control) MOs were designed and purchased from Gene Tools Inc, labelled with fluorescein at the 3' end. The optimal sequence (based on hybridization properties, likelihood of hairpin formation and self-dimerization etc.) of each morpholino was determined by Gene Tools using their proprietary software (not available for public use). The sequences of the MOs are

listed below. The numbers in square brackets indicate the start and end position of each morpholino relative to the miR-430 promoter TSS.

```
miR430_MO1_antisense      CACACGCATCTTGTTGTCTGCTGTT [15-39]
miR430_M02_antisense      GCAGCAAGAGAAATCATTGGCAGGC [46-70]
miR430_MO3_antisense      AGCAGATTAAAACCTTGTCCGTTCA [79-103]
miR430_M01_sense     AACAGCAGACAACAAGATGCGTGTG [39-15]
miR430_M02_sense     GCCTGCCAATGATTTCTCTTGCTGC [70-46]
miR430_M03_sense     TGAACGGACAAGGTTTTAATCTGCT [103-79]
```

2.2.8 Zebrafish maintenance

All zebrafish strains were maintained in designated facility (according to UK Home Office regulations) in a recirculating system (ZebTEC, Tecniplast) at 26°C in a 10-hour dark, 14-hour light photoperiod and fed 3 times daily.

Zebrafish experiments were restricted to early embryos and adults were only used for natural breeding. Animal work presented in this study was carried out under the project licences 40/3681 and P51AB7F76 assigned to the University of Birmingham, UK

Zebrafish embryos were obtained by putting together a pair of zebrafish (male and female separated by a divider) in specifically designed crossing cage in the evening. On the next morning after start of the light period, the dividers were removed, bringing the two sexes together which initiates spawning. Embryos

were collected shortly (5-15 min) after laying by using either a small net or tea strainer.

2.2.9 Dechorionation of zebrafish eggs

After collection, freshly laid zebrafish eggs were transferred into a small (6 cm) petri dish and the excess liquid removed, ~1 ml pronase (10mh/ml in E3 media) was added and eggs were incubated until the one/few eggs come out from the chorion, then immediately transferred into 500 ml beaker with system fish water and washed 3 times, to ensure complete removal of the pronase. Dechorionated eggs were transferred into agarose coted petri dishes and used in downstream applications like microinjections.

2.2.10 Microinjection of zebrafish embryos

Single-cell stage embryos were injected with 1-2 nl solution containing the desired reagents (reporter plasmid DNA 20 ng/µL, morpholino oligonucleotides 7 µM, mRNA) supplemented with 0.1% of phenol red (Sigma) as injection marker. Injections were performed through the chorion (or if necessary, for the downstream applications embryos were dechorionated prior injections) into the cytoplasm of zygotes using an analogue microinjector MINJ-1 (Tritech Research). After injection Embryos were kept in E3 medium at 28.5 °C until they reach the desired developmental stage.

2.2.11 Isolation of zebrafish genomic DNA

Zebrafish genomic DNA was prepared using the DNeasy Blood & Tissue Kits (Qiagen), following the provided manual. In brief, about 50 AB wild type zebrafish embryos at 24 hpf were homogenised in buffer ATL proteinase K solution was added and the lysate incubated at 55 °C for 1 hour. The genomic DNA was purified using the provided spin columns according to the instructions in the manual. Miniprep of plasmid DNA

2.2.12 Isolation of plasmid DNA

Plasmid DNA preparation was carried out using QIAGEN-tip 20 miniprep kit (QIAGEN) following manufacturer's instructions, which are described briefly below. 5 ml overnight LB bacterial cultures were pelleted in a standard benchtop microcentrifuge for 30 seconds at 11,000 g. After discarding the supernatant, 300 µl of P1 Buffer supplemented with RNase A was added to resuspend the cell pellet. 300 µl of P2 Buffer was then added to lyse the cells and the tube was inverted several times until the lysate was cleared. After 5 minutes of incubation 300 µl of P3 Buffer were added and the tube inverted several times until white precipitated was formed and separated from the lysate. The lysate was clarified by centrifugation at 11,000 g for 5 minutes. The supernatant was loaded onto a tip20 anion-exchange column and after 2 washes with wash buffer and the plasmid dan was eluted from the column with elution buffer and precipitated by adding 0.8 volumes isopropanol and centrifugation at 15000 g for 30 min. DNA pellet was washed with 70% ethanol

2.2.13 Polymerase Chain Reaction (PCR)

Amplification of DNA sequences was performed by PCR using a Sensoquest Thermocycler (GeneFlow, UK). For cloning purposes CloneAmp HiFi PCR Premix (TaKaRa) was used. For other purpose like colony screening MyTaq™ Mix (Bioline) was used. PCR reactions were assembled and performed according to the provided protocols of the enzyme mix manufacturers.

2.2.14 Restriction digest

All restriction digests were performed using New England Biolabs (NEB) High-Fidelity (HF®) Restriction Enzymes. Up to 5 µg of DNA were digested in 50 µl total rection volume. If more than 5 µg of DNA was used, the rection volume was scaled up to 100 µl. 1-2 U of restriction enzyme per µg of DNA was used and together reaction mixture was incubated for 2-3 hours at the recommended for the enzyme temperature (in most cases 37 °C). The digested DNA was assayed by agarose gel electrophoresis and subsequently purified for downstream applications.

2.2.15 Agarose Gel electrophoresis

Agarose gels were prepared by dissolving agarose in Tris-acetate -EDTA (TAE) buffer to a final concentration ranging from 0.8 to 1.5% depending on the fragment size. To allow visualization of DNA under UV light 10ng/ml of ethidium bromide was added to the gels. As marker 1 kb DNA Ladder Plus from NEB was used. Gels were run in TAE 1X Buffer electrophoresis buffer at 3-5 V/cm until

desired separation of DNA or RNA (probes for *in situ* hybridisation or *in vitro* transcribed mRNAs) bands was achieved.

2.2.16 Cloning of mir430 promoter region

The miR430 promoter region fragment (~650 bp) was PCR amplified from zebrafish using 100 ng genomic DNA 1x CloneAmp HiFi PCR Premix (TaKaRa) with 0.5 µM of each in a 50 µl reaction. The cycling parameters were as follows:

```
Initial denaturation: 98°C 1 min
Cycle denaturation: 98°C 15 sec
Aneling: 60°C 15 sec                 - 30x
Extension: 72°C 45 sec
Final extension: 72°C 2 min
```

The PCR products were purified using NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel) and cloned into modified *pJet:attB:mCherry* vector (Roberts et al. 2014) containing phiC31 attB site and mCherry reporter (plasmid maps provided in appendix 8.8), thus making suitable for use in stable transgenic reporter lines generation, using the phiC31 integrase mediated targeted transgene integration system (Roberts et al. 2014, Hadzhiev et al. 2016).

The cloning procedures were performed using In-Fusion HD Cloning Kit (TaKaRa/Clontech) according to the manufacture instructions. In brief the pJet vector was digested with Acc65I and NotI restriction enzymes (NEB) and purified

with NucleoSpin Gel and PCR Clean-up Kit (Macherey-Nagel). The linearized vector and purified miR430 promoter fragment in 1:3 molar ratio were incubated with 1x In-Fusion HD cloning mix at 50 °C for 15 min and 2 µl of the cloning reaction was transformed into the provided with the kit Stellar strain competent cells. Positive colonies, carrying plasmid with the expected insert length were identified by PCR screening using standard T7 primer which is present in forward and reverse orientation flanking the multiple cloning site (see map).

In brief the colony PCR was performed as follows: Single colonies were picked up and individually resuspended in 50 µl LB-medium with the appropriate antibiotic and incubated for 3-4 hours at 37 °C. One microliter of the mini culture was used as template in 20 µl PCR rection assembled using the MyTaq™ Mix (Bioline) with 1 µM T7 primer (TAATACGACTCACTATAGGG) with the following cycling parameters:

Bacteria lysis and Initial denaturation: 94°C 5 min

Cycle denaturation: 94°C 30 sec

Aneling: 55°C 40 sec

Extension: 72°C 2 min

} - 30x

Final extension: 72°C 5 min

Plasmid DNA was isolated using QIAGEN-tip 20 miniprep kit (QIAGEN) following the provided protocol, verified by restriction digest and Sanger sequencing (service provided by Source BioScience).

Primer sequences used (lower case letters correspond to the vector homology arms, upper case letters are specific to the miR-430 promoter regions):

miR430_0.6Prom_FP	actcacgtgtggtaccCTTAGTCACCTCTGCCCACCAAG
miR430_0.6Prom_RP	catcgataagcggccgcTCCGTTCATAGTCTTTAGCAGCAAG

The control reporter plasmids *pCS2:mCherry* and *pDB896:-0.8gata2a:mCherry* were previously generated by a colleague (maps provided in appendix)

2.2.17 Synthesis of antisense RNA probes for *in situ* hybridization

Antisense RNA probes were labelled with digoxigenin by in vitro transcription with T7 polymerases using a DIG RNA labelling kit from Roche Diagnostics. Probe against the mCherry reporter was invitro transcribed using linearized with BamHI *pCS2+:mCherry* vector as template. Gal4VP16 antisense probe was synthesised by using modified by a colleague pDB783:crygc:Gal4VP16 plasmid (gift from Darius Balciunas, Temple University. Plasmid map provided in appendix 8.8), linearized with Acc65I, as template.

Transcription reactions 20 µl) were assembled according to the provided protocol. In brief, 500-1000 ng DNA template, 1x Transcription buffer, 1x DIG-labelling NTP mix with 2 µl of T7 polymerase were incubated at 37 °C for 2 hours. The

DIG-labelled probe was further purified by filtering it through Illustra MicroSpin G25 Columns. The probe quality was checked by agarose electrophoresis.

2.2.18 *In vitro* transcription of *PhiC31* mRNA

Messenger RNA *of* the *PhiC31* integrase containing the zebrafish *nanos* 3'UTR, directing the mRNA localisation predominantly into the primordium germ cells (PGS), thus increasing the chances of germline transmitted integration events, was *in vitro* transcribed using the mMESSAGE mMACHINE™ SP6 Transcription Kit (Invitrogen, AM1340). As template NotI linearized *pCS2+PhiC31o-nos1-3'UTR* plasmid (Roberts et al. 2014) was used. The transcription reaction (total volume of 20 µl), containing 1 µg linearized plasmid template, was assembled according to the provided with the kit protocol and incubated for 2 hours at 37°C. The template DNA was removed by treatment for 15 min at 37°C with TurboDNAse I (provided with the kit). The mRNA was purified using the Monarch RNA Cleanup Kit (NEB T2050S) following to the provided instructions, quantified on NanoDrop and integrity confirmed by agarose gel electrophoresis.

2.2.19 Whole mount in situ hybridisation

Zebrafish embryos were fixed in 4% paraformaldehyde in PBS overnight at 4°C, transferred into 100% methanol on the next day and kept at -20°C until further processing. Embryos were rehydrated in dilutions of methanol in PBST (75%, 50%, 25% methanol-PBST) at RT for 5 minutes in each solution.

Embryos were pre-hybridized in 500 µl of hybridization buffer for 2-4 h at 65°C and hybridized overnight at 65°C, in 500 µl of hybridization mix containing 1ng/ul DIG-labelled RNA probe.

The next day embryos were washed with series of hybridization buffer in 2X SSC dilutions (75% 50% 25%) at 65°C,15 minutes for each dilution, then twice in 0.2X SSC at 65°C for 30 minutes. Next, 0.2 X SSC was replaced with PBST through 2 x 5 minutes washes at room temperature. Embryos were incubated in blocking buffer for 2-4 hours at room temperature. The blocking solution was replaced with Anti-Digoxigenin-AP antibody (Sigma Aldrich, #11093274910) containing blocking buffer in 1:4000 dilution and incubated overnight at 4°C. Next day embryos were washed with gentle agitation in PBST 4 times, for 30 minutes each wash. Embryos were washed once with staining buffer for 5 minutes and transferred to a 24-well plate and staining buffer was replaced with 1 ml of fresh staining solution. The colorimetric reaction was monitored closely under the microscope and embryos were kept in the dark between checks.

When the staining intensity and pattern desired was reached, the reaction was stopped by adding stop solution for 3 minutes. Embryos were washed 3 times in PBST and then fixed in PFA 4% for 30 minutes at 4°C and dehydrated through a series of washes with methanol diluted in PBST. Whole mount *in situ* hybridisation were imaged on Nikon SMZ1500 stereo microscope with DMX1200 colour camera.

2.2.20 Generation of stable transgenic zebrafish reporter lines.

Zebrafish reporter lines were generated by following the PhiC31 integrase mediated targeted transgene integration protocol as described in (Hadzhiev et al. 2016). In more detail one cell stage docking line embryos were microinjected with 1 nl solution containing the reporter plasmid (20 ng/µl) supplemented with *PhiC31-nanos* 3'UTR mRNA (30 ng/µl). The efficiency of integration events was determined by frequency of lens colour change (from green to red) in the injected embryos at 72-96 hpf. If lens colour change is observed in more than 5 % of the embryos in an injection batch all the embryos from the batch are grown to adulthood, including those without lens colour change, because integration events may have happened with higher frequency in the PGCs due to nanos 3'UTR driven localisation of the PhiC31 integrase mRNA occurring predominantly in the PGCs.

The adult zebrafish (F0 generation), when approximately 3 months old, where crossed pairwise and the and germline transmission was determined by lens colour change in the F1 embryos at 96 hpf. Identified F0 founder fish with germ line transmission were subsequently outcrossed to wild type fish, embryos with red lens were selected and grown to adulthood to establish a stable transgenic line.

2.2.21 Live embryo imaging.

Imaging of live embryos injected with fluorescent labelled morpholino oligonucleotides was performed on Zeiss Lightsheet Z1. Embryos were mounted in 1% low melt agarose column using size 3 glass capillaries and incubated in E3 media at 28°C during imaging. Z-stacks of approximately 200-300 slices in 0.5-1 µm steps were acquired every 30-60 seconds for 1.5-2.5 hours, with 20x objective. Post-processing of the acquired images was performed using Zeiss ZEN Black/Blue Software.

Embryos injected with miR430 promoter construct were screened for expression and imaged on Zeiss Axio Zoom V16 stereo microscope.

3 CHARACTERISATION OF A MODEL GENE FOR REAL-TIME VISUALISATION OF TRANSCRIPTION IN THE DEVELOPING EMBRYO

Foreword:

The results presented in this chapter have been partially published in (Hadzhiev et al. 2019).

3.1 Introduction

Development in most metazoan embryos begins with a set number of cell divisions absent in transcription, instead driven by maternal factors deposited in the oocyte, until global genome activation of the zygote occurs. However, there is accumulating evidence that several genes escape this global transcriptional block and are activated before the global onset of transcription (Vastenhouw et al. 2019). Included amongst the first genes to be expressed in the developing zebrafish embryo are factors that play a role in activation of the zygotic genome such as miRNA genes (Heyn et al. 2014), responsible for facilitating clearance of maternal RNAs (Giraldez et al. 2006).

Nuclear topology and conformation of chromatin are essential for coordination of gene expression (Zheng et al. 2019), yet it is not well understood how such nuclear organisation is related to dynamics of transcription in both time and space. Therefore, to address how the first wave genes can circumvent transcriptional repression before N/C ratio is reached at the MBT there is a need for development of a model where transcription can be monitored at single cell resolution, in real time, in the developing embryo natively, without need for transgenic manipulations.

3.2 Aims

Using *mir-430*, a multicopy highly expressed gene pre-MBT as a model to visualise transcription both *in vivo* and in real time on a native gene using morpholino oligonucleotides during the maternal to zygotic transition.

Due to its highly repetitive nature, we anticipate that it in addition to providing a suitable locus for utilisation and manipulation for the study of visualisation of gene expression it will also act as a proxy for the exploration of promoter dynamics of the earliest expressed genes in the embryo.

3.3 Computational analysis of *mir-430* locus for model validation

Structurally, the *mir-430* locus is composed of 4 different subtypes of precursor gene: a, b, c and the 'a' variant 'i', organised into triplets (see **Figure 3-1**). The current genome assembly, danRer11, has 9 repeating units, each containing between 2 and 3 *mir-430* triplets. A conservation of the miR-430a subtype has been found both in Fugu and Tetraodon (Chen et al. 2005) although the triplet structure appears not to be conserved. Despite this, the such a structure is not specific to zebrafish, frog *mir-427* is also predicted to be organised in a string of tandem repeats, each consisting of a series of 4 distinct miRNA subtypes (Lund et al. 2009). miRNAs clusters similar in sequence to miR-430 are also present in higher eukaryotes, mouse and human (Houbaviy et al. 2003, Suh et al. 2004). This distinctive structure with highly repetitive nature makes the locus challenging to study yet offers potential to utilise this repetition for enhancement of localisation signal, facilitating visualisation of the locus, particularly in relation to the study of transcription.

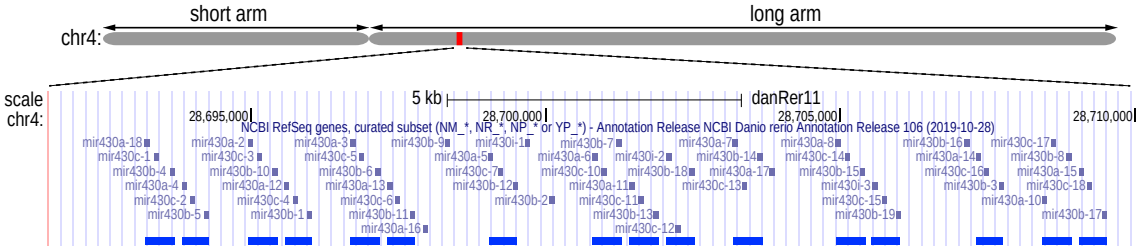


Figure 3-1 miR-430 endogenous locus structure from assembly GRCz11

The endogenous locus structure of miR-430 locus according to the genome assembly GRCz11 using RefSeq version 106 annotation. Annotated miR-430 subtypes are arranged into triplets of three miR-430 subtypes 'a/i', 'c', 'b' on the bottom-most line represented by solid blue bars.

Heyn et al. (2014) found miR-430 is one of the earliest and most highly expressed genes pre-MBT making it an ideal candidate model gene locus for visualisation of transcription dynamics in the early embryo. Likewise, in frog, miR-427 transcripts were found to accumulate at a rate near 1000-fold higher than any other early transcribed gene (Owens et al. 2016).

3.3.1 Mapping of NGS data revealed 8 functional promoters

The complication of dealing with repetitive sequences comes from the resulting ambiguity, it is often not clear which sequences are distinct from one another. This can be problematic both when assembling contiguous sequences, such as chromosomes, and also when mapping sequencing data onto the repetitive loci.

The most common strategy of dealing with repeats is to either remove them or take either the first or best alignment. Assignment of reads unambiguously relies on the identification of SNPs. More SNPs present in a sequence the higher the confidence will be when mapping reads to a particular location. Another option is to keep all of the reads aligning to particular loci, this results in over-amplification of signal, but can be useful for qualitative analysis (Treangen et al. 2012).

In order to analyse promoter dynamics at the miR-430 locus, CAGE tags from whole cell extract (WCE) (Nepal et al. 2013) were multi-mapped without restriction on number of alignments permitted. For the purpose of analysing the temporal dynamics of the gene cluster this qualitive mapping approach was necessary (**Figure 3-2**), it was not possible to assign CAGE tags to the locus unambiguously due to the lack of SNPs present in the 27 bp length of the CAGE tag.

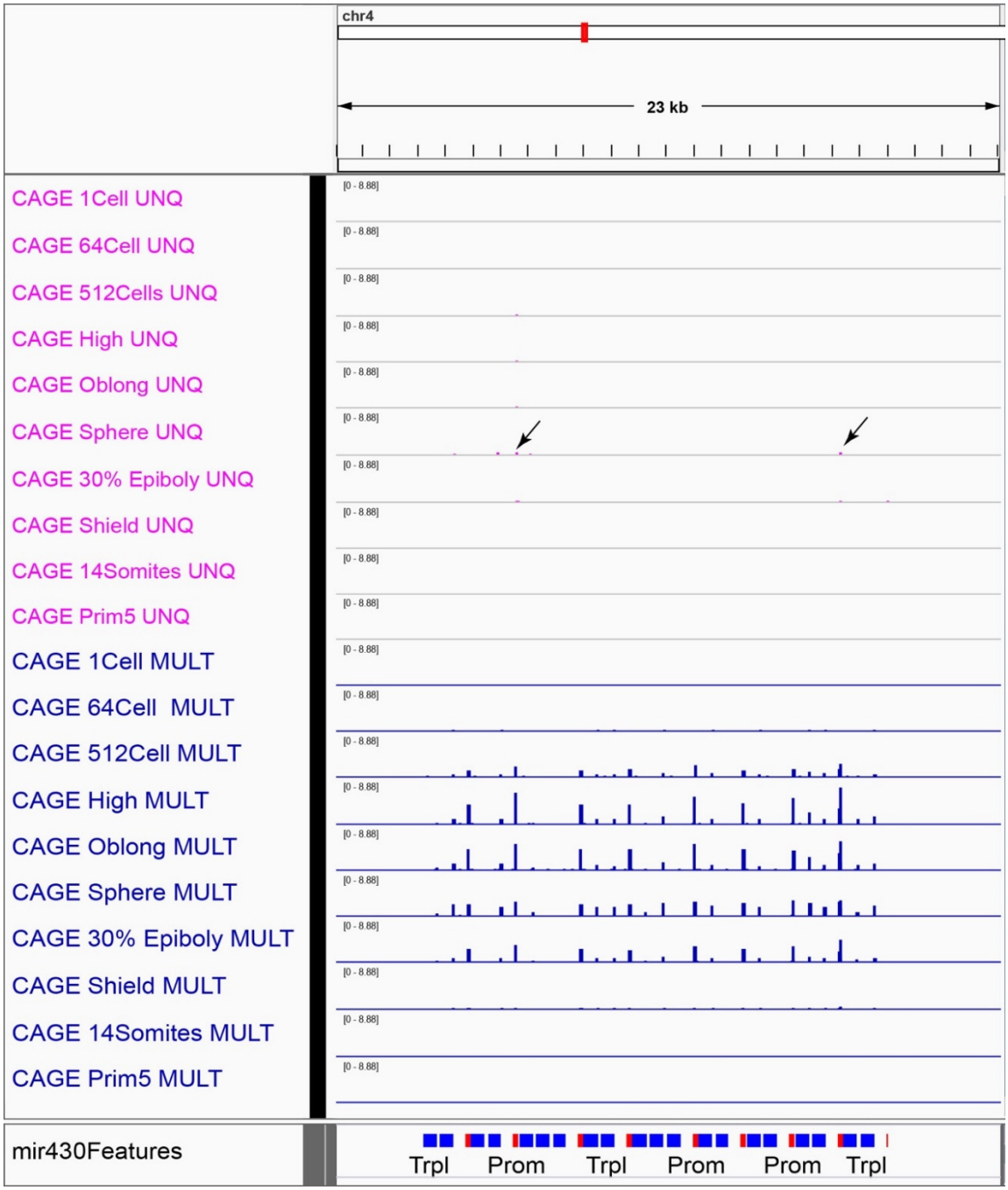


Figure 3-2 Demonstration of differences in CAGE-seq coverage between unique and multi-mapping approaches

Genome browser view of miR-430 locus on chr4 on genome assembly GRCz11/danRer11. Upper pink coloured rows demonstrate coverage with default unique mapping approaches, with examples of uniquely mapped CAGE-seq data marked with arrows. In contrast, lower blue rows highlight the coverage achieved when mapping the same CAGE-seq data with described multi-mapping approach.

In general, for downstream analysis of NGS data, reads which are assigned to multiple locations are either discarded or ignored for the purpose of achieving more robust normalisation of the data. Yet in the case of repetitive regions, such as mir430, this hinders the analysis, as demonstrated in **Figure 3-2**, where the difference in coverage from both unique and multi-mapped mapped reads is demonstrated.

Mapping of CAGE tags to zebrafish genome (**Figure 3-3**) demonstrated 8 possible TSSs at the miR-430 locus for gene units 1-8 (P1-P8), although CAGE tags did not align upstream to P1. Analysis of the underlying sequence at the start of the locus revealed a truncation of the upstream structure, resulting in a lack of promoter sequence with no evidence of an upstream promoter combined with a truncation of the first miR-430 ‘a’ precursor, together providing an explanation for the lack of CAGE signal at the first gene unit (see **Figure 3-4**, **Figure 3-5** and **Figure 3-6**).

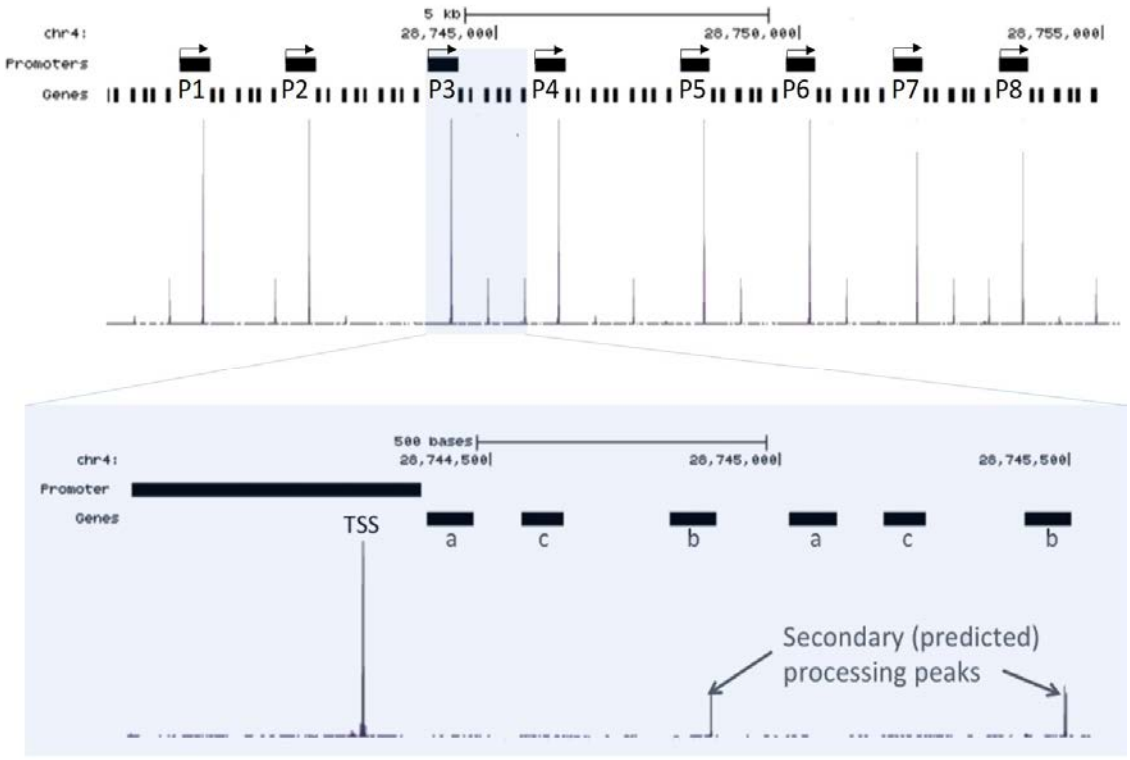


Figure 3-3 Overview of miR-430 locus structure showing mapped CAGE tags demonstrating 8 predicted promoters

Schematic representing miR-430 locus structure with miR-430 genes and predicted promoters shown as small and large black boxes respectively. Lower panel shows an enlarged view of gene unit P3 with the TSS and predicted post-processing peaks indicated by mapped CAGE tags.

Secondary CAGE tag peaks are also present at the end of many triplets (**Figure 3-3**); it was found that this position aligns with a predicted cleavage site of one of the miRNA processing apparatus, Drosha. It has been shown by Nepal et al. (2015) that in addition to other post-transcriptionally processed RNA products (Fejes-Toth et al. 2009, Nepal et al. 2013) CAGE-seq can detect such so-called micro-RNA offset RNAs (moRNAs) resulting from pri-miRNA post-transcriptional

processing, previously described in a simple chordate *C. intestinalis* (Shi et al. 2009) in addition to mammals (Langenberger et al. 2009).

3.3.2 *miR-430* locus has a classical promoter architecture

Analysis of DNA sequence surrounding the promoter revealed presence of the TATA box and 'C-A' initiator core promoter elements in all eight promoter containing gene units (**Figure 3-4**). Through combination of CAGE profiles with analysis of surrounding sequence motifs it has been possible to develop a promoter feature profile for this region, strongly indicating likely presence of 8 functional promoters with a classical architecture.

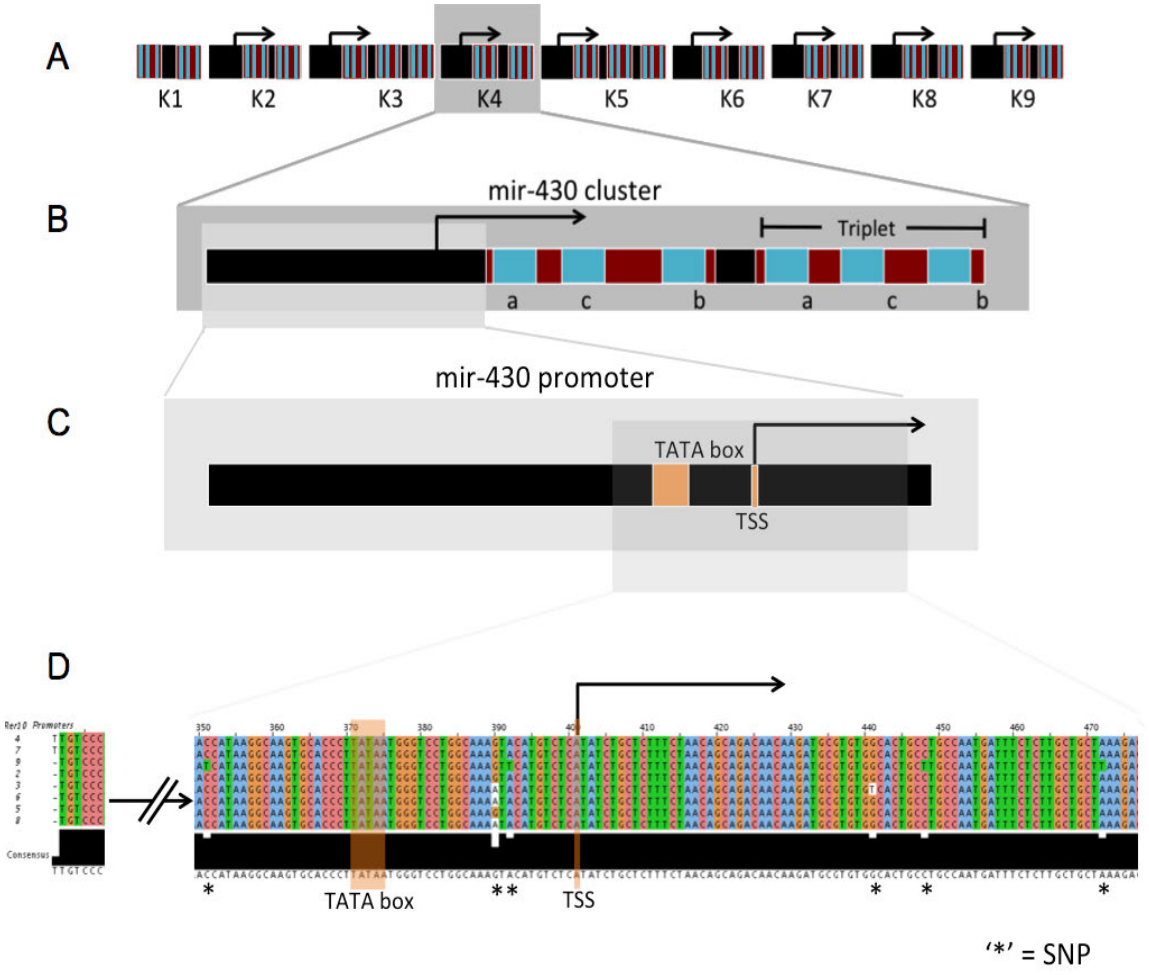


Figure 3-4 miR-430 is representative of a highly repetitive, gene dense, early expressed locus

(A) *Schematic showing predicted structure of miR-430 locus based on TSS mapping by CAGE* ***(B)*** *Enlargement of K4 showing clear representation of triplet arrangement of mir-430 subtypes;* ***(C)*** *Schematic of promoter region with detected TATA box and TSS;* ***(D)*** *Multiple sequence alignment of 500 nts surrounding the TSS of GRCz11/danRer11 zebrafish genome assembly. Colours are indicative of base composition (blue = 'A', red = 'C', green = 'T', orange = 'G') with each row representing one of the 8 promoter containing gene units. Sequence consensus represented by continuous black region – gaps are highlighted by asterisks and represent single nucleotide polymorphisms (SNP).*

A more in-depth analysis of sequence conservation between each element of the *miR-430* locus has revealed striking sequence similarity – impeding attempts to distinguish between transcripts using single nucleotide polymorphisms (SNP) (promoter similarity shown in panel D, **Figure 3-4**). In order to assess the potential functionality of the promoters present, several SNPs were identified in both the up- and downstream regions flanking the TSS through multiple sequence alignment. Unfortunately, due to the short length of CAGE tags (20 nts) it was not possible to assign tags to individual promoters.

A second attempt at SNP analysis was made with nuclear CAGE data (from our lab, unpublished). This dataset has been generated using the no-amplification non-tagging CAGE (nAnTi-CAGE) protocol (Murata et al. 2014). In contrast to the classic tagging-CAGE protocol (Takahashi et al. 2012), in the nAnTi-CAGE protocol, the enzymatic cleavage step with *EcoP15I* limiting the tag length to 27 bp is omitted, thus the tag length is limited only by the sequencing technology used. The nuclear nAnTi-CAGE has been sequenced on Illumina Hiseq 2000 platform with 50 bp single end reads, yielding 47 bp CAGE tags, after trimming of the 3 bp barcode. The longer tag length, almost twice the length of the WCE CAGE tags (Nepal et al. 2013) allowed the identification of SNPs, predicted to be present at 40 bp downstream of the TSS (**Figure 3-4**) from analysis of GRCz11/danRer11 zebrafish genome assembly data. Unfortunately, it was not possible to detect this SNP in any of the sequenced tags and there were no novel SNPs present in a representative number of transcripts. Demonstrated by the fact

that on average each SNP was only represented between 1 and 5 times over several million reads.

3.3.3 Temporal dynamics of miR-430 expression during ZGA

Several studies indicate that transcription of genes may occur as early as 64 cell stage at 2 hpf, before the global onset of ZGA at 1000 cell stage/MBT at 3 hpf (Mathavan et al. 2005, Heyn et al. 2014). One such gene purported to be the most highly expressed gene pre-MBT is miR-430 (Heyn et al. 2014), known to facilitate the deadenylation and clearance of maternal mRNAs during the transition between maternal and zygotic controlled transcription (Giraldez et al. 2006).

The expression of miR-430 steadily increases until it peaks at dome stage, during the transition to the zygotic-specific expression programme, after which it is then successively degraded and only found in trace amounts in larvae (**Figure 3-5**). Detection of secondary processing peaks (indicated in **Figure 3-3**) by CAGE-seq alongside promoter-associated tags demonstrates that post-transcriptional processing of the pri-miR-430 transcript is occurring co-transcriptionally.

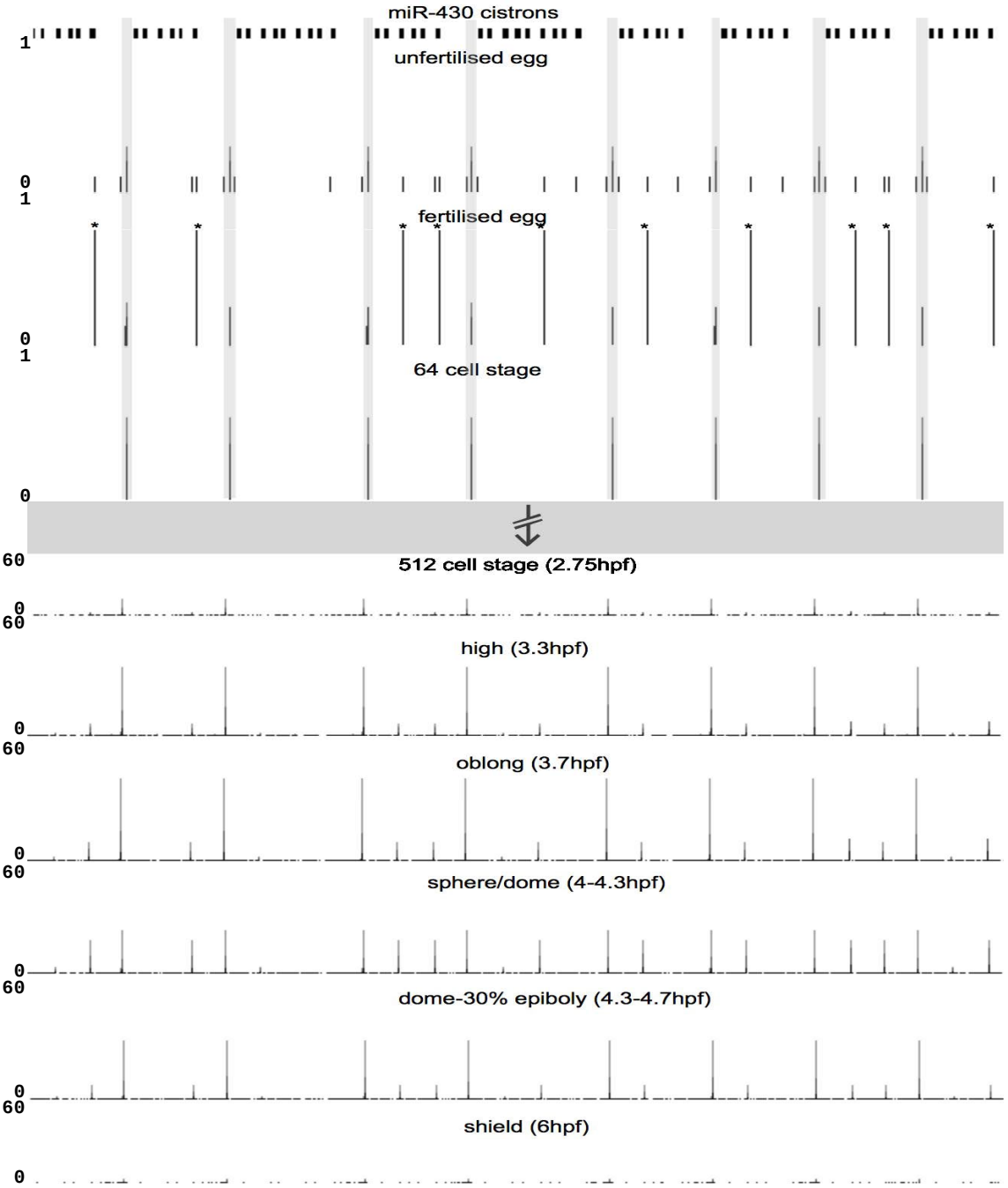


Figure 3-5 Temporal dynamics of miR-430 expression during development

Two separate fixed scales are used to highlight dynamics between similar time points: for unfertilised-64 cell stage and for 512 cell stage-shield. TSSs are highlighted and asterisks mark secondary peaks for unfertilised-64 cell stages. At unfertilised egg stage both TSS peaks and small secondary peaks (marked with asterisks) are detected, although after fertilisation there is a substantial preference for increase in secondary

peaks relative to TSS and by 64-cell stage the transcript abundance at TSS begins to increase.

Further analysis of the earliest stages via CAGE-seq data enriched for nuclear RNA has shown a steady increase in transcript abundance between 64-cell stage and 256-cell stage (**Figure 3-6**), providing evidence via an independent technology to corroborate others indications regarding the possibility of transcription of *miR-430* occurring pre-ZGA.

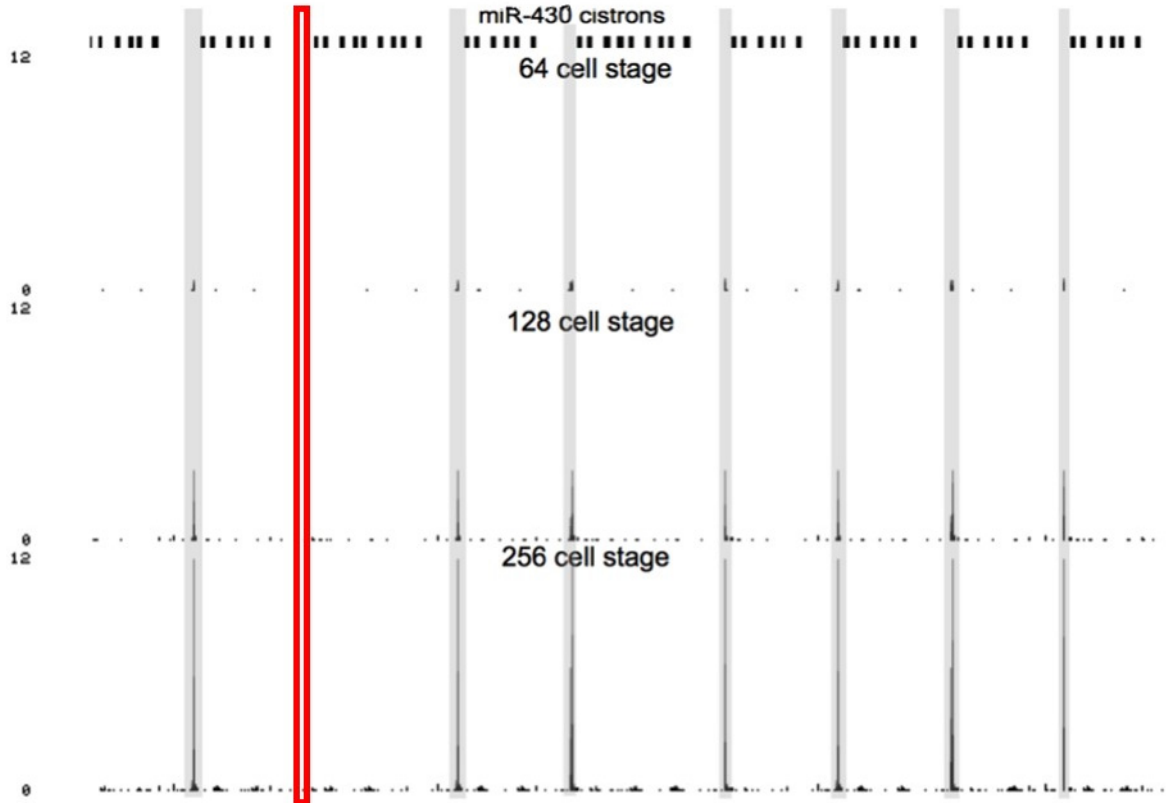


Figure 3-6 Evidence for mir-430 transcription pre-MBT

CAGE-seq data specific to nuclear extracts is shown mapped to the miR-430 locus on chr4. TSSs are highlighted in grey with pre-ZGA stages 64 -256 cell stage shown. Location of predicted promoter region that does not have an assigned CAGE tag peak is demarcated by a red box which surrounds the expected peak location.

A disparity in SNP profiles between the genome sequence of the third promoter and nuclear CAGE tags blocked the binding of CAGE tags to that particular promoter (**Figure 3-6**). Despite this discrepancy, the nuclear CAGE data confirms two observations, firstly the expression dynamics shown by the WCE CAGE in **Figure 3-5**, in particular the suggestion of an increase in promoter activity before the onset of ZGA. Secondly it confirms the observation of a classical promoter architecture associated with adult tissue specific genes, consisting of a TATA box motif and Inr sequence in the core promoter sequence with a characteristic single start site of transcription giving a sharp profile (Lenhard et al. 2012, Haberle et al. 2016, Haberle et al. 2018) as detected by both CAGE-seq sets.

3.4 A model for spatio-temporal visualization of transcription initiation during early development *in vivo*

In order to assess whether detectable transcription could be observed indirectly through accumulation of fluorescence signal at an endogenous locus in real time, an array of morpholino oligonucleotides (MOs) specific for the miR-430 promoter were designed to bind to the 5' end of the unprocessed pri-miRNA transcript. Localised enrichment of signal was predicted to occur due to a series of three adjacent MOs (see appendix 8.2) binding to the repeating promoter units within the cluster, resulting in amplification of the signal. Amplification of signal would then occur across all transcribing clusters (**Figure 3-7**).

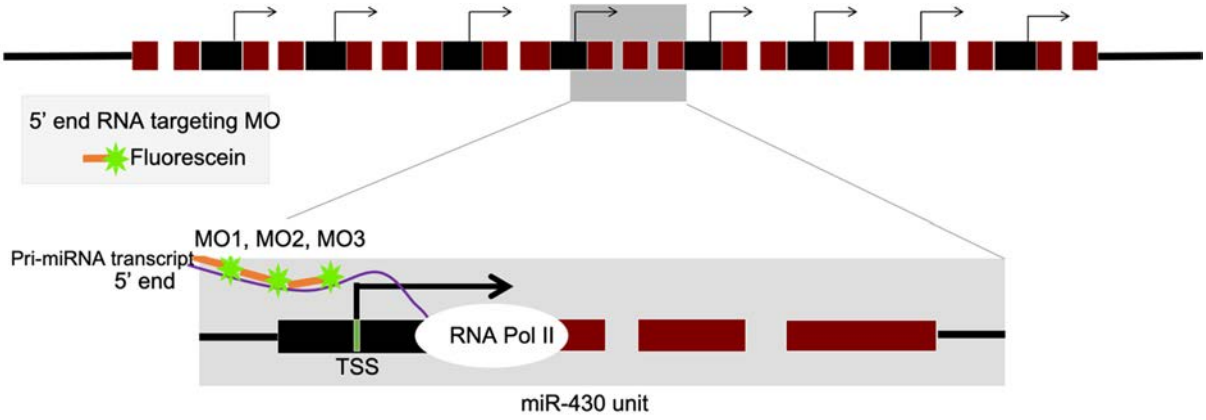


Figure 3-7 a model for visualisation of transcription in real-time during early genome activation

Schematic of miR-430 locus with enlargement of a promoter containing unit. Illustration of three fluorescent morpholino oligonucleotides designed to bind to the 5' end of the transcript during active transcription by RNA Pol II.

Following injection of embryos with the series of miR-430 promoter targeting MOs it was possible to detect two distinct transcription foci, via lightsheet imaging, within the nuclei of the developing embryos both in real time and prior to onset of ZGA. Lack of signal detection in the sense/DNA binding MO demonstrates that this event is dependent on RNA (**Figure 3-8**) and furthermore, the dependence on transcription has been shown by a colleague via loss of signal upon transcription inhibition with triptolide (Hadzhiev et al. 2019). Foci were predicted to originate from the two chromosomal alleles with the timing of signal detection correlating with what has previously been shown for temporal dynamics of miR-430 activity by both CAGE-seq and RNA-seq (**Figure 3-9**).

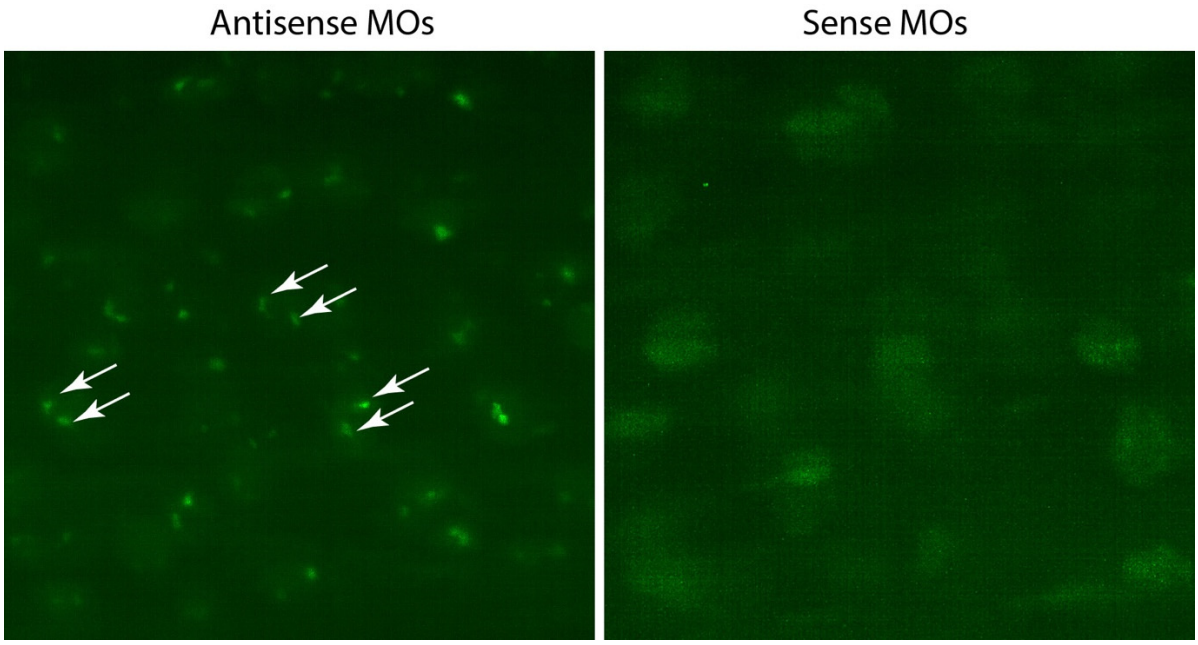


***Figure* 3-8 miR-430 MOs specifically bind to miR-430 transcript**

Detection of two miR-430 containing foci in live embryos at high/oblong stage, via binding to the miR-430 transcript with 3x array of antisense green fluorescent morpholino oligonucleotides (MOs). Lack of signal in Sense (DNA binding) control MOs demonstrates specificity of experimental (antisense) MOs to miR-430 RNA. Signal detected by lightsheet imaging.

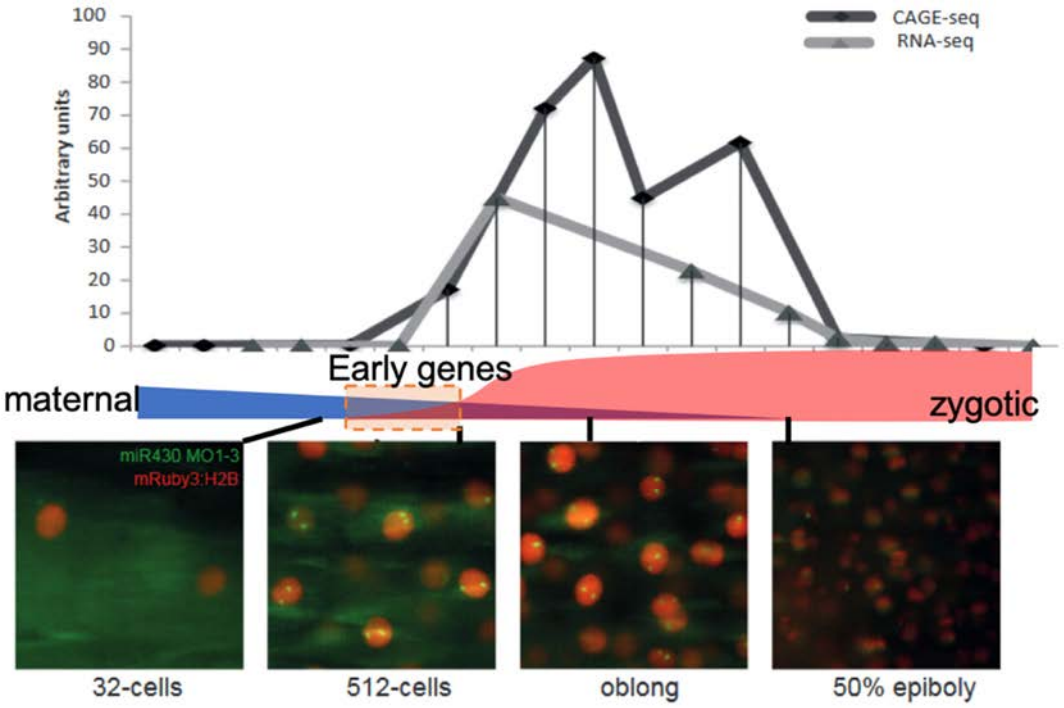


Figure 3-9 miR-430 expression across the maternal to zygotic transition

Expression of miR430 during development detected by genome wide transcriptome assays (multimapped WCE CAGE-seq and RNA-seq data (White et al. 2017)) alongside the corresponding detection of transcripts by miR-430 targeting morpholinos in vivo. Green signal is representative of miR-430 expression and nuclei are marked by H2B mRuby fusion, shown in red. Images taken by Yavor Hadzhiev (Hadzhiev et al. 2019).

3.5 Discussion

In order to begin to elucidate the mechanisms of promoter-mediated regulation of transcription initiation, we developed a model (Hadzhiev et al. 2019) for transcription detection utilising the most highly transcribed miRNA family in the developing zebrafish early embryo: miR-430 (Chen et al. 2005, Giraldez et al. 2005). Responsible for directing degradation of maternal transcripts, miR-430 activity is fundamental for zygotic genome activation to occur (Giraldez et al. 2006).

In this chapter I have characterised the miR-430 promoter, in regard to promoter architecture, and provided evidence, complementary to others (Mathavan et al. 2005, Heyn et al. 2014), that miR-430 is expressed before the MBT, when the zygotic genome becomes active globally. This was achieved through utilisation of a multimapping approach for various CAGE-seq datasets which revealed presence of 8 predicted promoters on sequentially repeating miR-430 gene units. Further analysis of the underlying promoter sequence has revealed a classical promoter structure consisting of both a TATA-box and canonical ‘C-A’ Inr motif.

One may surmise that early active repeats may be necessary simply to promote accumulation of transcripts in a short time frame, where there is a restriction on the amount of time for transcription due to short cell cycles. Such a structural feature could be integral for accumulation of factors fundamental for progression of development past the MBT, indeed it is known that genome activation in

humans and mice involves multicopy transcribed repeats (Macfarlan et al. 2012, Grow et al. 2015). This hypothesis is supported further by the fact that the miR-430 multicopy nature is conserved across species (Watanabe et al. 2005, Lund et al. 2009, Tani et al. 2010) evidence that the repetitive structure is of functional importance.

Here I have demonstrated, through design of a tandem array of three fluorescent morpholino oligonucleotides specific for the miR-430 model gene, visualisation of native transcription from the endogenous locus. Achieved by utilising the multi-copy nature of the gene cluster for signal amplification, without the need for transgenic manipulation.

Through utilisation of morpholino oligonucleotide-based visualisation, I found that transcripts of miR-430 accumulate into two distinct foci. Which later was shown to be associated with accumulation of active Pol II, together forming a transcription body with size similar to the nucleolus, yet representing a distinct subnuclear compartment (Hadzhiev et al. 2019). One possible role for such a transcription body could be in formation of an environment enabling transcriptional activation in an otherwise transcriptionally quiescent embryo via isolation of specific genes from repressive factors.

Recent *in vivo* work in mammalian cell models using transcription imaging technologies has revealed formation of sizeable transcription hubs, indicated via accumulation of Pol II Ser2P into distinct foci (Cisse et al. 2013, Conic et al. 2018)

and specifically in the context of genome activation, such pairs of Pol II Ser2P-containing foci are detected in the nucleus as sites for transcription in Drosophila during the MBT (Chen et al. 2013, Blythe et al. 2015, Hug et al. 2017), similar to what we have observed in the zebrafish embryo.

Interestingly, large transcription compartments are also formed by viral DNA. Viral replication compartments formed during infection by Herpes Simplex virus were found to sequester host Pol II into membraneless compartments reminiscent of transcription hubs, believed to be facilitated by the increased accessibility of the viral genome to basal transcription machinery conferred by low nucleosome occupancy (McSwiggen et al. 2019). Overall, these observations suggest that active genes are generally organised into distinct nuclear topologies.

4 CHARACTERISATION OF THE CHROMATIN STATE AND MODIFICATION DYNAMICS OF THE MIR-430 CLUSTER

4.1 Temporal dynamics of chromatin states and transcription

The main wave of ZGA is characterized by activation of thousands of genes as a result of reaching a threshold nucleocytoplasmic ratio, representing competition between maternally deposited repressing chromatin factors, which get diluted by cell divisions and a rate limiting availability of transcription machinery (Amodeo et al. 2015, Joseph et al. 2017). However, this is just one of the several mechanisms which influence both the timing and extent of genome activation. These mechanisms, are likely parallel acting and include: (i) limitation of availability of the transcription machinery, such as general transcription factors, which regulate transcription initiation at the core promoter and are gradually produced from maternal mRNAs (for example TBP (Veenstra et al. 1999)), (ii) delayed availability of transcription factors (Lee et al. 2013), (iii) sequential increase in chromatin accessibility (Blythe et al. 2016, Pálfy et al. 2020), (iv) and that of transcriptionally competent epigenetically marked chromatin (Lindeman et

al. 2011, Chan et al. 2019), the establishment of some of which (e.g. H3K4me3 at promoters) is transcription independent (Haberle et al. 2014) and (v) elongation of the cell cycle, which regulates the appearance of G phase, permissive for elongating transcription (Collart et al. 2013, Blythe et al. 2015). These mechanisms (reviewed in Lee et al. (2014), Wragg et al. (2016), Pálfy et al. (2017)) together may provide permissive conditions for genes at the main wave of genome activation.

However, miR-430 among several other developmentally essential genes escape the global repression (Heyn et al. 2014, Abe et al. 2018) and get activated well before the rest of the genome. It is currently unknown what mechanisms are involved in activation of the first wave genes, including miR-430, and therefore it is important to characterize their epigenomic features, which may inform on various genome activation mechanisms involved. We (Hadzhiev et al. 2019) found that (i) multicopy microRNA genes occupy a transcription compartment, (ii) do not respond to manipulation of the nucleocytoplasmic ratio (unpublished data by Qureshi et al., and Chan et al. (2019)). Others have shown that miR-430 activation is dependent on BRD4 and competent chromatin marked by H3K27ac (Chan et al. 2019), (iii) is marked by CTCF and Rad21 (Meier et al. 2018) and (iv) that its high transcriptional activity triggers displacement of the local chromatin, potentially by formation of the transcription body by phase separation (Hilbert et al. (2018) – preprint). As it forms an extremely large ‘transcription body’, that contains large amount of transcriptionally active PolII Ser2P and colocalises with RNAs from several active gene clusters (Hadzhiev et al. 2019), it represents an

accessible model to address the mechanisms of the activation of the very first genes in the embryo.

Evidence from our study (Hadzhiev et al. 2019), pertaining to the pair of miR-430 enriched transcription foci of considerable size (peaking at 16 μm^3), suggests a high level of transcriptional output from the gene cluster to account for the density of miR-430 transcripts present. Furthermore, Nanopore long read DNA sequencing carried out in our laboratory revealed several single reads with more than 40-50 miR-430 gene units, a finding that conflicts with the current assembled version GRCz11/danRer11 (8 complete gene units), suggesting that the gene cluster is larger than is currently assembled. The existence of approximately 100 miR-430 precursor copies distributed over two clusters: 30 kb and 17 kb respectively, on unassembled chromosomal regions was described previously by (Chen et al. 2005) on a legacy version of the zebrafish genome, Zv4. Possible shortcomings in the assembly of the long arm of chromosome 4, including the miR-430 locus, cannot be excluded, due to the highly repetitive nature of the genomic sequence that makes it extremely challenging to assemble (Howe et al. 2013, Howe et al. 2016).The results described led to surmise that due to the locus being represented in an incomplete state, gene units may have been omitted. Therefore, it was necessary to perform locus-specific reassembly of the miR-430 region using our long read Nanopore DNA sequencing data, followed by a more precise reannotation.

4.2 Aims

To characterise the epigenetic profile of the most highly expressed gene cluster at the first wave of genome activation, miR-430, in comparison to other gene sets to establish whether genes that are transcribed pre-ZGA follow the same chromatin state dynamics and transcription activation mechanisms.

4.3 Analysis of chromatin state on a revised miR-430 cluster

To analyse the chromatin state and modification dynamics of the miR-430 cluster a draft assembly, generated by a colleague in our laboratory from 3rd generation long-read sequencing data (PacBio and Nanopore) was used. The miR-430 cluster assembled from the long-read datasets appeared to be significantly larger than in the current reference genome (GRCz11/danRer11), spanning almost 250 kb and containing approximately 120 gene units (promoters linked to triplet structures), see **Figure 4-1**. As described in Chapter 3, structurally the miR-430 cluster is composed of 3 primary subtypes of precursor gene: ‘a’, ‘b’ and ‘c’, which are organised into triplets ordered as following: ‘a,c,b’. The most common organisation of a gene unit in the assembled miR-430 contig is two triplets downstream of a promoter, which will be referred to as a gene duplex. For comparison, in the current reference assembly the miR-430 cluster is roughly 16 kb in length and contains only 8 complete gene units (see **Figure 4-1** A, and **Figure 3-2**).

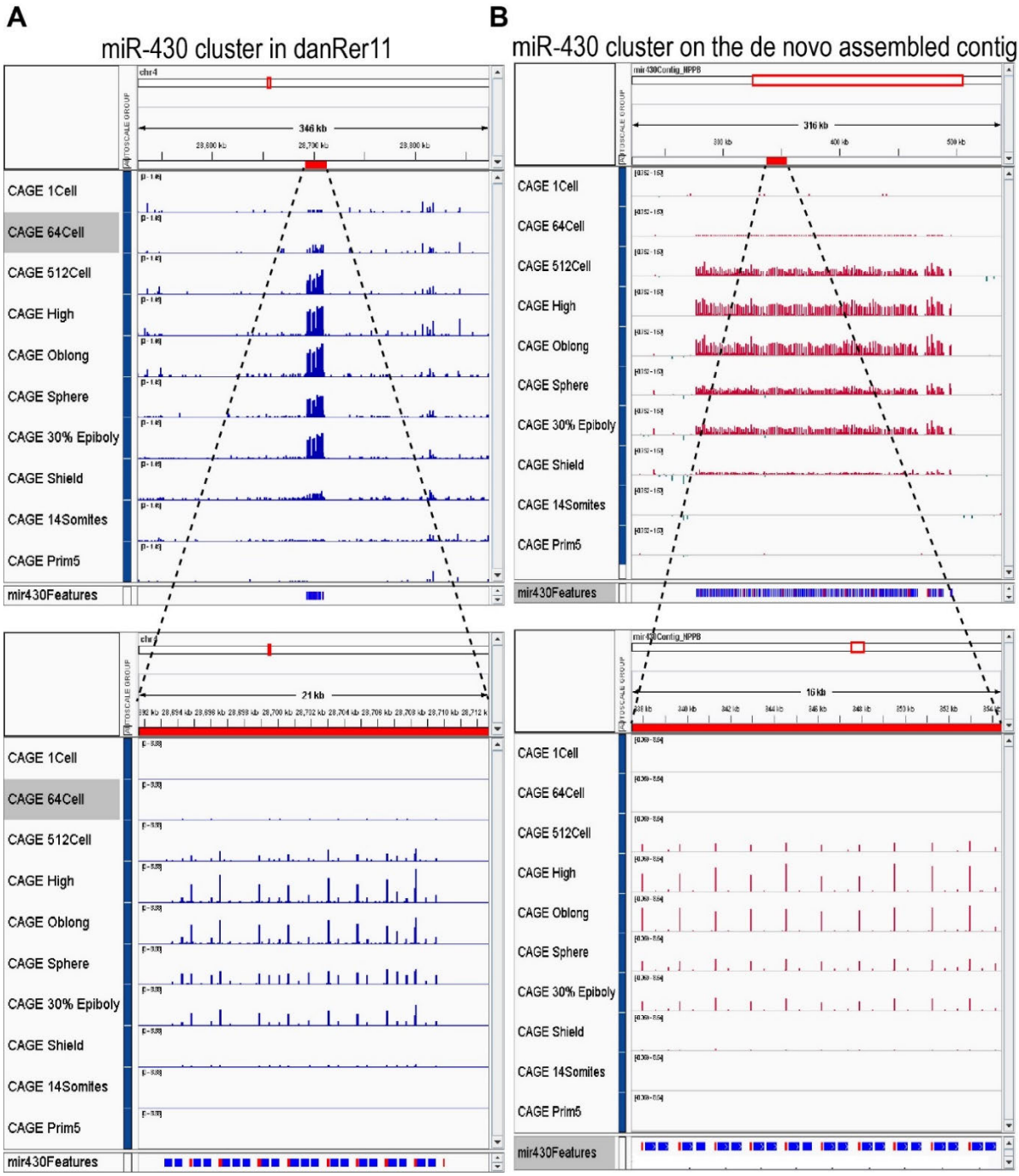


Figure 4-1 Schematic of miR-430 cluster assembled from long read DNA-seq

Comparison between miR-430 clusters on the reference genome (danRer11) (A) and the de novo assembled contig from PacBio and Nanopore reads (B). Top panels show an overview of ~300 kb region with an enlargement of a 15-20 kb on the bottom panels as indicated by dashed lines. CAGE-seq signal for several developmental stages marks

the TSS (the highest peak) and shows expression dynamics. Annotations shown represent identified promoter (red bar) and triplet (blue bar) structures.

Publicly available epigenomic datasets (ATAC-seq, H3K4me3-, H3K27me3- and H3K27ac-ChIP-seq), covering as many as possible developmental stages of interest per dataset were selected for analysis (see appendix 8.1).

The selected datasets were mapped to a custom genome, created by appending the miR-430 cluster containing contig from the long-read assembly to the current zebrafish genome reference (GRCz11/danRer11) sequence file. The mapping was performed using the short-read aligner segemehl (Hoffmann et al. 2009, Otto et al. 2014), reporting up to 2000 valid (minimum 90% accuracy) alignments per read (see methods for all parameters). The segemehl aligner was used as the best trade-off between accuracy, speed, and ability to find and report multiple alignments per read as the later was crucial for the analyses, due to the tandem repeat structure of the miR-430 cluster. Of all the aligners compared, STAR (Dobin et al. 2013) was the fastest, but was not sensitive enough in mapping reads to multiple locations. Bowtie1/Bowtie2 (Langmead et al. 2009, Langmead et al. 2012) performed well in finding multiple alignments per read but was very slow when instructed to detect and report more than 200-300 alignments. The segemehl aligner on the other hand was less sensitive in finding multiple alignments than bowtie2 but proved to be more accurate and most importantly significantly faster (3-4x) than bowtie2, when requested to look and report high numbers (>1000) of valid alignments per read.

Genome coverage normalised to library size (reads per million, rpm) was calculated for uniquely mapped reads only and both uniquely and multi-mapped reads combined using the STAR aligner by providing the alignment (bam) files as input.

To investigate the role of epigenomic features in the early genome activation, their signal levels and dynamics during early development (blastula stages) were compared among 5 distinct gene sets listed below. These gene sets were selected based on their onset of activation, with the aim to be able to identify distinguishing epigenomic features characteristic to a particular wave of genome activation.

1. **miR-430** genes: The earliest and most highly transcribed gene cluster during the first (minor) wave of zygotic genome activation. 80 *mir-430* duplex genes (containing a promoter and two triplets, which is the most common gene structure) from the *de novo* assembled miR-430 cluster were selected.
2. **First Wave** genes: reported to be activated during the first (minor) wave of the zygotic genome activation (Heyn et al. 2014). Only zygotically expressed (no maternal contribution) were selected. It is important to noted that the pre-MBT expression for most of these genes are solely based on the data (Ethynyl Uridine incorporation RNA-seq) published in the above manuscript (see **Appendix 8.4**). In many cases the pre-MBT expression could not be detected at significant levels by conventional RNA-seq and CAGE-seq (on **Figure 4-2**, A). This was the gene number

limiting set, i.e., the maximum number of genes which matched the selection criteria was 80. To have the same gene number in every set, 80 genes which matched the selection criteria for the dataset were chosen at random.

3. **Main Wave** genes: activated at the main (MBT) wave of the zygotic genome activation and have not been reported to be transcribed during the first wave by (Heyn et al. 2014), 80 exclusively zygotic genes were selected (see **Appendix 8.5** for full gene list).
4. **Constitutive** genes: mainly housekeeping genes having maternal contribution and expressed throughout development. These genes use two different promoter codes present on the same core promoter platform. The first TATA-like motif called W-box is utilised in the oocyte which later is replaced by a GC boundary downstream of the TSS at zygotic genome activation (Haberle et al. 2014). 80 genes which showed relatively constant expression levels throughout development (form fertilised egg to Prim 20) were selected, shown in **Appendix 8.6**.
5. **Post Gastrulation** genes: activated after gastrulation during early segmentation stages, 80 zygotic only genes were selected (as shown in **Appendix 8.7**).

Expression dynamics for all gene sets were determined based on CAGE-seq dataset (Nepal et al. 2013) mapped to GRCz11/danRer11, ENSEMBL v95 gene annotations and is shown on **Figure 4-2** and **Figure 4-3**.

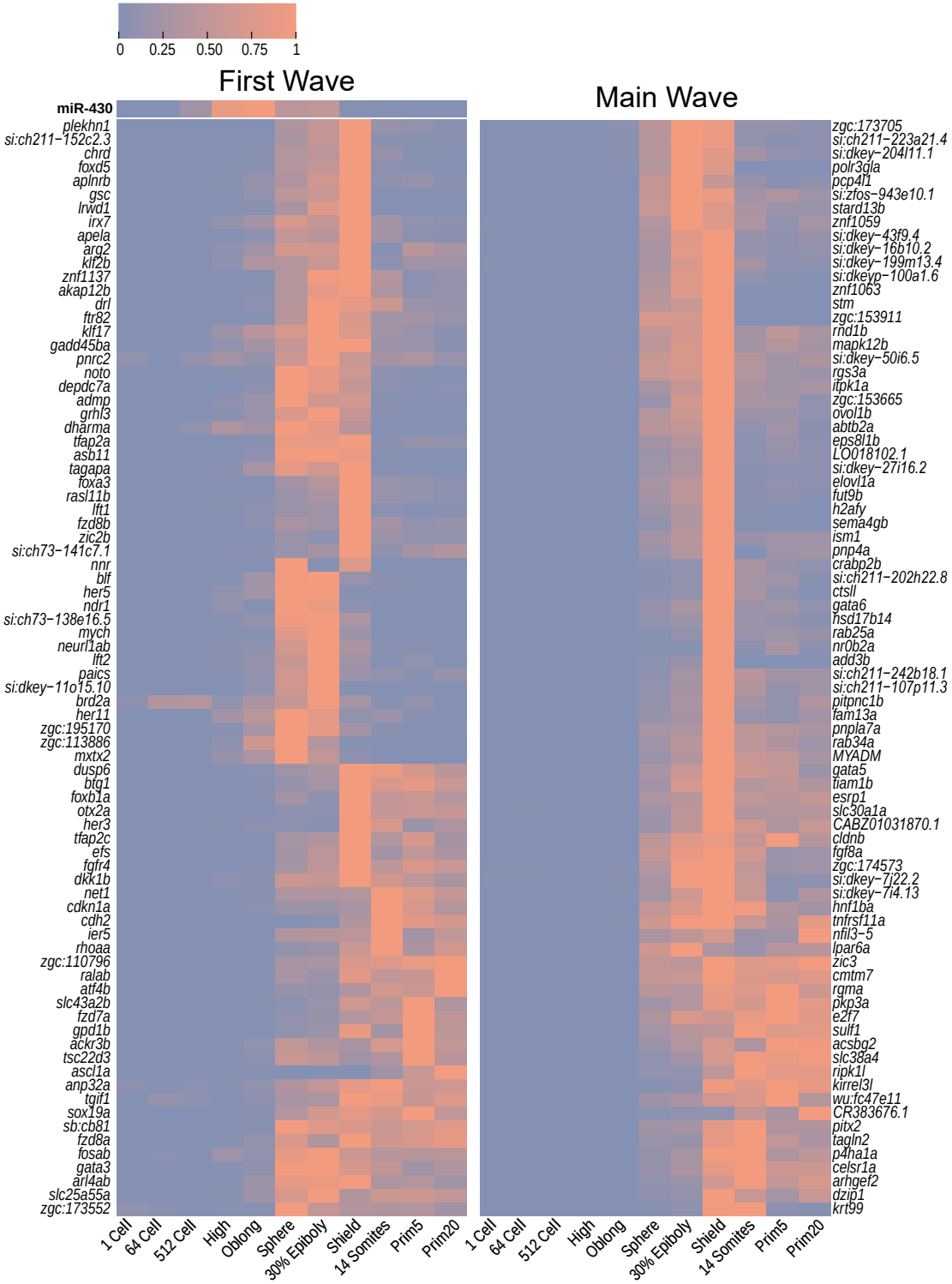


Figure 4-2 Expression dynamics for first wave and main wave gene sets

Dynamics of expression of genes within contrasting gene sets, as determined from CAGE-seq data. Colour keys indicate expression levels with low / no expression to high expression represented by a colour shift from purple to red respectively.

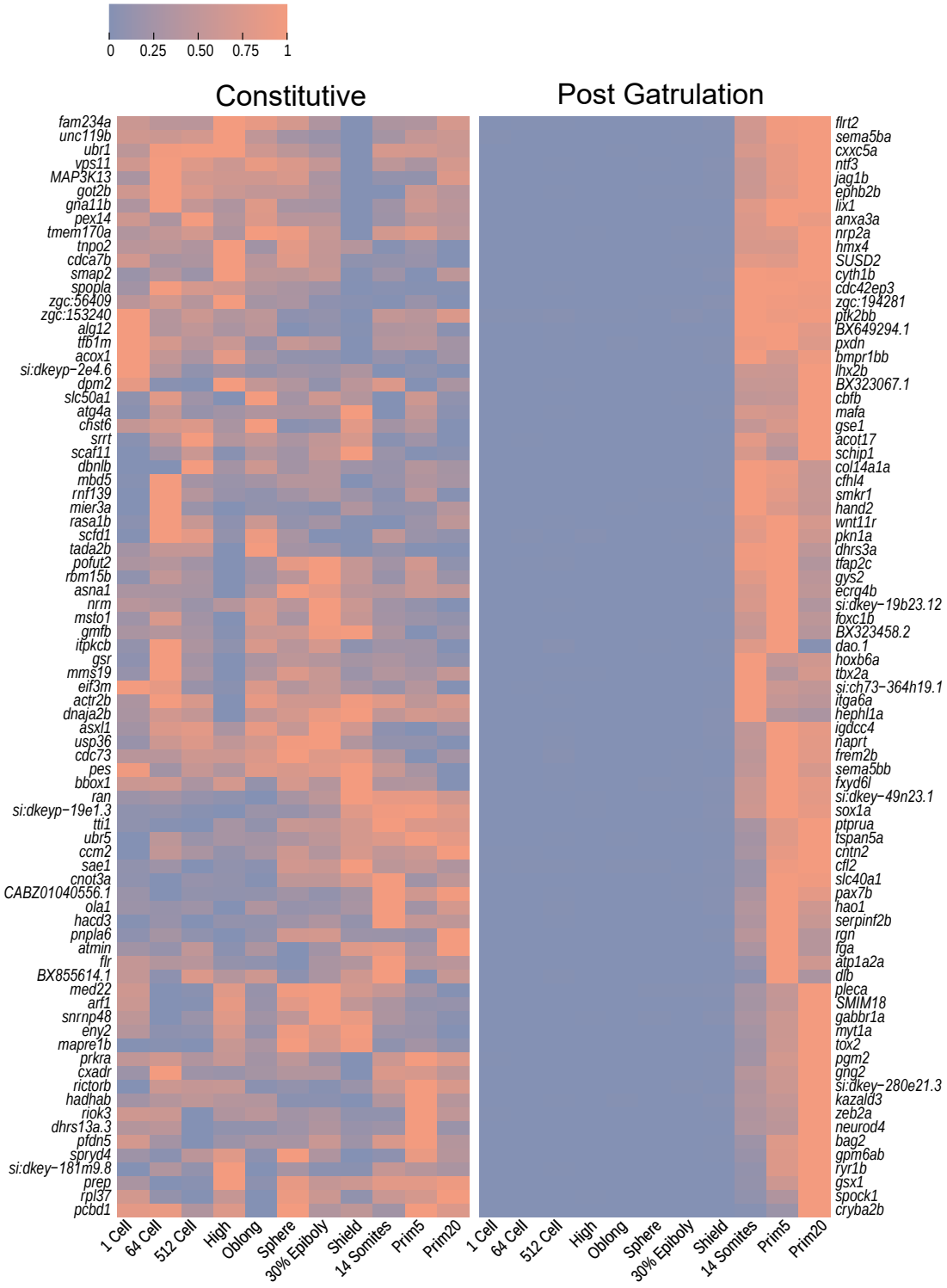


Figure 4-3 Expression dynamics for constitutive and post gastrulation gene sets

Dynamics of expression of genes within contrasting gene sets, as determined from CAGE-seq data. Colour keys indicate expression levels with low / no expression to high expression represented by a colour shift from purple to red respectively.

Genome browser screenshots summarising the signal coverage of all epigenetic features for a representative example gene from each set is shown in **Figure 4-4 - Figure 4-8**.

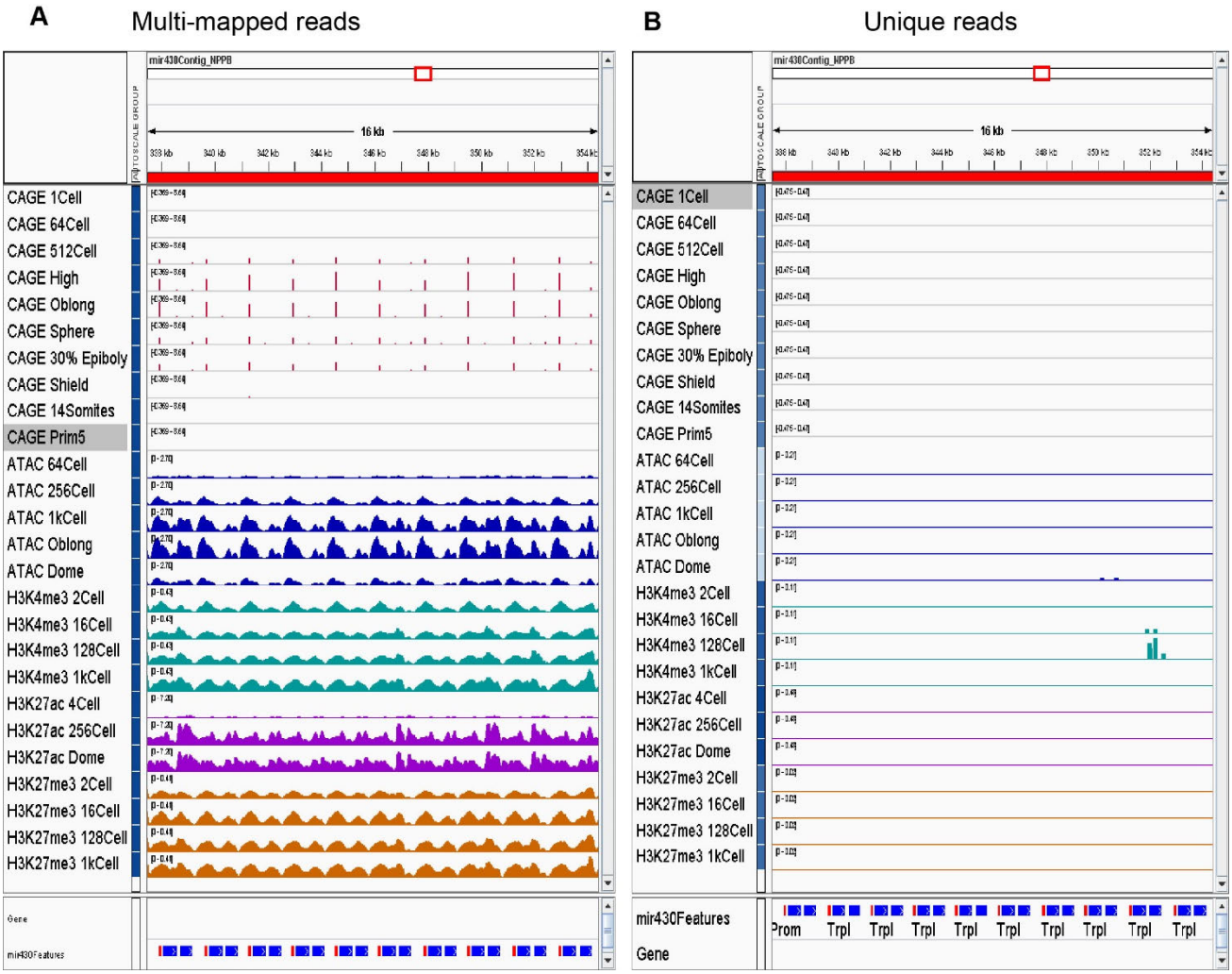


Figure 4-4 Summary of dynamics of epigenomic data at the miR-430 assembly

Summarised representation of the epigenetic datasets used in this analysis as coverage tracks on part of the long read draft assembly for the miR-430 cluster. CAGE tracks (top) mark the TSS and expression dynamics of the miR-430 genes. ***A*** *multi-mapped reads,* ***B*** *Unique mapped reads.*

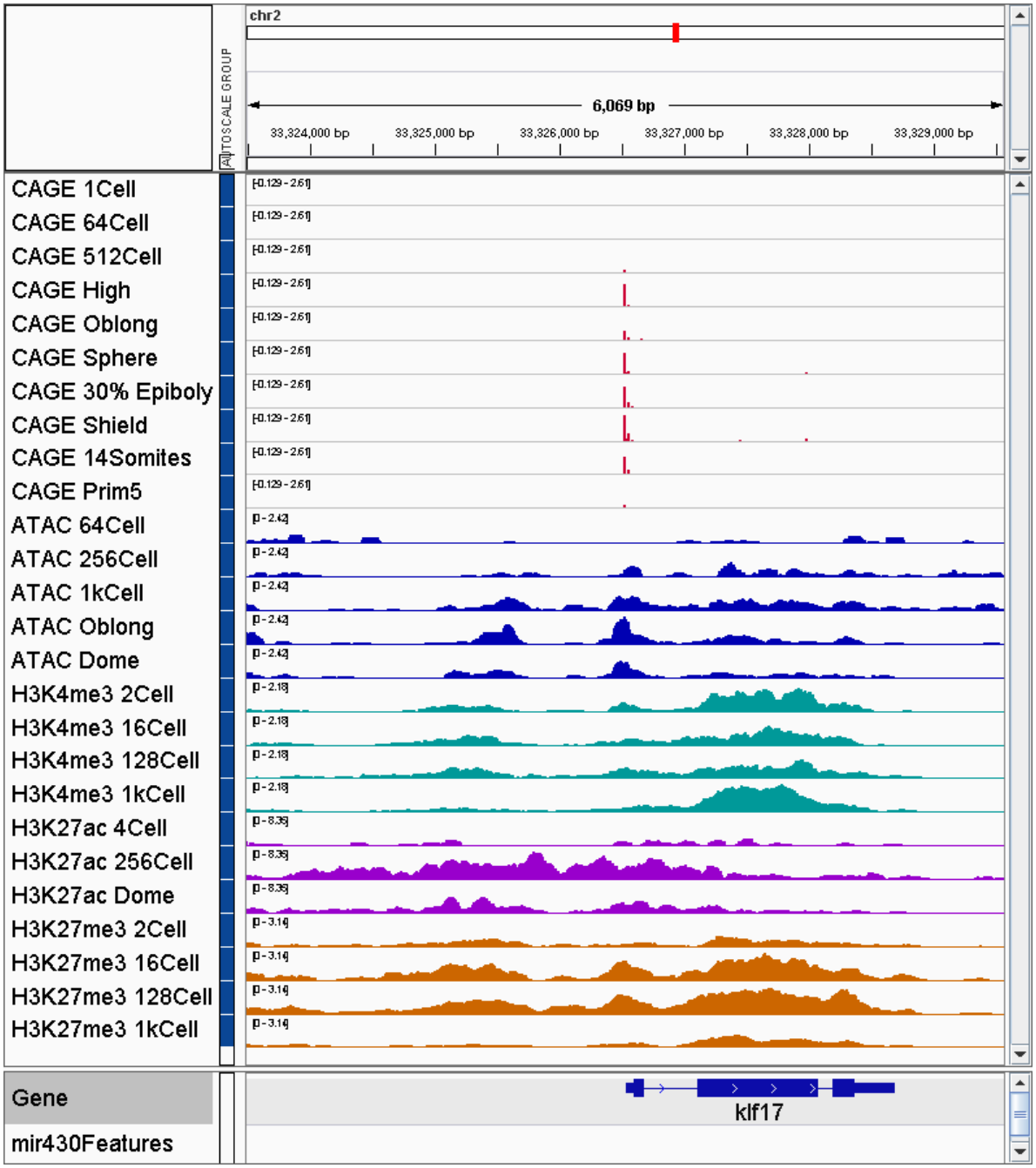


Figure 4-5 Summary of dynamics of epigenomic data on example first wave gene

Summarised representation of the epigenetic datasets used in this analysis as coverage tracks klf17 a representative gene from the first wave gene set. CAGE tracks (top) mark the TSS and expression dynamics of the first wave gene example.

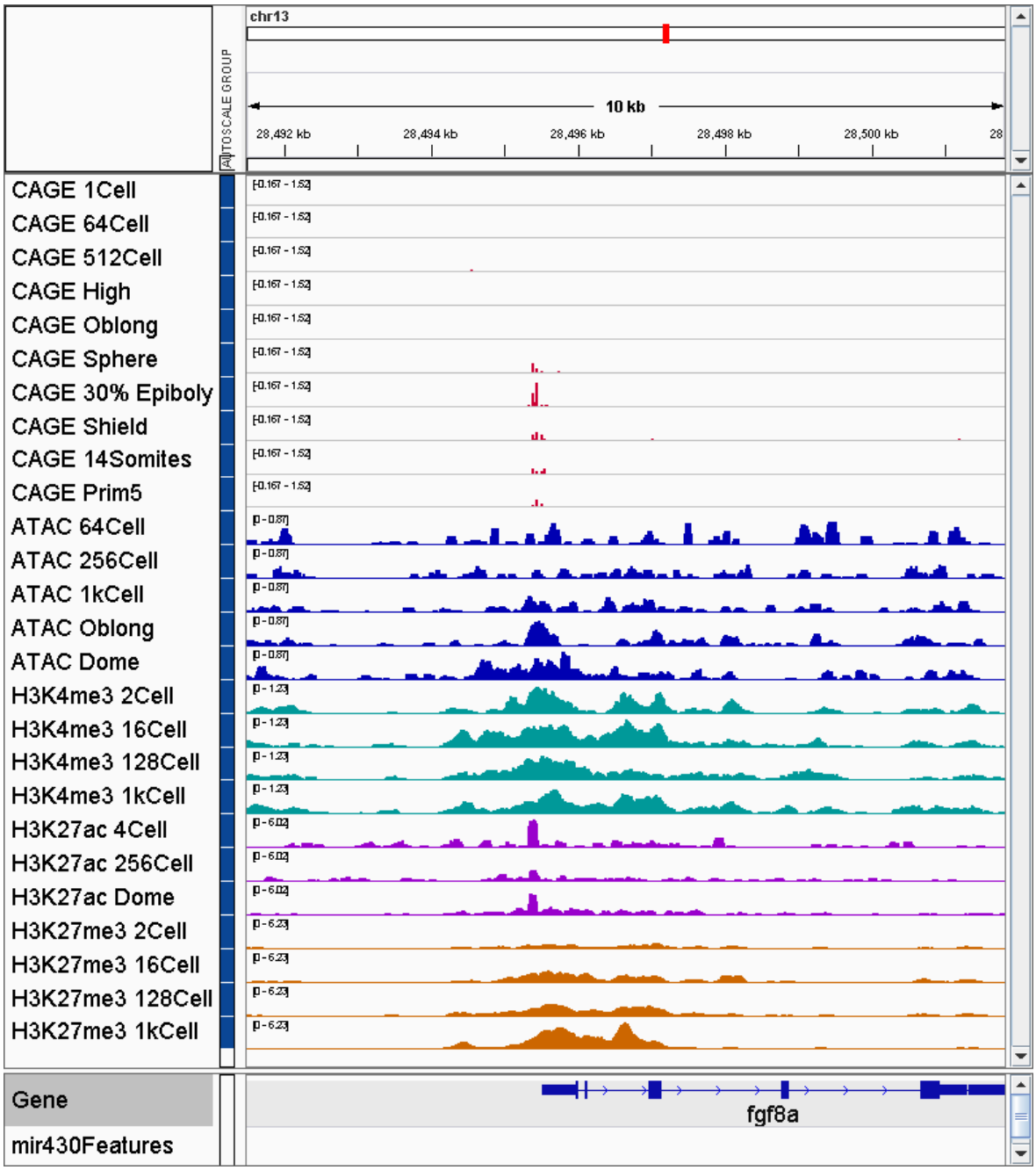


Figure 4-6 Summary of dynamics of epigenomic data on example main wave gene

Summarised representation of the epigenetic datasets used in this analysis as coverage tracks fgf8a a representative gene from the main wave gene set. CAGE tracks (top) mark the TSS and expression dynamics of the main wave gene example.

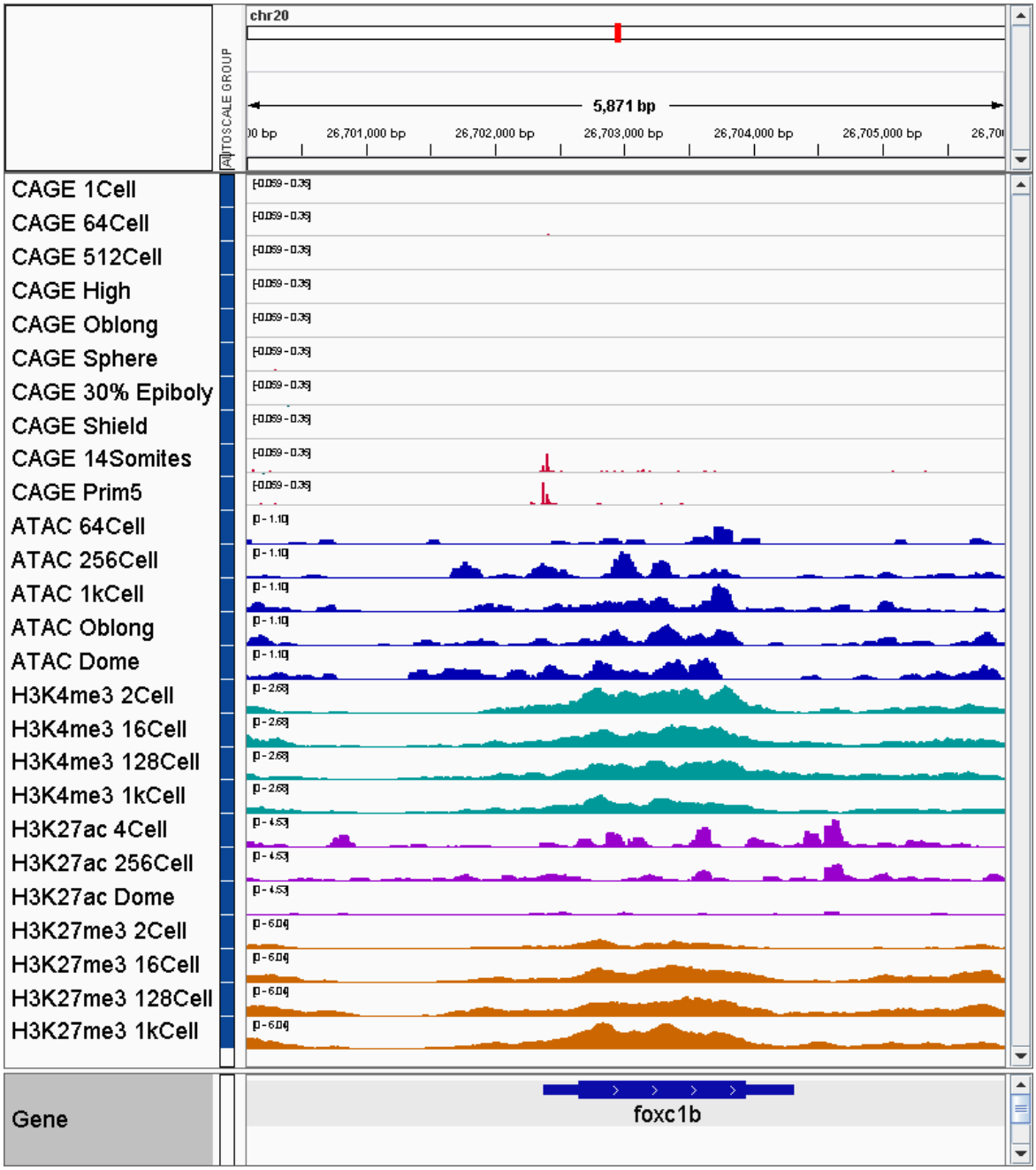


Figure 4-7 Summary of dynamics of epigenomic data on example post gastrulation gene

Summarised representation of the epigenetic datasets used in this analysis as coverage tracks foxc1b a representative gene from the post gastrulation gene set. CAGE tracks (top) mark the TSS and expression dynamics of the post gastrulation gene example.

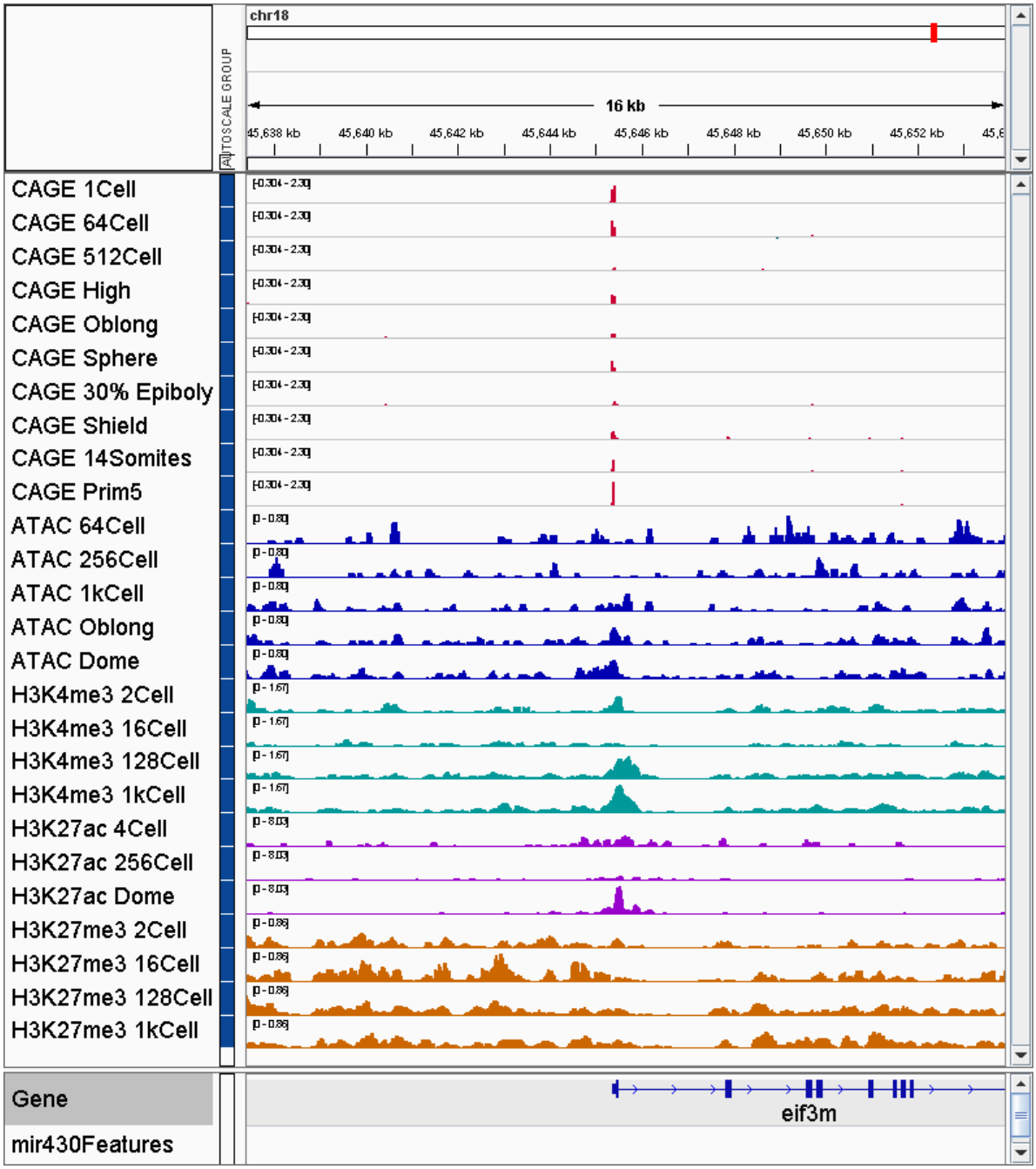


Figure 4-8 Summary of dynamics of epigenomic data on example constitutive gene

Summarised representation of the epigenetic datasets used in this analysis as coverage tracks eif3m a representative gene from the constitutive gene set. CAGE tracks (top) mark the TSS and expression dynamics of the constitutive gene example.

To compare the dynamics of the epigenetic features between the above datasets, the distribution of average normalised signal relative to the transcription start site for each set and feature was visualised on aggregation plots (**Figure 4-9**-**Figure 4-13**).

The transcription start site was determined using the above-mentioned CAGE-seq dataset at a developmental stage where there was a sufficiently high signal to accurately determine the dominant TSS for each gene in the set or to minimize interference from the maternal contribution in the case of the constitutive gene set. For the miR-430, first and main wave gene sets high, sphere and 30% epiboly were used, respectively. For both the constitutive and post-gastrulation sets prim5 stage was used.

The plots were generated using the SeqPlots Bioconductor R-package (see methods). The signal for plotting of the early and late zygotic gene sets was calculated from the coverage of uniquely mapped reads only, whereas for mir-430 genes, coverage from both uniquely and multi-mapped reads were used to determine overall signal. A reduced window size was used for analysis of miR-430 for all datasets, limiting the upstream sequence. Necessary due to the repetitive nature of the gene cluster to prevent inclusion of sequence from flanking triplets.

The state (open/closed) of chromatin at the cluster was examined from the ATAC-seq datasets. The distribution profile of the signal was as expected for all datasets

peaking at around 100-150bp upstream of the TSS where the nucleosome free region of active promoter is. The signal levels on the miR-430 genes, as well as on the early and late zygotic gene sets, correlated with the gene expression dynamics, i.e. high levels of ATAC-signal, indicating more open chromatin when genes are expressed and low when weakly or not expressed (**Figure 4-9**). Profiles for open chromatin observed in all gene sets overall correlated with expected onset of activation i.e. miR-430 is very active early and shows the most open chromatin (**Figure 4-9** A,F), whereas the first wave genes were generally more open and become open at earlier developmental stages than the main wave genes (**Figure 4-9** B,C,F). The main wave genes were comparable to the constitutive gene set (**Figure 4-9** C,D, F) in line with their onset of zygotic expression at and post-MBT (**Figure 4-9** C,D, F). The most closed profile was associated with the post gastrulation set which are not expressed at the stages shown (**Figure 4-9** E, F).

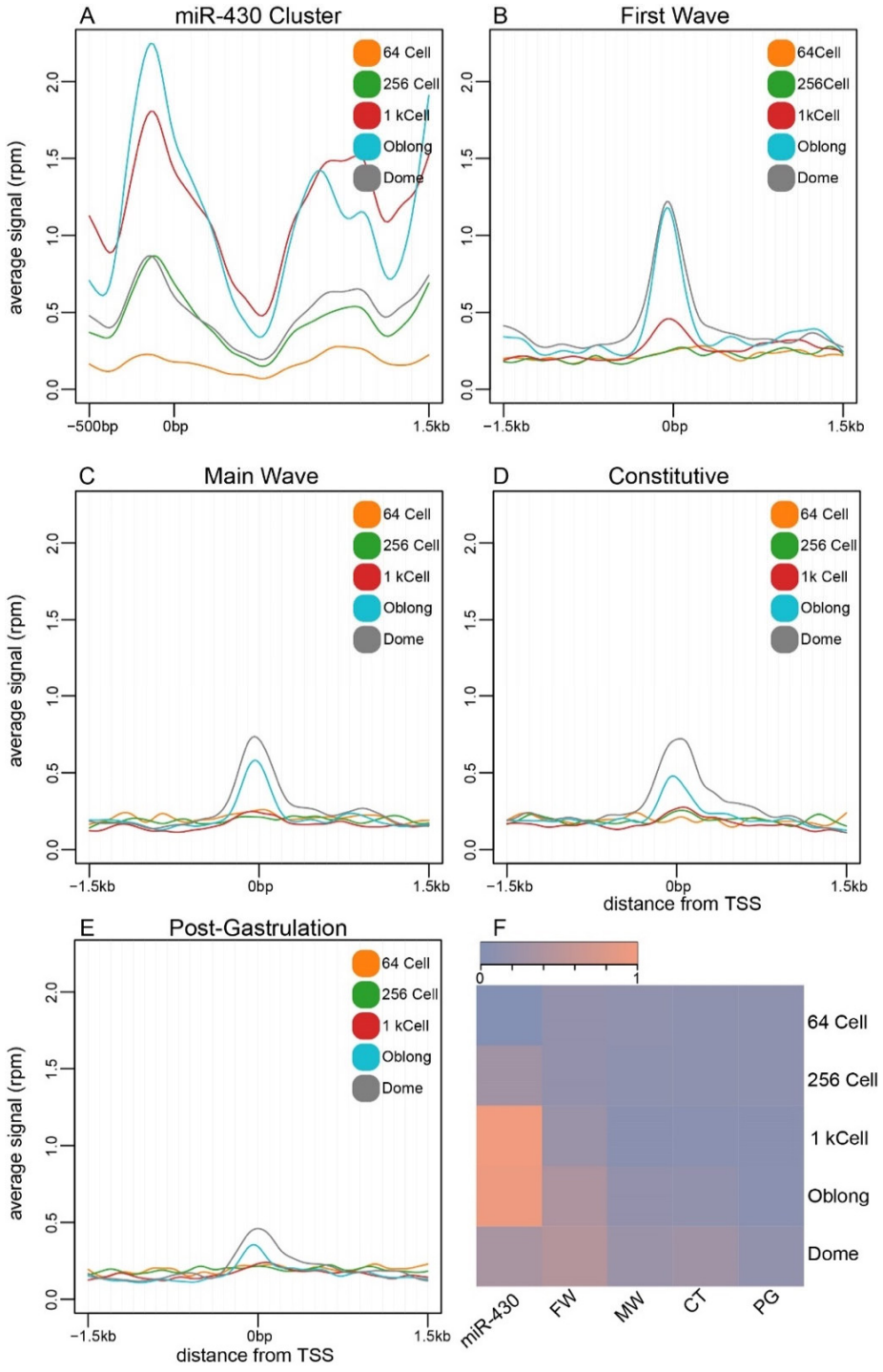


Figure 4-9 ATAC-seq dynamics and distribution relative to TSS

Plots represent aggregated signal levels (mean, y axis) relative to the TSS (x axis) for a given gene set and data set. For visualisation purposes a smoothing spline function has been applied to all curves in addition to light grey vertical guidelines placed at 100 bp intervals on each plot (A-E), (A) miR-430 has a reduced window size due to lack of upstream sequence and an additional downstream peak due to the repetitive nature of

the gene cluster. (F) Heatmaps represent area under curve for each feature of the plots (A-E) and was scaled from 0-1, for visualisation purposes

The H3K27ac, a mark for actively transcribed regions corelated well with the open chromatin state. The miR-430 showing again the highest signal, already at 256-Cell stage (**Figure 4-10** A,F). The first wave gene set also showed higher levels and broader signal distribution than the main wave (**Figure 4-10** B,C,F), which in contrast to the first wave set were more comparable to the constitutive genes (**Figure 4-10** C,D,F). Similar to the ATAC-seq profile, the lowest levels were observed in the post gastrulation set, in line being inactive at the stages shown (**Figure 4-10** E,F).

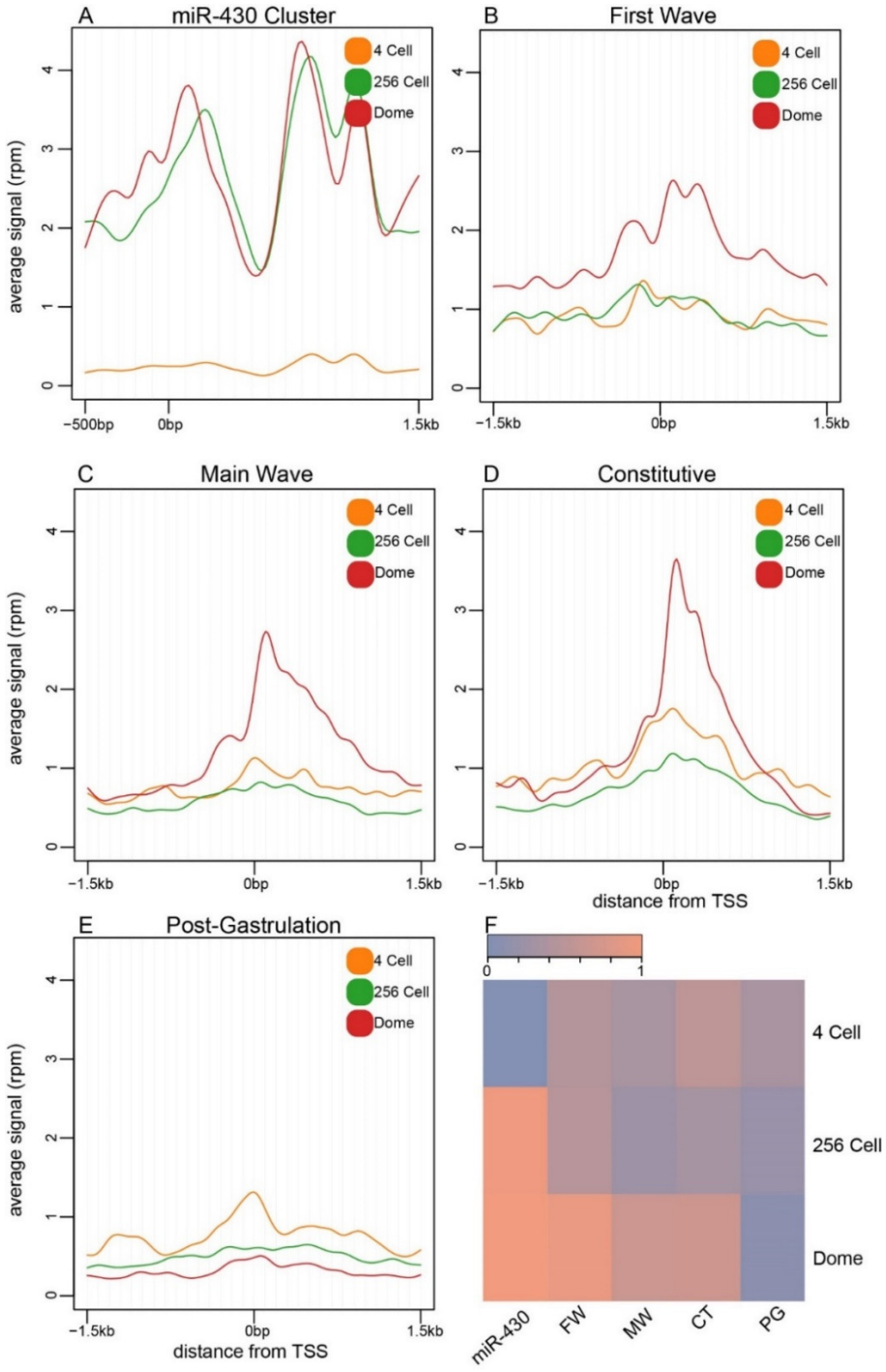


Figure 4-10 H3K27ac dynamics and distribution relative to TSS

Plots represent aggregated signal levels (mean, y axis) relative to the TSS (x axis) for a given gene set and data set. For visualisation purposes a smoothing spline function has been applied to all curves in addition to light grey vertical guidelines placed at 100 bp intervals on each plot (A-E). (F) heatmaps represent area under curve for each feature of the plots (A-E) and was scaled from 0-1, for visualisation purposes

Despite the high degree of open chromatin and enrichment in H3K27ac, corresponding to the high expression levels at early blastula stages, the miR-430 cluster showed significantly lower levels of the active promoter mark H3K4me3 in comparison to all other gene sets. Although when signal levels were compared internally, between stages, the signal levels correlate with miR-430 activity (**Figure 4-11** A,F). The low levels of H3K4me3 could be due to a lack of CpG island in the miR-430 promoter, as it is known that the deposition of H3K4me3 follows the extent of the promoter associated CpG island, playing major role in blocking DNA methylation (reviewed in Rose et al. (2014)) and might not be required for its activation.

The highest levels H3K4me3 were observed in the constitutive and first wave gene sets, with the latter also having a broader distribution of signal (**Figure 4-11** B,D,F) in line with these genes being active/activated at these stages which also aligns with observation that these genes sets (the first wave set in particular) have the highest proportion of genes with a promoter associated CpG island as well as longer CpG islands on average (**Figure 4-12**)

The main wave showed lower signal, which was expected, as their expression is initiated at slightly later stages (sphere/dome). Although not expressed at early blastula and getting active at much later (early segmentation) stages, the post gastrulation genes showed relatively high H3K4me3 signal, comparable to the main wave genes (**Figure 4-11** C,E,F). One explanation for this could that they

have a certain level of bivalency, where both the H3K4me3 activation and the H3K27me3 repressive marks are present on the same nucleosomes, keeping the genes in a “poised” state, yet pre-marked to get activated at later stage. This is to be expected for this gene set as it is enriched for developmentally regulated genes known to be the main targets of the polycomb repression complex (Loubiere et al. 2019).

An interesting dynamic of H3K4me3 signal also was observed between developmental stages. Relatively high levels were observed at 2-cell stages with a significant drop at 16-cell, most prominent in the constitutive and main wave gene sets and to degree at the miR-430 locus (**Figure 4-11** A,C,D,F). This dynamic indicates parental inheritance/pre-marking, that is followed by subsequent erasure of the mark at 16-cell stage with re-deposition at later stages, an observation which is in line with the trend reported in (Zhu et al. 2019) by whom the dataset was generated. This “reprogramming” was not observed in the first and post gastrulation gene sets. (**Figure 4-11** B,E,F). As mentioned above, like the first wave genes the post gastrulation is also enriched for developmentally regulated genes and as this class of genes are expected to be polycomb targets the deposition of active/repressive marks is likely regulated by a different mechanism. The distribution of the signal in the first wave and post gastrulation sets extends significantly more downstream into the gene body compared to both the main wave and in particular the constitutive gene set, where sharper peak around the first downstream positioned nucleosome was observed. This aligns with the fact that first wave genes have larger CpG islands (**Figure 4-12**), which

also suggests that the set is enriched for polycomb target genes, although the post gastrulation gene set do not have longer CpGs than other sets on average.

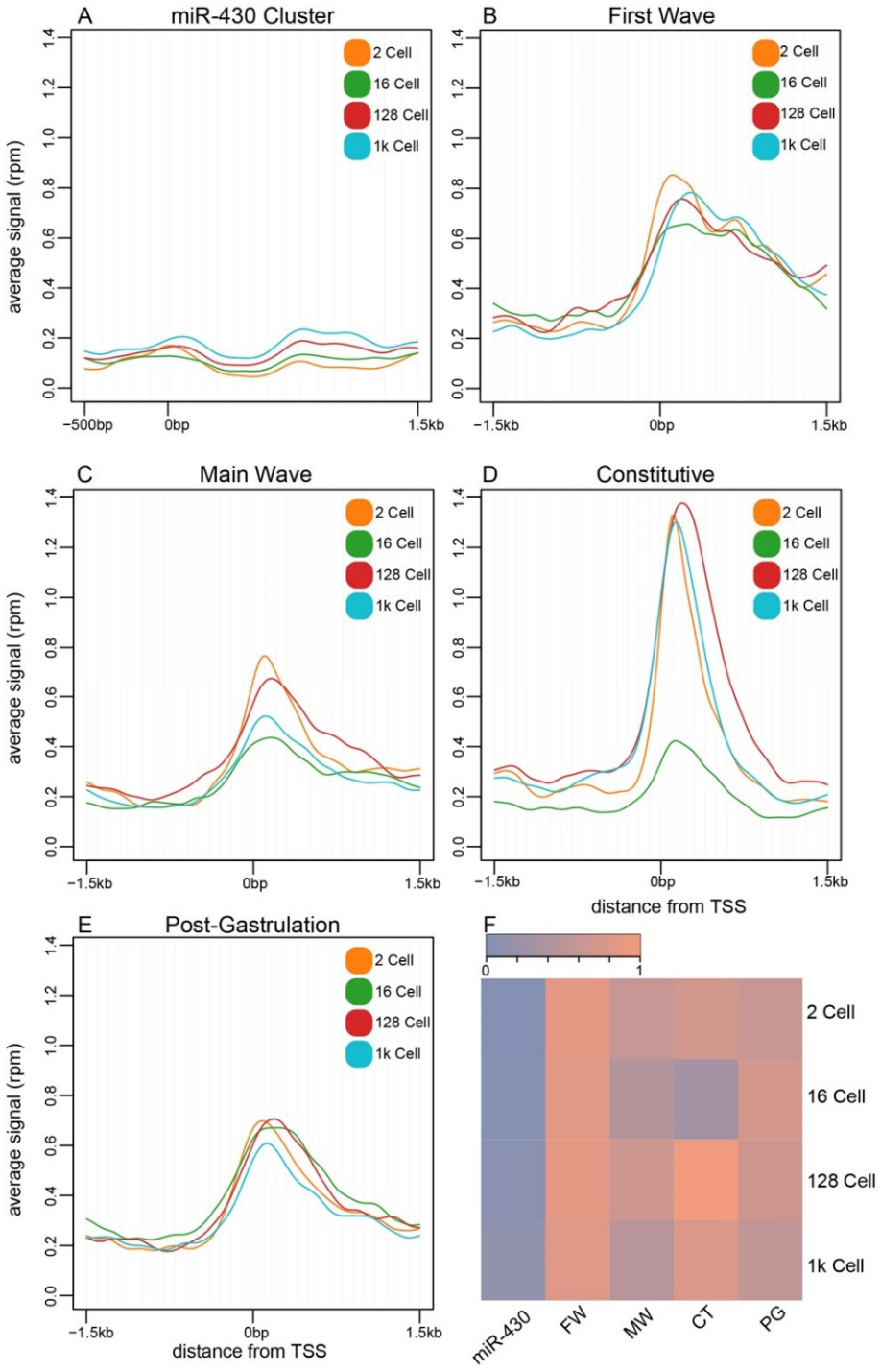


Figure 4-11 H3K4me3 dynamics and distribution relative to TSS

Plots represent aggregated signal levels (mean, y axis) relative to the TSS (x axis) for a given gene set and data set. For visualisation purposes a smoothing spline function has been applied to all curves in addition to light grey vertical guidelines placed at 100 bp intervals on each plot (A-E), (F) heatmaps represent area under curve for each feature of the plots (A-E) and was scaled from 0-1, for visualisation purposes.

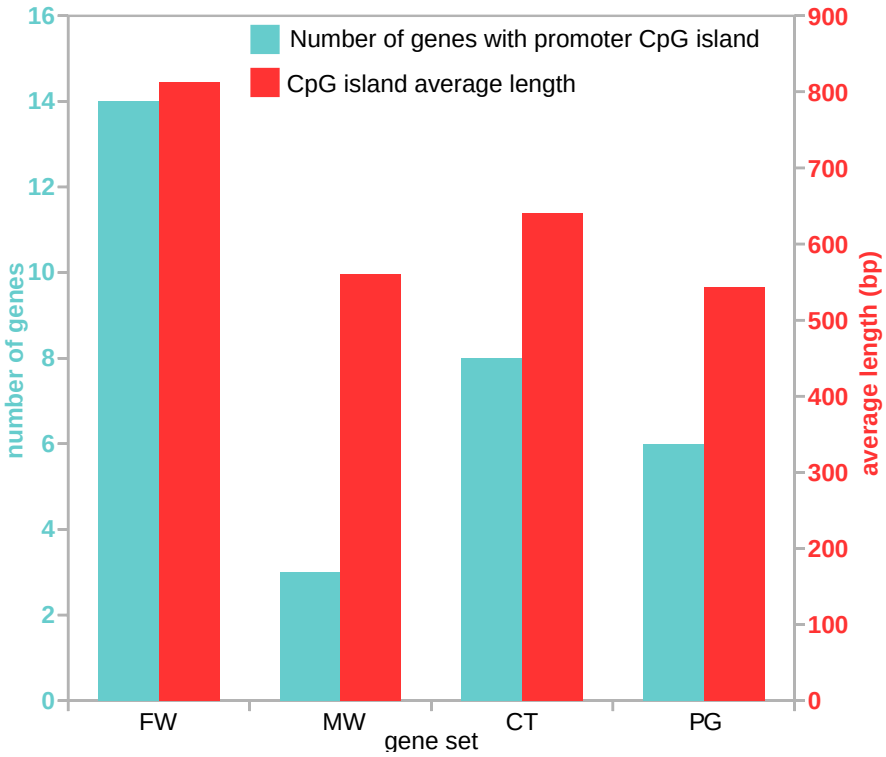


Figure 4-12 first wave and constitutive genes have a higher occurrence of promoter associated CpG islands in comparison to other gene sets.

For each gene set shown: first wave (FW), main wave (MW), constitutive (CT) and late/ post gastrulation (PG) both the number of promoter associated CpG islands is quantified (blue) and also the average length of such CpG islands (red).

The two data sets, the first wave and post gastrulation genes being polycomb targets is also suggested by the signal levels and dynamic of the repressive mark (**Figure 4-13** B,E,F). As shown on **Figure 4-13** (B,E,F) the two sets have the highest signal levels of H3K27me3. This is to be expected for the post gastrulation genes as they are not expressing at these stages, but unusual for the first wave genes as they are supposed to be actively transcribed during this time period. At the same time the first wave genes have high levels of H3K4me3 (**Figure 4-11** B,F) suggesting high degree of bivalency, without being in a

“poised” state. Nevertheless, at 1k cell stage they showed lowest levels of H3K27me3 **Figure 4-13** (B, F) and the highest of H3K4me3 (**Figure 4-11** B,F)

As expected the constitutive gene set, which comprises mostly of housekeeping genes expressed throughout development has the lowest levels of H3K27me3 (**Figure 4-13** D,F). The miR-430 locus has similar low levels (**Figure 4-13** A,F) suggesting that it is not a polycomb target and at later stages is repressed by a different mechanism. The main wave genes showed moderate levels of H3K27me3 with the expected dynamics.

A summarised heatmap showing the temporal dynamics of all analysed histone marks and from each dataset is shown in **Figure 4-14**.

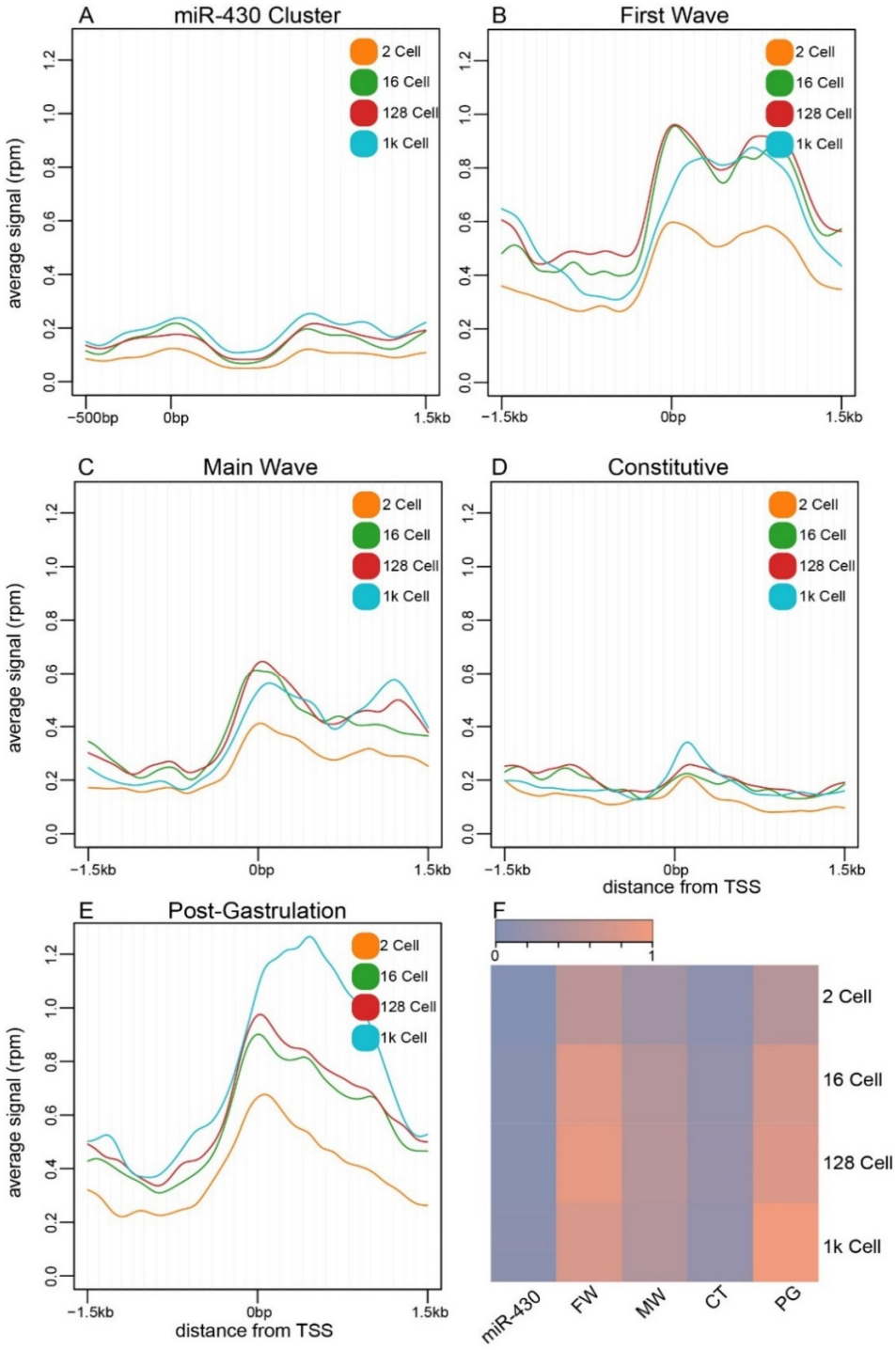


Figure 4-13 H3K27me3 dynamics and distribution relative to TSS

Plots represent aggregated signal levels (mean, y axis) relative to the TSS (x axis) for a given gene set and data set. For visualisation purposes a smoothing spline function has been applied to all curves in addition to light grey vertical guidelines placed at 100 bp intervals on each plot (A-E), (F) heatmaps represent area under curve for each feature of the plots (A-E) and was scaled from 0-1, for visualisation purposes

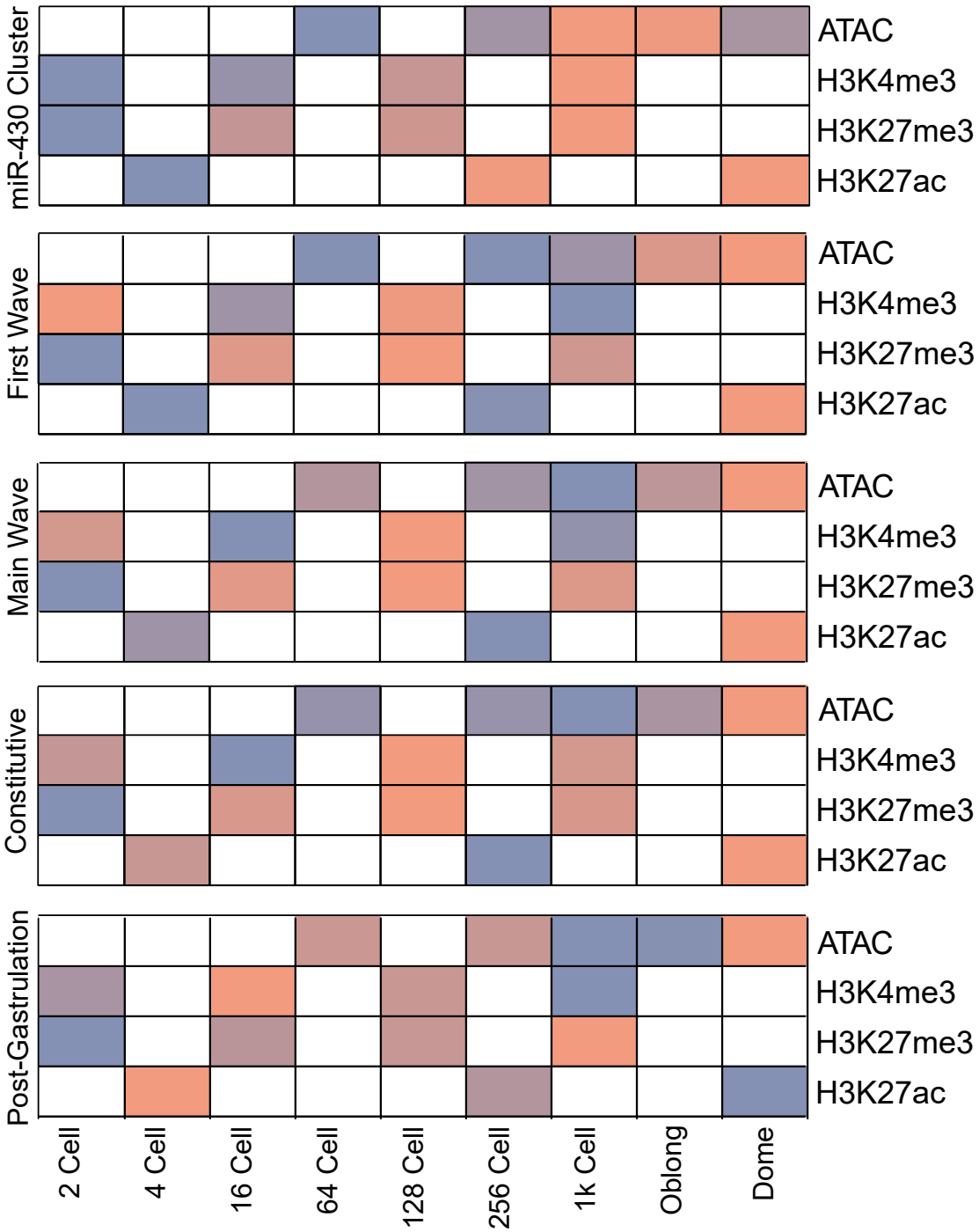


Figure 4-14 summary of epigenetic mark dynamics

The heatmap represents a summary of figures 3-6. For each stage (rows) and epigenetic mark (columns). Note that values for each epigenetic mark were scaled between 0 (purple) – 1 (red) within a dataset for visualisation purposes.

4.4 Discussion

Results within this chapter demonstrates first wave genes have high levels of both H3K4me3 and H3K27me3, suggesting potential bivalency, while also being suitably enriched in H3K27ac demonstrating active transcription within the region, although without single cell resolution it is not possible to say whether these marks (H3K4me3 and H3K27me3) are present in the same cell. The predicted bivalency is in line with the observation that the first wave (FW) gene set is enriched for developmental and tissue specific TFs, expression of which are restricted to specific cell types and tissues post-ZGA. Therefore, the presence of both marks for actively transcribed regions and repression may show that they are repressed in some cell types yet active in others, however the reason for their activity prior to MBT is still not known. Some FW TFs, like mxtx2 and klf17, are likely already needed for specification of one of the earliest defined cell types in the embryo, the extraembryonic yolk syncytial layer (YSL) and enveloping layer (EVL) respectively (Xu et al. 2012, Kotkamp et al. 2014). Despite this, many FW genes are reported to be required later at gastrulation stages, although it can be surmised that they may already play a role in specification of certain cell types, thus requiring their expression pre-ZGA.

In case of miR-430, multimapping was necessary to be able to visualize likely epigenetic profiles, but due the nature of this approach, it is not possible to verify whether the observed peaks are indeed present on each individual promoter and genic feature. Additionally, the multimapping results in arbitrarily enhanced peak heights, which blocks direct comparison of peak amplitudes to the other gene

sets used in the study. Nevertheless, temporal dynamics within gene sets can be compared and analysed in comparison to the miR-430 cluster. Unique mapping of the datasets in the revised genomic locus is similarly scarce as was presented in Chapter 1 for the CAGE-seq mapped on the reference genome. This is due to low number of SNPs in the repetitive sequence of the locus.

Taken together, the multimapping results on the revised miR-430 locus suggest a likely arrangement where we predict that each promoter is marked by promoter-associated post-translational histone modification marks such as H3K27ac and an excessively high marking density spreading over several hundred kilobases may contribute to a transcriptionally favourable environment in the otherwise globally repressed nuclear environment. This model explains the unusually strong H3K27ac signal detected by immunohistochemistry ((Chan et al. 2019) and by *in vivo* antibodies (Sato et al. 2019)

The ATAC-seq profiles at miR-430 indicate a generally high degree of openness spreading beyond the promoter region and extending to the flanking gene, raising the possibility that the whole locus is generally open and functions both as binding surface for transcription factors as well as template for transcription. Temporal dynamic of ATAC peaks broadly correlate with expression timing among the analysed genes and the amplitude of chromatin opening associated signals also appear to correlate with their expression dynamics. The pattern of chromatin opening at the TSS shows the expected profile for most gene groups, with the open region covering roughly 1 nucleosome width area surrounding the TSS.

Intriguingly, there is a dramatic drop in signal intensity in both ATAC open chromatin and H3K27ac signals reaching a trough at around +400bp from TSS on the miR-430 primary transcript gene.

The consistent drop in signals at an unexpected position and which shows striking difference from the first, main wave and constitutively expressed genes either indicates a unique chromatin topology or mapping artefact. To answer this question, mapped read counts on genomic sequence were analysed and a corresponding drop in mappable reads were found also centering at the same position where the signal drop was detected (data not shown), which suggests that the reason for ATAC and H3K27ac signal drop is due to low mapability affecting these sequences.

The amplitude of H3K27ac at dome stage globally correlates with the expression timing of the gene groups analysed with the earliest expressed gene miR-430 showing the highest signals. However, the temporal dynamics of H3K27ac does not correlate with expression timing of the studied gene groups. There is a weak but notable presence of H3K27ac at 4-cell stage in the constitutive and late gastrulation genes, which is in contrast with the lack of signal at the miR-430 at this stage which is expressed much earlier. This unexpected result warrants further studies into how H3K27ac may be deposited soon after fertilisation and what its role may be. While H3K27ac has been shown to be required for activation of early expressed miR-430 (Chan et al. 2019), presence of H3K27ac at genes

which are not expressed or at least are not detected at the 4-cell stage suggest that H3K27ac is required but not sufficient for gene activation.

The most notable difference between miR-430 and other first wave genes is the apparent complete lack of H3K4me3 at this locus. This result suggest that miR-430 carries distinct promoter features from the rest of the first wave genes, which typically carry H3K4me3 signals. This result may reflect the lack of sufficient CpG dinucletide density which is targeted for deposition of H3K4me3 (Guenther et al. 2007, Mikkelsen et al. 2007, Long et al. 2013) or that H3Kme3 which stabilizes TAF3 and thus supports GTF binding as prerequisite for Pol II recruitment (Vermeulen et al. 2007, Lauberth et al. 2013) is not required at this promoter and may suggest that distinct promoter recognition complexes bind to this promoter. These observations suggest that miR-430 is likely regulated in a manner distinct from the other first wave genes.

5 CHARACTERISATION OF PROMOTER SEQUENCE FEATURES AND MOTIFS OF EARLY EXPRESSED GENES

5.1 Stem cell factors are vital for onset of ZGA

Transcription factors implicated in genome activation vary among animal models, suggesting more than one, non-conserved mechanisms in gene activation in animal zygotes (reviewed in Schulz et al. (2019)). As was reviewed in Chapter 1, in Drosophila, a zinc finger protein, Zelda, which presently has no described homolog in vertebrates, has been identified as the key regulator of genome activation. In contrast, work in mouse and human embryos identified Dux-family proteins as regulators of genome activation (De Iaco et al. 2017), although it has recently been questioned whether these factors are indeed essential for genome activation (De Iaco et al. 2020).

The OSKM stem cell factors *Oct4*, *Sox2*, *Klf4* (*Krüppel-like factor 4*), and *c-Myc* regulate gene activation in embryonic stem cells (Zeng et al. 2005, Soufi et al. 2015). Intriguingly, ZGA in fish has been shown to be mediated by the stem cell factor homologs: *Oct4* homolog *Pou5f1 (also known as Pou5f3)* and homologs to *Sox2 (Sox2, Sox19b, SoxB1 (Leichsenring et al. 2013, Veil et al. 2019))*, an additional TF *Nanog* has been implicated although the necessity of its function in embryonic cells is controversial (Lee et al. 2013, Leichsenring et al. 2013,

Gagnon et al. 2018). These stem cell factors are also called pioneering factors, characterised by their ability to both bind and subsequently open up nucleosome occupied DNA at promoters (Veil et al. 2019), providing chromatin accessibility necessary for association of basal transcription machinery at the main wave of genome activation. In their study, Veil et al. (2019) demonstrated how flanking binding sites of *Pou5f1* and *Nanog,* when occupied, shift the adjacent nucleosome downstream, and subsequent binding of *Sox19b* together with *Pou5f1* evict the nucleosome, to allow regulatory element opening. Similarly, Pálfy et al. (2020) studied open chromatin during main wave of genome activation and combined loss of function of *Pou5f1*, *Sox19b* and *Nanog* resulted in decreased accessibility indicating their pioneering role during main wave of genome activation.

The role of maternal pioneering factors *Pou5f3* and *Sox3* in predefining first zygotic responses to inductive signals was demonstrated in *Xenopus* (Gentsch et al. 2019), suggesting a selective role for stem cell factors in opening promoters of genes which are targets of regionalising signalling. They found *Pou5f3-Sox3* TFBS enrichment in pre-MBT Pol II occupied regions. The TFs were found to be synergistic by using loss of function experiments. Their model suggests opening by these factors then subsequent additional binding of developmental context-dependent pioneering factors, together leading to activation of genes. These factors are deposited maternally in the embryo as mRNA and may be translated sufficiently early in development to facilitate opening of chromatin at promoters of first wave genes, thus they are candidates for their activation. Early translation

of maternal factors has been suggested to be key for first wave of genome activation by Chan et al. (2019), in zebrafish.

The above described results indicate that promoters are differentially opened during genome activation and it remains to be explored how the first wave of genes are regulated on the promoter level. Promoter level regulation in this context may include regulator function of specific transcription factors which selectively bind to promoters but can also be a subject of core promoter level regulation, mediated by general transcription factors. As was reviewed in the general introduction, core promoters can be broadly classified into several categories based on their promoter architecture (transcription start site distribution, such as broad and sharp peak promoters), but also based on their sequence composition, indicative of positionally constrained binding of general transcription factors, which contrary to their collective name, have been shown to selectively bind subsets of promoters, e.g. TATA-box containing promoters, W-box or TATA-like element containing promoters or non-TATA CpG enriched promoters to name just a few key categories, reviewed in (Lenhard et al. 2012, Haberle et al. 2018).

It has already been demonstrated that the core promoter plays a role in regulation of genome activation and distinct regulatory mechanisms have been proposed for core promoter level regulation during oocyte development, which is the source for maternally deposited RNAs in the early embryo and subsequent regulation of the same genes during main wave of genome activation. The zebrafish oocyte

utilises W-box mediated definition of TSS positions indicating the function of a TBP family member (likely TBP2 (Bartfai et al. 2004, Akhtar et al. 2009)) in regulating these genes in the oocyte, and a distinct nucleosome positioning signal dependent TSS definition during the main wave of genome activation, suggesting a potentially different initiation complex formation, which does not require binding to a TATA-box or TATA like sequence by a TBP family protein (Haberle et al. 2014). The results indicate at least two different types of initiation complexes functioning in defining the transcriptome during the maternal to zygotic transition. Little is known about the characteristics and activation principles of the first wave genes of genome activation. The only study so far on this has been done in Drosophila, in which approximately 100 genes were identified to be active before main wave of ZGA before the MBT and they were shown to, in contrast to the main wave genes, contain a TATA element, suggesting regulation by TBP (Chen et al. 2013).

5.2 Aims

In regard to the metazoan promoter, sequence composition primarily determines which factors i.e., initiation complexes are most likely to regulate these promoters. Currently a specific classification for early expressed genes has not been described, therefore this chapter will explore whether genes active before the MBT have a distinct promoter architecture, whether they are more likely to be bound by known stem cell factors and thus surmise whether these promoters may be activated by an alternative mechanism.

Aims have been to:

- Characterise the promoters of first wave genes in order to determine potential enrichment of promoter sequences specific to first genes to be expressed in the embryo that may suggest differential transcriptional regulation.
- In order to better understand how early genes are distinct from main wave genes, in relation to how they can be activated prior to global onset of ZGA at the MBT, presence of transcription factor binding sites was explored with focus on the stem cell factors to better determine whether a distinct chromatin accessibility environment may be conferred through the core promoter sequence.

5.3 Early promoter classification through analysis of core promoter elements present in first wave genes

In order to investigate potential differences in composition of core promoter elements between four different gene sets, (first wave (FW), main wave (MW), post gastrulation (PG) and constitutive (CT), first described in chapter 4) that would suggest regulation by a different PIC, the core promoter regions of the different genes sets were first compared in regard to identification of distinct differences in core promoter element composition. For this purpose, the core promoter sequences of all genes in each set were extracted and centre-aligned to the dominant TSS. Position weight matrices (PWM) for each core promoter set was calculated from the sequences using the Bioconductor “DiffLogo” package and motif sequence logos were visualised with the R “ggseqlogo” package.

As demonstrated on **Figure 5-1** there was clear enrichment of a TATA-box like motif at the expected position i.e. 30 bp upstream of the TSS in the first wave gene sets compared to the other sets (**Figure 5-1** A). In the main wave set there was a weaker TATA-box like motif, more resembling a W-box, which was less prominent in the post gastrulation set. As expected, no TATA-like motif was identified in the constitutive gene set (aligned to their zygotic TSS) (**Figure 5-1** D).

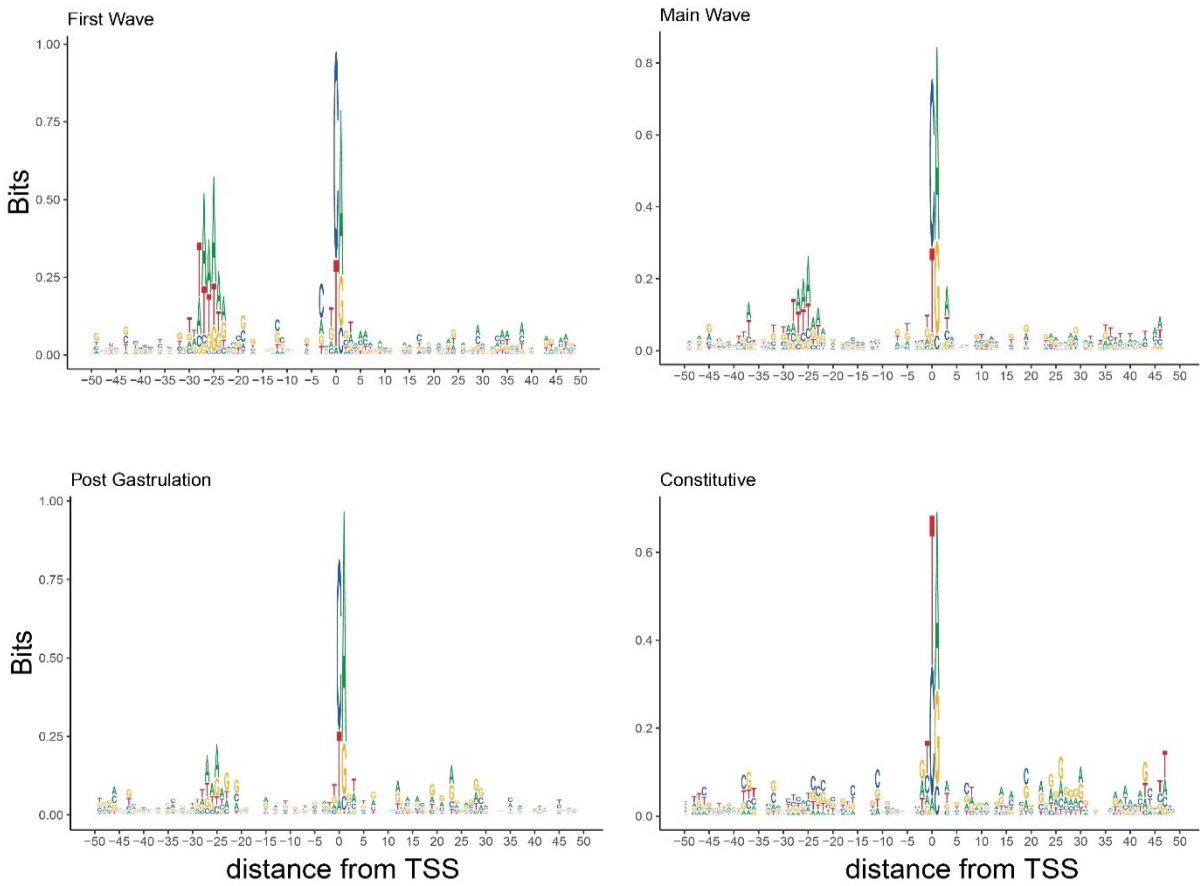


Figure 5-1 Position motifs composition around the TSS in promoter sequences show enrichment in TATA-box for first wave genes

Base-specific information content of promoter sequences spanning -50 bp upstream/ downstream of the TSS demonstrates most the common bases in four pre-determined gene sets.

To confirm this enrichment for TATA-box like motif in the first wave genes compared to the rest, a positional motif enrichment analyses using the canonical TATA-box (TBP) motif position weight matrix, obtained from JASPAR, was performed with the CentriMo tool (Bailey et al. 2012), from the meme software suite (http://meme-suite.org/). This tool is designed to look for enrichment of

motifs around particular position i.e. TSS. It also can show if the local enrichment is significant relative to control (negative) set of sequences.

As shown on **Figure 5-2**, the enrichment for canonical TATA-Box at the expected location (the peak centres around 26 bp upstream of the TSS) in the first wave genes was confirmed. There were 26 genes having canonical TATA-Box in the first wave set, compared to 6 in the main wave (**Figure 5-2** A,D), 3 in the post gastrulation (**Figure 5-2** B,D) and none in the constitutive set (**Figure 5-2** C,D).

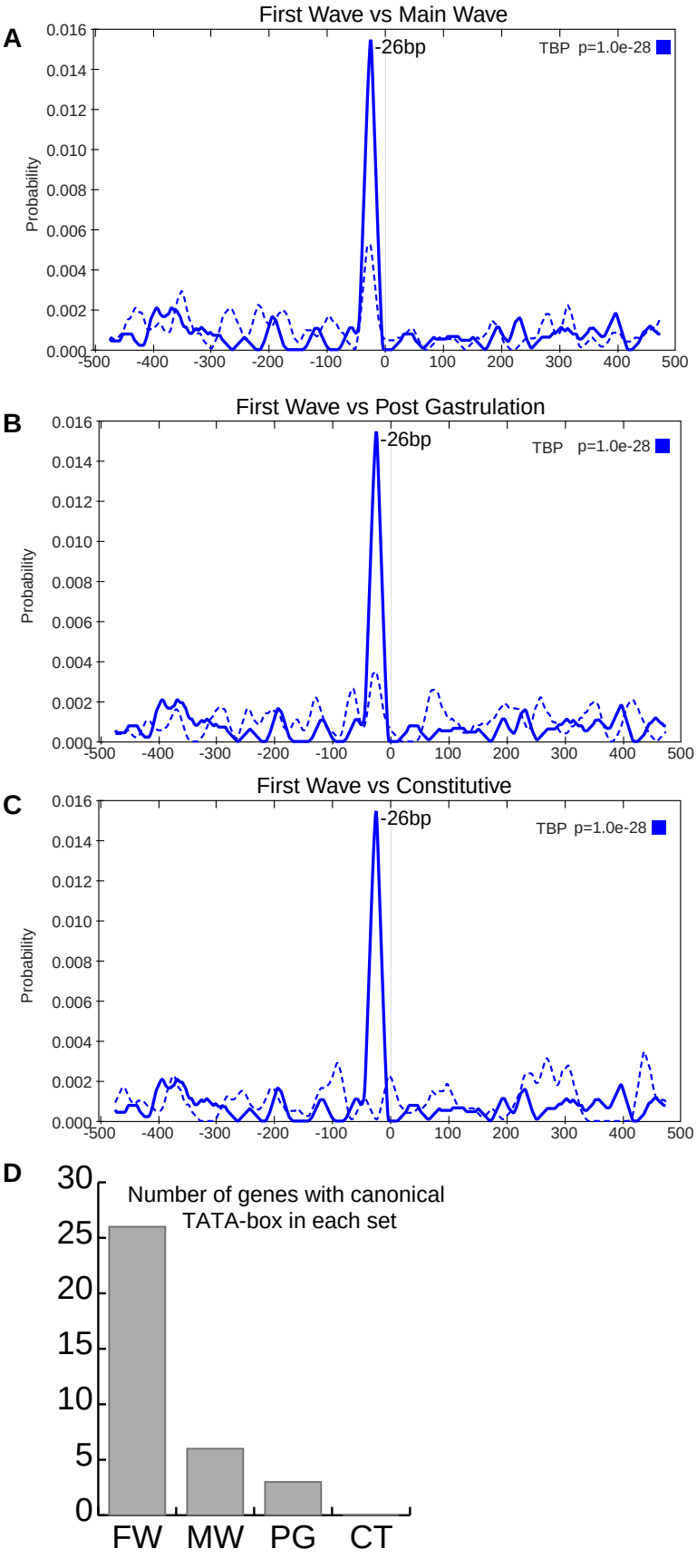

A
First Wave vs Main Wave
-26bp
TBP p=1.0e-28
Probability
B
First Wave vs Post Gastrulation
-26bp
TBP p=1.0e-28
Probability
C
First Wave vs Constitutive
-26bp
TBP p=1.0e-28
Probability
D
Number of genes with canonical TATA-box in each set
FW
MW
PG
CT

Figure 5-2 TBP binding site motif enrichment around the TSS in promoter sequences of first wave genes

Enrichment for TATA-binding protein (TBP) motif is shown on aggregation plots demonstrating probably of motif presence in all sequences. (A-C) Distribution of TBP motifs of first wave genes is determined using main wave (A), post gastrulation (B) and constitutive (C) gene sets respectively, as controls. Dotted lines represent distribution of TBP motif probabilities in control set which varies per plot in respect to the varying distribution in control gene sets. Chart below plots (D) correspond to counts of number of genes with identified TATA-box in four gene sets: first wave (FW), main wave (MW), post gastrulation (PG) and constitutive (CT).

The above analyses show the core promoters of first wave genes are distinct form the other sets by being enriched for TATA-box type promoters, generally associated with a sharp promoter architecture, caused by transcription initiated from single site at the promoter, due to the positional constraints imposed upon PIC assembly (Ponjavic et al. 2006).

An example of a representative gene for each of the four promoter architectures, for further information on promoter architectures please see section 1.2.3. A demonstration of the characteristic transcription initiation patterns is shown in **Figure 5-3.** A further example of the distinction between a TATA-box and W-box promoter can be seen in panels A and B in **Figure 5-3.**

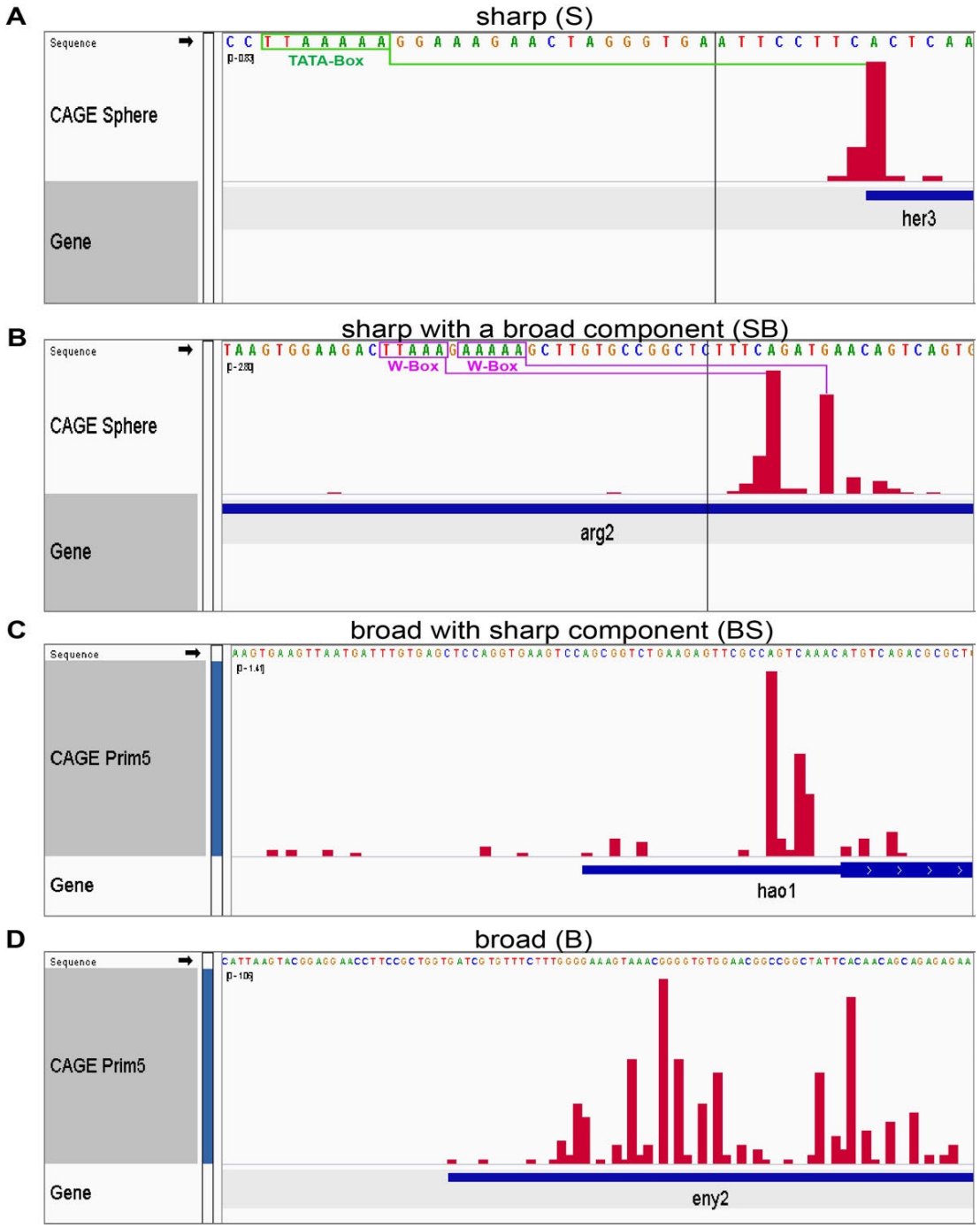


Figure 5-3 Transcription initiation patterns characteristic for known promoter architectures

*Promoters with interquartile width up to 3 bp were classified as **(A)** sharp (S), **(B)** between 4 and 10 bp as sharp with a broad component (SB), **(C)** between 11 and 20 bp as broad with sharp component (BS) and **(D)** above 20 bp as broad (B). Four*

architectures are shown with representative gene demonstrating the characteristic profile of TSS initiation at an appropriate stage when the gene is active.

To check if the first wave genes also have sharp promoter architecture, akin to miR-430 and other type I genes, the promoter type (i.e. architecture) was determined for each gene in the four datasets, using the 0.1-0.9 interquartile width (the length of the region in bp where 10%-90% of the CAGE signal is concentrated) of the promoter associated CAGE-tags consensus clusters. Promoters with interquartile width up to 3 bp were classified as sharp (S), between 4 and 10 bp as sharp with a broad component (SB), between 11 and 20 bp as broad with sharp component (BS) and above 20 bp as broad (B).

The frequency of each promoter type with gene set is shown on **Figure 5-4**. There was a clear enrichment for sharp promoters in the first wave genes, consistent with the above observation that they are more similar to type 1 TATA-box promoters.

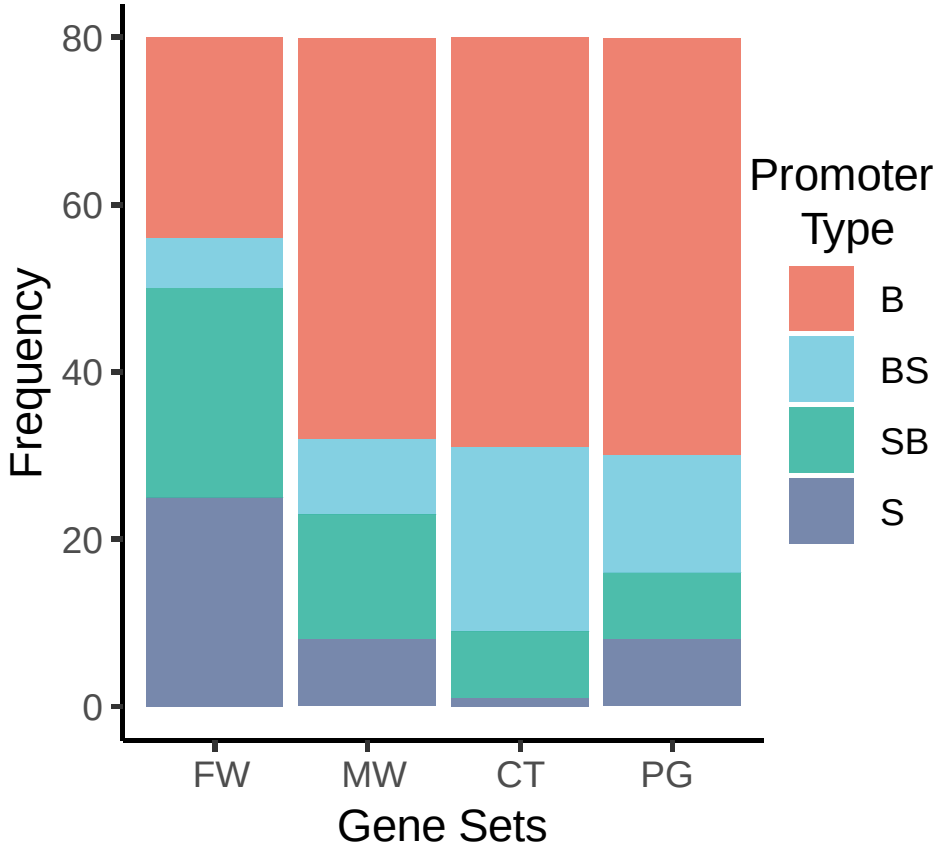


Figure 5-4 Distribution of promoter architectures reveals predominately sharp initiation for first wave genes

Frequency of different promoter architectures, promoters with interquartile width up to 3 bp were classified as sharp (S), between 4 and 10 bp as sharp with a broad component (SB), between 11 and 20 bp as broad with sharp component (BS) and above 20 bp as broad (B). Frequency is determined from counts of genes (of a total 80) classified into one of four architectures. Characterisation is shown for four gene sets: first wave (FW), main wave (MW), constitutive (CT) and late/post gastrulation (PG).

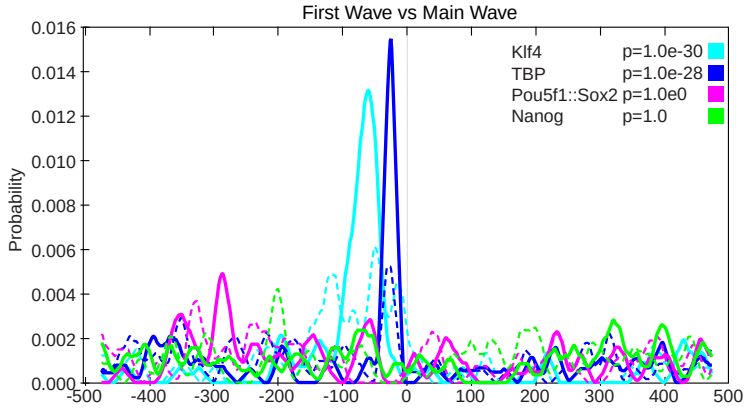
Further investigation into whether the first wave gene set shows enrichment of binding sites of stem cell factor known be required in early genome activation (Lee et al. 2013, Leichsenring et al. 2013) was initially performed via global enrichment analysis for stem cell factor binding sites: *Nanog*, *Sox2*,

Pouf5f1::Sox2 and *Klf4* motifs (obtained from JASPAR core vertebrate database, see appendix 8.3 for matrix IDs), comparing the first wave genes versus the other sets with the AME tool from the MEME suite which searches for motif enrichment on a global scale regardless of position. The analysis was performed on different region sizes (5 kb, 3 kb and 1 kb) up and downstream from the TSS, but no statistically significant enrichment of any of the above motifs was found.

I hypothesised that these factors might be locally enriched around promoters during zygotic genome activation, due to being reported to bind in the vicinity of the proximal promoter while suggested to drive displacement and eviction of nucleosomes for both initial opening and later (post-ZGA) for maintaining open chromatin state (Veil et al. 2019). For this purpose, local motif enrichment analyses of the core and proximal promoter regions (500 bp upstream and 500 bp downstream of the TSS) was performed using the aforementioned CentriMo tool using the position weight matrices (PWM) for the above stem cell factors and the canonical TATA-box motif (TBP) as a positive control (**Figure 5-5**). Unlike AME, CentriMo looks for enrichment positionally in relation to various genomic landmarks such as ChIP-seq peak enrichment and TSSs.

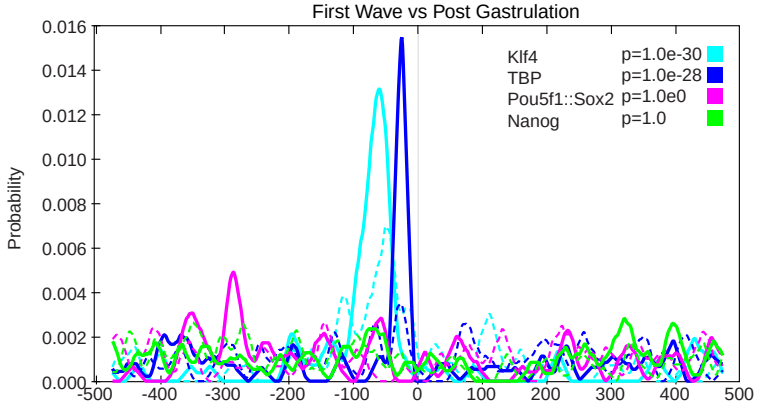
Interestingly, the *Klf4* motif was found to be also enriched in the first wave genes at a particular position (~70 bp upstream) of the TSS in a relatively narrow region of 55 bp (**Figure 5-5** A,B,C). A slight enrichment for the *Pouf5f1::Sox2* motif at ~300 bp was observed, but was not statistically significant according to the p-values. *Nanog* and *Sox2* motifs were not enriched.

A



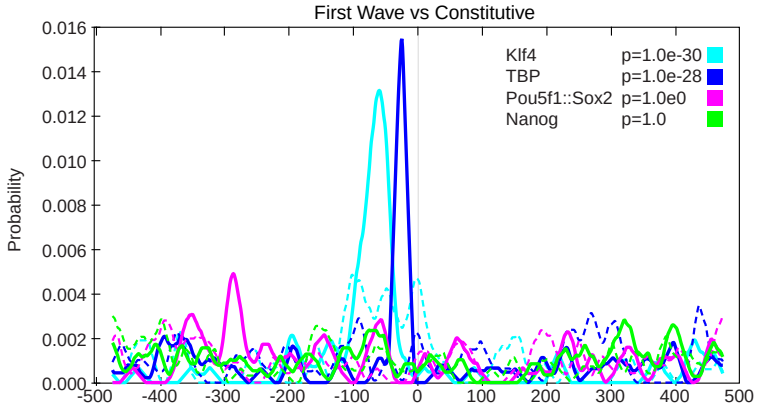
TF	Fisher E-value	p-value	Negative p-value	Region Center	Region Width	Region Matches	Negative Region Matches
Klf4	2.4	1E-30	1	-69.5	55	44	14
TBP	5	1E-28	1	-26.5	8	26	6
Pou5f1::Sox2	5	1	1	-284.5	24	10	3
Nanog	5	1	1	-495	1	0	0
Sox2	5	1	1	-496	2	0	0

B



TF	Fisher E-value	p-value	Negative p-value	Region Center	Region Width	Region Matches	Negative Region Matches
Klf4	2.1	1E-30	1	-69.5	8	26	3
TBP	5	1E-28	8.1E-3	-26.5	55	44	18
Pou5f1::Sox2	5	1	1	-284.5	24	10	1
Nanog	5	1	1	-495	1	0	0
Sox2	5	1	1	-496	2	0	0

C



TF	Fisher E-value	p-value	Negative p-value	Region Center	Region Width	Region Matches	Negative Region Matches
Klf4	3.8E-3	1E-30	1	-69.5	8	26	0
TBP	5	1E-28	8.1E-3	-26.5	55	44	12
Pou5f1::Sox2	5	1	1	-284.5	24	10	2
Nanog	5	1	1	-495	1	0	0
Sox2	5	1	1	-496	2	0	0

Figure 5-5 transcription factor binding site motif enrichment around the TSS in promoter sequences of first wave genes

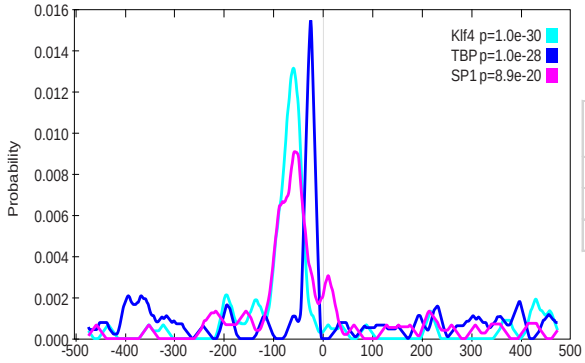
Enrichment for three vertebrate motifs: TATA-binding protein (TBP), and stem cell factors Klf4, Pou5f1 in combination with Sox2 and Nanog are shown on aggregation plots demonstrating probably of motif presence in all sequences. (A-C) Distribution of TF motifs of first wave genes is determined using main wave (A), post gastrulation (B) and constitutive (C) gene sets respectively, as controls. Dotted lines represent distribution of motif probabilities in control set which varies per plot in respect to the varying distribution control gene sets. Tables below plots correspond to scoring for each motif.

The mammalian Klf4 motif, representing recognition sequences of Klf4, one of the OKSM (Oct4, Klf4, Sox2 and cMyc) stem cell factors used to reprogram somatic cells to iPSCs (Takahashi et al. 2006) was detected ~70 bp upstream of the TSS. The motif was found to span a 50 bp region in the first wave genes which was much more positionally constrained compared to other gene sets (**Figure 5-5**). KLF4 is a member of a larger Klf/SP1 family of C2H2 zinc fingers, which recognise similar binding sites. In particular, SP1 is known to be enriched on promoters in 21% of human genes (Hughes 2011) and has highly similar recognition motif to Klf4, therefore in order to determine if the Klf4 enrichment around the TSS is due to SP1 motif enrichment, positional motif enrichment analysis was performed for each dataset separately, with inclusion of the vertebrate SP1 motif from the JASPAR database **Figure 5-6**.

From the aggregation plots in **Figure 5-6** it was found that although both SP1 and Klf4 motifs were enriched at a significant level at promoters of the first wave gene

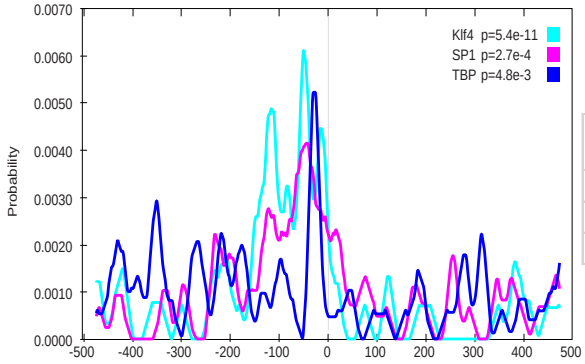
set, better scoring (lower p-value) and higher enrichment was found for Klf4 motif (**Figure 5-6** A). Furthermore, Klf4 was also found to have higher enrichment over SP1 in the main wave gene set (**Figure 5-6** B), in comparison to both the post gastrulation and constitutive sets which had a higher enrichment for SP1 over Klf4 (**Figure 5-6** C,D), demonstrating that genes activated around the MBT have higher enrichment for stem cell factor Klf4 binding motif. Interestingly, both Klf4 and SP1 motifs in first wave gene sets showed a stronger positional constraint, concentrated within ~50 and ~80 bps respectively (**Figure 5-6** A). Whereas in other gene sets, recognition sites were spread over more than 100 bps. This observation is in line with that fact that these promoters are sharp initiating, TATA-Box promoters. Also due to its relative proximity to the TATA-box (about 30-40 bp upstream) it can be surmised that whichever KLF family member binds this motif may play a role in initiation by directly interacting with the PIC.

A



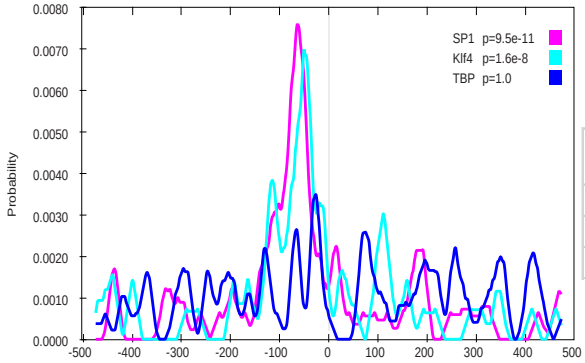
TF	p-value	Region Center	Region Width	Region Matches
Klf4	1E-30	-69.5	55	44
TBP	1E-28	-26.5	8	26
SP1	8.9E-20	-59.5	81	41

B



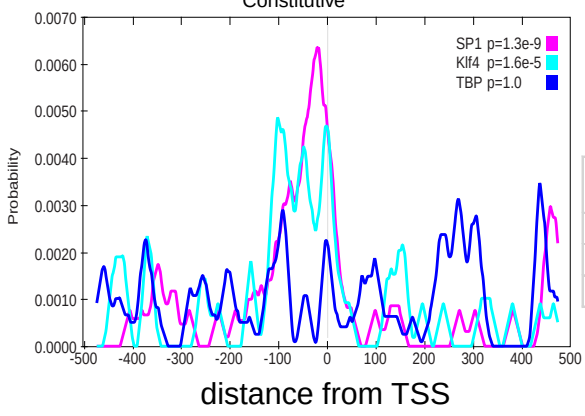
TF	p-value	Region Center	Region Width	Region Matches
Klf4	5.4E-11	-79.0	148	40
TBP	4.8E-3	-27.5	10	10
SP1	2.7E-4	-99.5	267	45

C



TF	p-value	Region Center	Region Width	Region Matches
Klf4	1.6E-8	-64.5	115	33
TBP	1.0	-493	1	0
SP1	9.5E-11	-60.5	161	44

D



TF	p-value	Region Center	Region Width	Region Matches
Klf4	1.6E-5	-48.5	147	29
TBP	1.0	-493	1	0
SP1	1.3E-9	-43.5	129	35

Figure 5-6 transcription factor Klf4 binding site motif is positionally constrained at the TSS of first wave genes

Aggregation plots for four gene sets (A-D) showing comparisons in probability for presence of Klf4 and SP1 binding sites, TBP is shown as an additional reference point indicating where the PIC associates with the core promoter.

In order to infer presence of potential *cis-regulatory* elements that regulate transcription, it was determined whether existence of a localised concentration of binding sites at the approximately 120 miR-430 gene units (miR-430 promoter lined to a triplet structure) of the repetitive miR-430 cluster, described and characterised in Chapter 4, could be predicted. For this purpose the miR-430 contig was analysed for potential enrichment of stem cell factor binding sites: *Pou5f1:Sox2*, *Sox2, Nanog* and *Klf4, (*datasets listed in **appendix 8.1**).

Cluster buster (CBust) is a software for locating clusters of pre-specified motifs in a DNA sequence (Frith et al. 2003). Cbust analysis of the pre-specified motifs on the miR-430 contig resulted in the miR-430 annotated region being identified as one single cluster, enriched for high scoring binding sites. Although Klf4 binding sites were identified, they mostly had scores below the default threshold of 6. In addition to the locus there were another 2-3 minor clusters identified. The **Figure 5-7**A is an overview of which encompasses the entire miR-430 cluster, with a zoomed in view (**Figure 5-7**B) of the locus showing the exact location of identified motifs.

Predicted binding sites for *Pou5f1*::*Sox2* and *Nanog* are present in the upstream region from TSS and interestingly, the *Nanog* motif corresponds with the start of the ChIP-seq signal (**Figure 5-7** B). The *Pou5f1*::S*ox2* site is of a similar score and position -300 bp from TSS as the motif enriched in the first wave genes. The Pou5f1 and SoxB1 ChIP-seq coverage is of a lower resolution than the Nanog coverage which may account for the broader distribution of signal across the *Pou5f1::Sox2* and *Sox2* motifs.

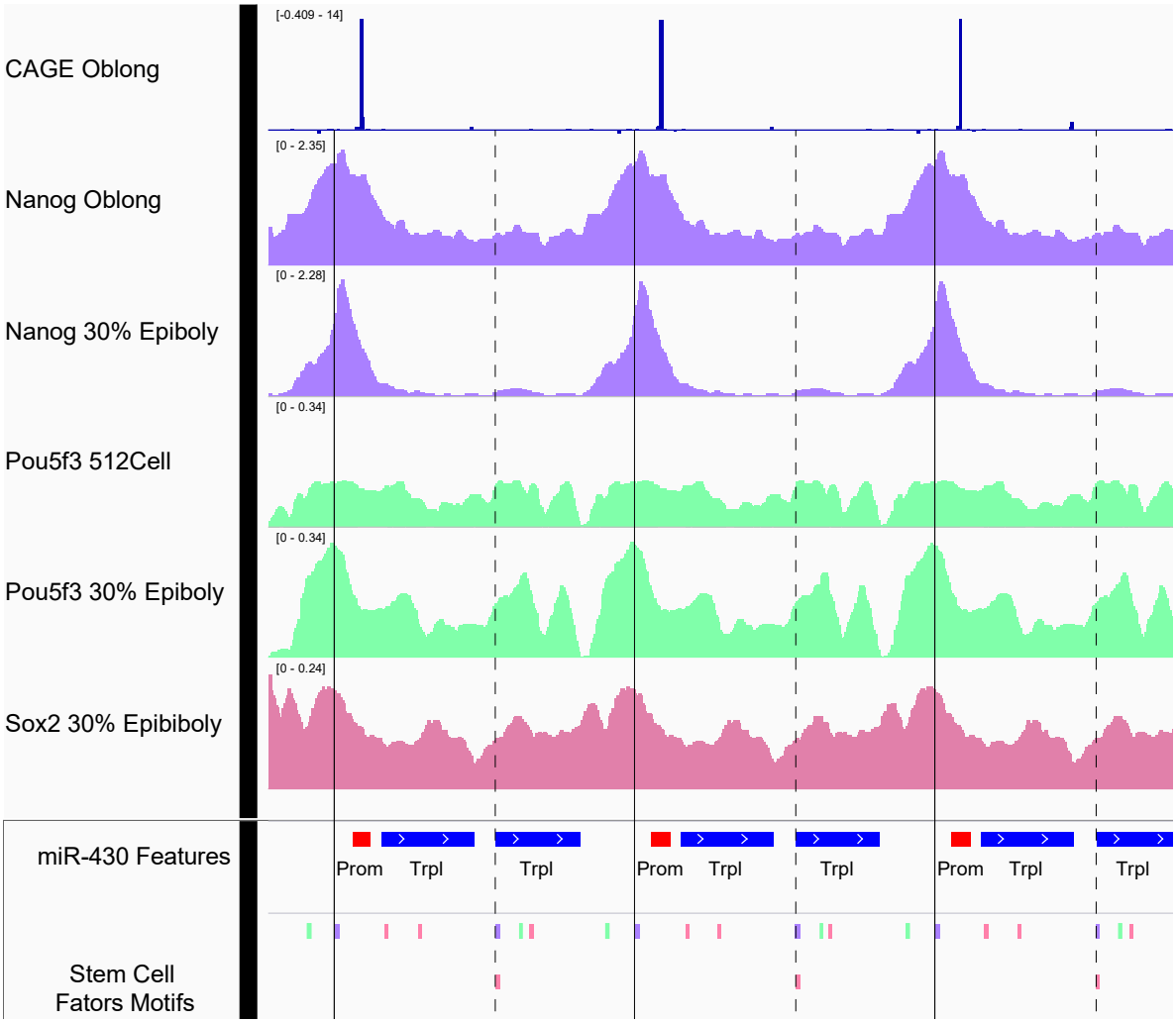


Figure 5-7 Stem cell factor binding motif distribution on miR-430 cluster

Predicted vertebrate binding motifs for three stem cell factors Pou5f1, Sox2 and Nanog with corresponding ChIP-seq coverage spanning across a representative portion of the miR-430 cluster. Binding sites and ChIP-seq coverage are coloured per specific factor i.e purple: Nanog, green: Pou5f1, red= Sox2. CAGE-seq determined CTSSs are displayed on the uppermost row to indicate start sites of transcription.

5.4 Discussion

It has been shown in this chapter that promoters of early genes have a distinct architecture, with increased enrichment for TATA-box motif in comparison to main wave genes with sharp initiation pattern, more akin to the profile of type 1 genes associated with adult terminally differentiated peripheral tissue (Lenhard et al. 2012, Haberle et al. 2016, Haberle et al. 2018). The promoter profile for first wave genes is distinct from genes active at the MBT which, in contrast, have an enrichment for a maternal W-box with a broad initiation pattern. This result is in line with findings in Drosophila (Chen et al. 2013) where a group of genes with a TATA element promoter profile, distinct from genes active at the MBT, were found to be active before global ZGA.

An interesting observation concerns the enrichment of a positionally constrained *Klf4* motif on genes active during ZGA, which incidentally, is most enriched on early genes and in particular, on first wave genes. *Klf4* is a member of the Klf/SP1 family of C2H2 zinc fingers, alongside SP1, which is known to commonly associate with CGI promoters (Brandeis et al. 1994). In combination with the GTFs, SP1 is believed to support nucleation of the PIC at the core promoter and subsequent maintenance of the hypomethylated state of the local CGIs.

Despite the four gene sets, FW, MW, CT and PG having enrichment for CGIs within the core promoter (**Figure 4-12**), there was a differential distribution of Klf/SP1 family member binding enrichment, with post-MBT gene sets consisting of constitutive genes and genes active post gastrulation showing a higher

preference for SP1 motifs and early gene sets expressed pre-MBT more enriched in *Klf4.* Although it is important to note that due to the high degree of similarity in recognition of binding sites between members of the Klf4/SP1 TF family, it is not clear whether *Klf4* or another family member altogether binds to TATA element containing genes. Despite this, the fact that SP1 is not usually positionally constrained and does not regulate TATA promoters argues against it being a regulator of this specific type of promoter, instead it can be surmised that a Klf/SP1 factor other than SP1 likely binds to genes active during ZGA, indicating a different regulatory mechanism specific to this type of promoter.

Analysis of enrichment for potential binding sites for stem cell factors Pou5f1, Sox2 and Nanog on the most highly expressed early gene, miR-430 revealed presence of all three binding sites on individual gene units, although levels did not exceed enrichment of other first wave genes. Thus, it is highly probable that the key defining feature is the repetitive nature, which results in an overall local high-density enrichment of consecutively placed binding sites which, if functional, could explain why this is one of the first active and most highly expressed genes in the developing embryo, with its unique structure strikingly reminiscent of a super enhancer.

Interestingly, most of the intron-lacking genes transcribed in the embryo prior to zygotic genome activation code for either clusters of histone or zinc finger transcription factors (Gentsch et al. 2019) and miRNAs such as miR-430/miR-427 (Giraldez et al. 2005, Lund et al. 2009, Tani et al. 2010, Heyn et al. 2014,

Owens et al. 2016), with the suggestion, as has been proposed here for miR-430 genes, that early genes may be situated in unique environments within the genome. This may function, in part, to allow mass utilisation of *cis*-regulatory elements; so called super enhancers (SE) (Whyte et al. 2013), promoting higher transcriptional output in regions containing clusters of similar genes shown to play roles in cellular identity not only during development, but also in disease such as cancer where SEs were found located at oncogenes in cancerous human cells *in vitro (Hnisz et al. 2013)*.

Taken together, the results in this chapter lead to the summation that genes active during zygotic genome activation have a distinct promoter architecture, making them divergent from other post-ZGA active genes, suggesting alternative regulation of genes active during ZGA. Yet despite this, it is also clear that first wave genes have features distinct from those expressed at the MBT, such as enrichment for TATA element over the maternal W-box which suggests regulation by TBP, yet it is known that TBP is not present in the embryo until it is transcribed at the MBT (Veenstra et al. 1999). Thus, one could hypothesise that since the zygotic machinery are not yet present, it is likely then that regulation of first wave genes is by maternally deposited factors.

Although all TATA-box binding proteins are fundamental for embryonic development, very few genes in the embryo are reliant on TBP (Jacobi et al. 2007). It has long been known that despite large quantities of maternal transcript presence TBP is fundamentally absent in the oocyte and is not translated until

activation of the zygotic genome occurs (Veenstra et al. 1999, Gazdag et al. 2007), although overexpression of TBP results in premature genome activation in the frog demonstrating the vital role it plays in transcription of genes of the main wave of activation (Veenstra et al. 1999).

The most highly related paralog of TBP is TBP2, with the conservation between c-terminal domains, in particular the binding domains, of such a degree that it is believed to have evolved as a duplication of the TBP gene at the origin of vertebrates (Persengiev et al. 2003). TBP2 has been demonstrated to be fundamental for embryonic development not only in the zebrafish, but also in the *Xenopus* and for mouse oogenesis, with deletion of TBP2 resulting in sterility in female mice (Bartfai et al. 2004, Jallow et al. 2004, Hart et al. 2007, Gazdag et al. 2009).

The suggestion that genome activation is continuous over time as threshold conditions are reached (Vastenhouw et al. 2019) does not take into account promoter-related contributions to regulation of transcription. In particular it is evident that group of genes are always active before the rest (Mathavan et al. 2005, Heyn et al. 2014), yet the mechanism by which these first wave genes are able to transcribe pre-ZGA is still elusive, although there is significant evidence for TBP-family proteins being crucial for correct gene regulation throughout ZGA, with each paralog required differentially throughout development. Despite this, their role in first wave of genome activation is yet to be described.

Due to the lack of zygotic machinery before global genome activation, resulting from lack of TBP transcription prior to the MBT (Veenstra et al. 1999), other mechanism(s) must be in place to facilitate transcription pre-MBT. TBP2 is a maternally deposited TBP paralog (Bartfai et al. 2004) that recognizes TATA box motifs in a similar manner to TBP (Bartfai et al. 2004, Jallow et al. 2004), making it a candidate factor.

Unpublished work by Tora *et al.* indicates that TBP2 is able to form a non-canonical initiation complex in association with TFIIA in mouse embryonic stem cells (mESCs), with such preference being distinct from TBP which has TFIID-regulated sequence requirements. Interestingly, they also suggest that the TBP2-specific TATA box preference is largely associated with a sharp promoter architecture, which would be in line with the result from this chapter that first wave genes have a strong TATA-box enrichment with a sharp initiation pattern. This result is in contrast to TBP/TFIID, which preferentially drives a broad TSS selection from TATA-box promoters as is seen in Tora *et al.* unpublished work in mESCs, again in agreement with the main wave gene profiles shown here, suggesting possible regulation of first wave genes by the maternally deposited TBP2/TFIIA non-canonical initiation complex in the developing embryo.

6 EARLY GENE PROMOTERS: DOES SEQUENCE OR ENVIRONMENT FACILITATE FIRST WAVE GENE ACTIVITY?

6.1 Aims

It was demonstrated in Chapter 5 that genes expressed before global ZGA have a promoter structure distinct from both the main wave and late zygotic genes, more akin to the structure seen in adult tissue specific genes. The aim of this chapter is to explore, through using transgenesis, whether the sequence of the most highly expressed pre-MBT promoter in *zebrafish, miR-430,* will be sufficient to drive pre-MBT expression of a reporter either transiently, or when integrated in a single copy on different genomic location, addressing whether circumvention of transcriptional repression pre-MBT in the zebrafish model organism is dependent on the endogenous genomic environment or whether the early promoter sequence is sufficient to drive expression at pre-MBT.

Foreword:

All plasmid maps for constructs described in this chapter are located in appendix 8.8

6.2 Functional validation of miR-430 promoter sequence in transient transgenic assay.

In order the determine whether the predicted promoter sequence of the miR-430 is able to support transcription, a 650 bp promoter fragment was selected for testing, because it extends in both directions to the upstream and downstream triplet structures respectively from the CAGE-seq determined TSS and was predicted to contain the proximal promoter elements required to drive transcription (**Figure 6-1**).

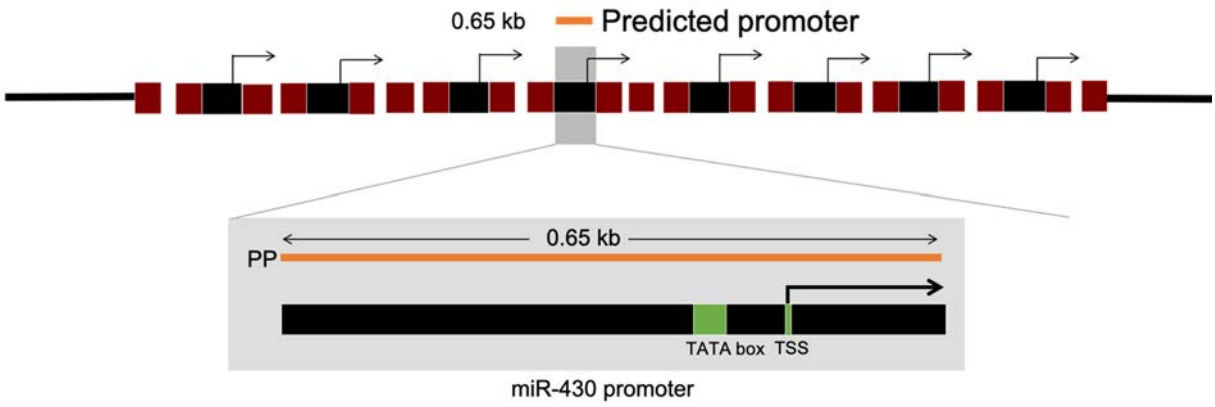


Figure 6-1 Schematic demonstrating predicted promoter region from miR-430 locus used for functional validation.

Black bar, 650 bp bar denotes predicted promoter (PP) fragment shown overlaid onto part of the miR-430 cluster; white boxes represent intergenic sequence, wine colour indicates triplets. Grey box spans fragment and represents predicted promoter region containing both TATA box motif and TSS.

The promoter sequence was PCR amplified from genomic DNA and cloned into a plasmid containing red fluorescent protein as a reporter. To ask whether the putative promoter sequence is able to support transcription in the embryo, the generated reporter constructs were injected into zebrafish zygotes and assayed for reporter activity *in transient* transgenic assays by fluorescent microscopy at different developmental stages. 157 embryos were injected, of which 65% showed fluorescence signal in a mosaic fashion of which the signal is characteristic to transient transgene activity, due to uneven inheritance of the transgene in the developing embryo. Fluorescence was detected from bud stage at 10 hpf indicating maturation of RFP.

Detection of transgene signal in live embryos from bud stage through to prim5 stages following injection of miR-430 promoter reporter construct demonstrates the competency of the cloned promoter fragments to drive transcription of the reporter gene. However, signal was not detected before the main wave of genome activation (**Figure 6-2**). This result could have two possible explanations, either the tested promoter fragment is not active before MBT due to a requirement for miR-430 promoters to be in an array structure or the maturation and accumulation of the reporter protein has not reached detection levels, which may have pushed forward the point at which fluorescent signal detection could occur, thus hindering temporal gene expression analysis.

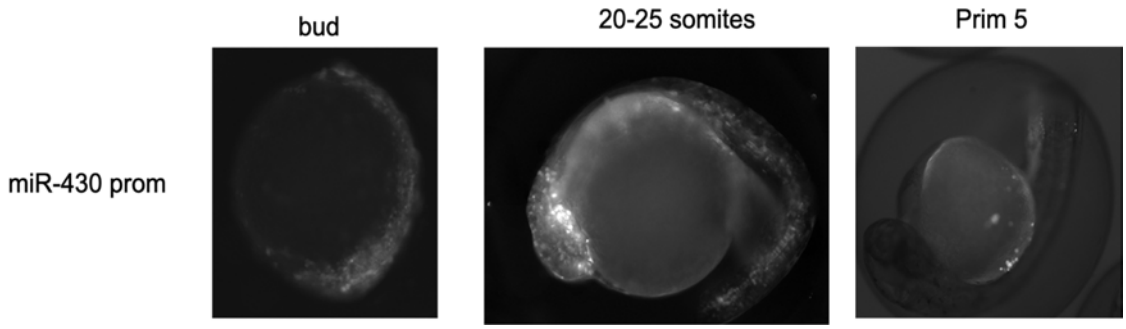


Figure 6-2 in transient reporter fluorescence detected post-MBT in vivo

Fluorescence of the red reporter gene detected in embryos injected with miR430 promoter reporter construct detected post-MBT, from bud stage to 24 hpf.

To determine whether transcription from the reporter construct precedes fluorescence reporter expression detection and to assess whether the construct was capable of activating transcription before the main wave of ZGA, whole mount *in situ* hybridization (WISH) experiments were performed to directly detect

expressed mRNA transcripts via targeting of the reporter sequence from reporter construct injected embryos.

As controls, embryos were injected with two constructs containing the same reporter but different promoters: CMV and *gata2* (Maeno et al. 1996), known to be active at and post-ZGA, respectively. Comparison of the temporal gene expression profile of endogenous mir430 and *gata2* were analysed by comparing WCE CAGE-seq defined TSS profiles as shown in (**Figure 6-3**) demonstrating the distinction in temporal dynamics between the endogenous genes and the transiently expressed reporter constructs (**Figure 6-4**)

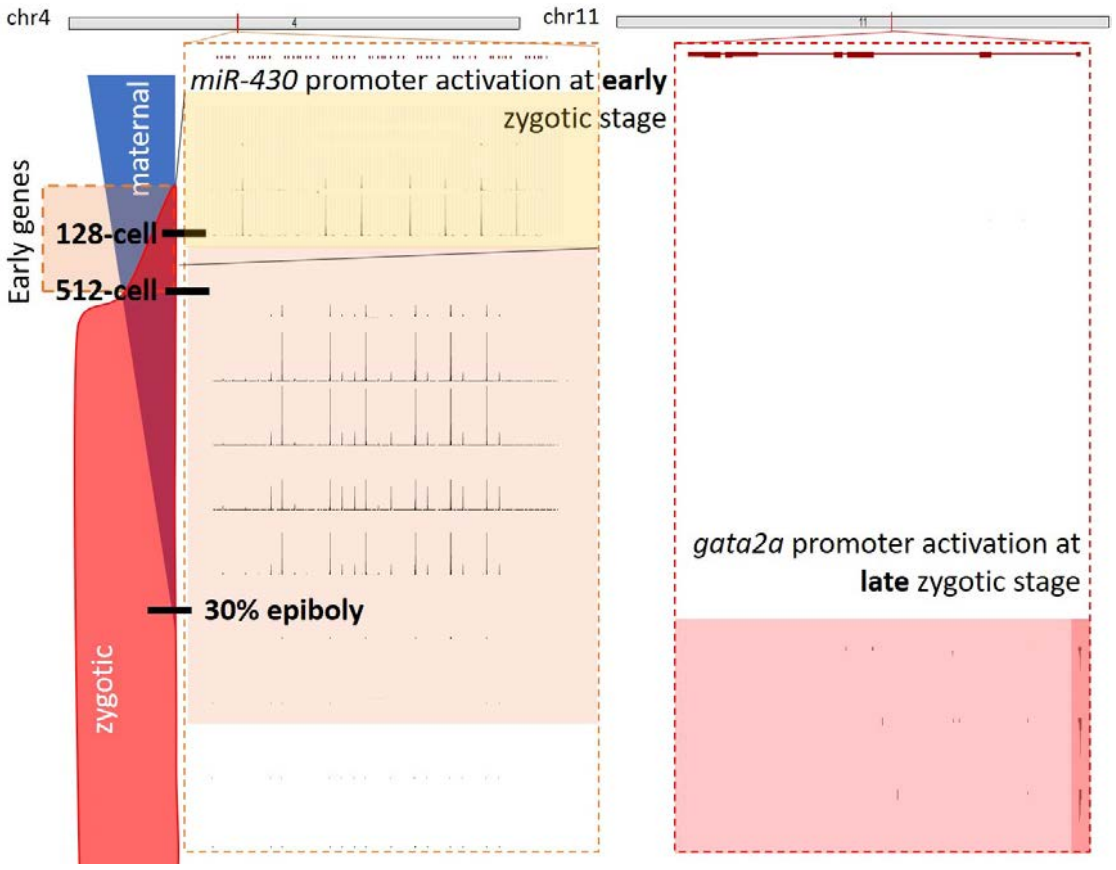


Figure 6-3 Demonstration of early activation of miR-430 promoter in comparison to late activated gata2a

CAGE profiles distinguish early miR-430 expression from late zygotic gata2a gene of which the promoter sequence was used as a control for in transient vector transgene detection via in situ hybridisation. Data is plotted on a fixed scale for comparison between stages. Gene models are indicated as red bars and lines. Minor and major wave of gene activation are highlighted as orange and red fields respectively.

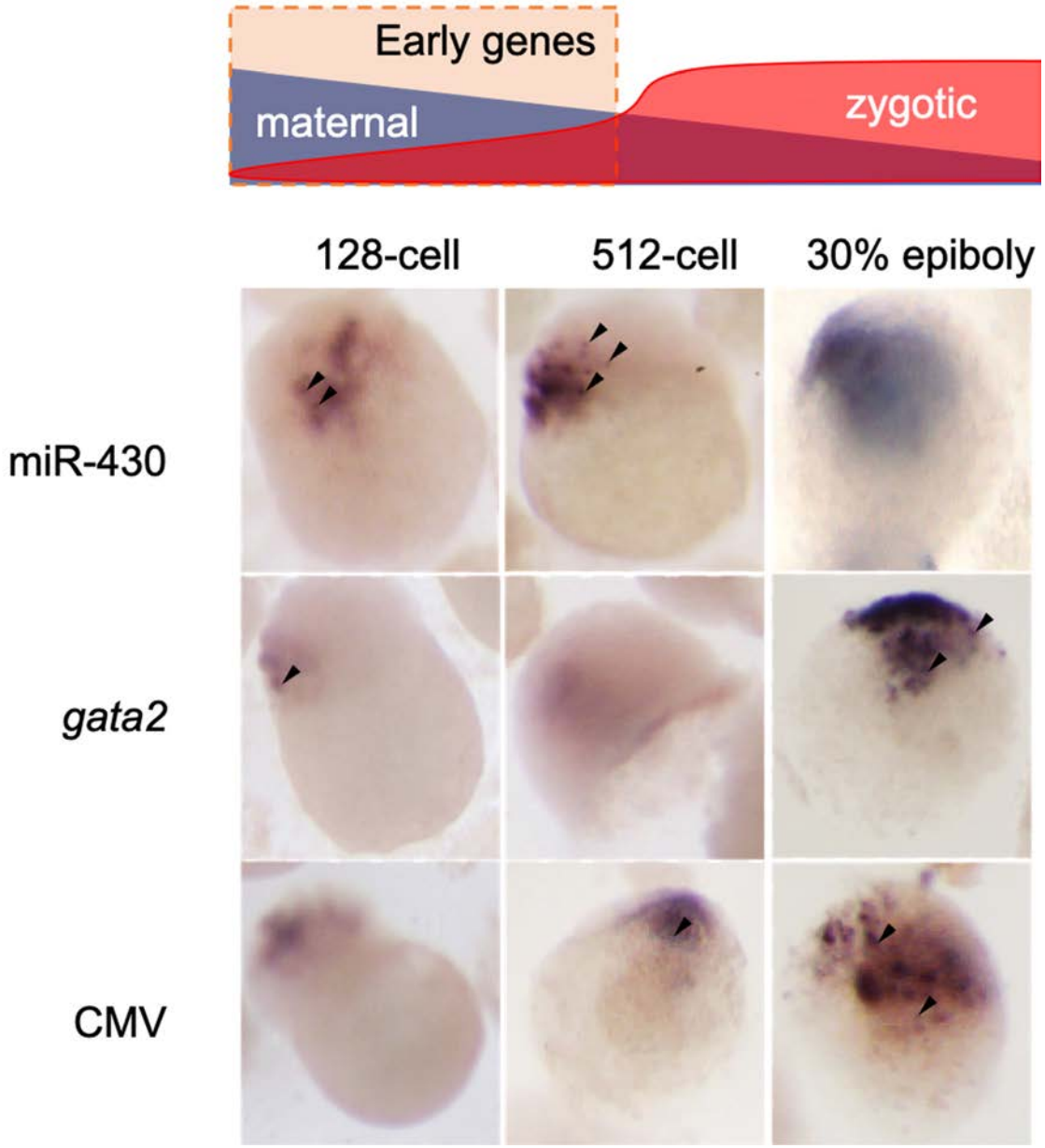


Figure 6-4 Lack of temporal control of transcription from episomal DNA before the main wave of zygotic genome activation

In situ hybridisation demonstrates indiscriminate expression of the transgene by miR-430 promoter and two control promoter sequences: gata2 and CMV via targeting of reporter sequence. Mosaic activity of transgene is indicated by purple colour in blastomeres. Punctate signal indicates nuclear localisations where visible as shown by arrowheads.

	128-cell		512-cell		30% Epiboly	
Promoter	**total**	**% expressing**	**total**	**% expressing**	**total**	**% expressing**
miR-430	**60**	**70**	**52**	**63**	**56**	**71**
gata2	**39**	**62**	**42**	**71**	**31**	**88**
CMV	**48**	**68**	**35**	**82**	**29**	**83**

Table 6-1 Expression frequency of transient injected embryos by WISH

Table shows cumulative amounts of embryos recorded from WISH with percentage score of those in which reporter expression was detected for three different promoter constructs.

The WISH results indicated that all promoters tested are active in transiently transgenic embryos and to a similar frequency (**Figure 6-4** and **Table 6-1**). The results indicate that the embryo is competent to express genes pre-ZGA already at 128-cell stage, denoted from the continuous expression of injected constructs detected by *in-situ* hybridization (**Figure 6-4** and **Table 6-1**). Expression was detected at all stages analysed including pre-ZGA stage (128-cell) and 30% epiboly stage, when endogenous miR-430 is still active, while endogenous *gata2* is not yet active (Maeno et al. 1996). CMV and *gata2* promoter controls not only express earlier than they would endogenously, but also at a time in development when most of the genome is in a state of transcriptional repression. Such a phenomenon has also been observed upon injection of plasmid into *Xenopus* embryos and likely results from episomal DNA lacking yet unexplored chromatin

regulation which is imposed upon the endogenous genes. Indeed such plasmids can be transiently expressed pre-MBT and again reactivated at MBT following silencing, demonstrating the competence of the basal transcriptional machinery pre-MBT (Newport et al. 1982, Prioleau et al. 1994, Almouzni et al. 1995).

The results with transgenes indicate that miR-430 promoter is capable of activation of reporters at the time when miR-430 endogenes are active and thus mimics the endogenous promoter activity. However this result cannot be interpreted as evidence for presence of DNA sequences sufficient for early wave activation as genes, since genes that are otherwise not activated at these early stages were also able to activate the reporter.

6.2.1 Generation of stable transgenic lines with the miR-430 promoter reporter constructs utilising the PhiC31 targeted integration system.

In light of the evidence from the *in transient* studies demonstrating incorrect temporal expression of the transgene in episomal constructs, an alternative transgenesis method is needed to be able to assess temporal control emerging from the DNA sequences embedded in the miR-430 promoter sequence. In order to determine the activation timing for the miR-430 minimal promoter element, the transgene needs to be under the same chromatin control as the endogenous gene. For this aim, integration into a genomic locus needs to be achieved, either *in transient* transgenic or in stable transgenic lines, in which the transgene has been passed through the regulatory mechanism impacting the gametes before fertilisation.

However, for comparative reporter gene analysis the transgenesis system needs to counter position effects. Position effects often considerably affect the activity of the integrated transgene and can manifest as either ectopic enhancement of fluorescent reporter activity or silencing (Jaenisch et al. 1981) due to the landing of the transgene in a regulatory environment of a nearby gene, which then influences activity of the transgene. While position effects are favourable in enhancer trapping (Korzh 2007), they are to be avoided in order to be able to assess regulatory information solely present in the transgene.

The PhiC31 transgenesis model system, which was established in zebrafish by our lab (Roberts et al. 2014), offers targeted integration of a transgene with the purpose of overcoming positional effects resulting from integration into variable chromosomal environments. This transgenesis system has two benefits. Firstly, it allows selection of embryos in which the integration has taken place through a reporter-switching cassette, thus offering analysis of donor transgenes in-sorted for integrants. Secondly, due to integration into a single docking site, the PhiC31 model system reduces drastically the variability between lines created from the same founding docking line, ultimately resulting in reduction of transgenic animals needed in order to confirm correct transgene expression (Rossant et al. 2011, Roberts et al. 2014).

Principally, the PhiC31 system allows targeted integration into predefined genomic location in a transgenic docking line shown in **Figure 6-5**.

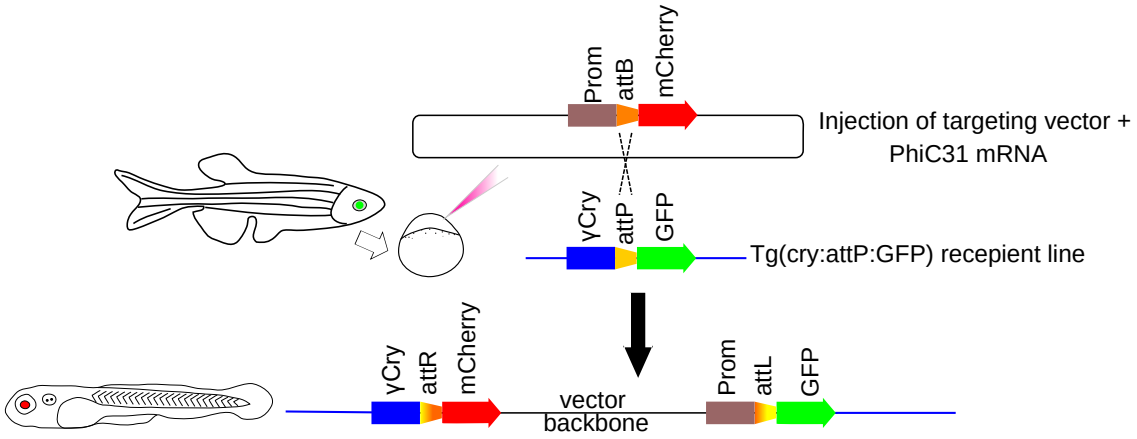


Figure 6-5 Schematic representation of PhiC31 integrase transgenesis system in the zebrafish model organism

Schematic showing injection of promoter-reporter construct into docking line fish. Upon injection of the targeting vector, in combination with an integrase mRNA, the translated PhiC31 integrase protein drives integration into the attP site present in the recipient/docking line. Both attB and attP recombine to form separate attL and attR sites, resulting in a replacement of the gamma crystalline promoter (γCRy) driven GFP reporter in the docking line with an mCherry reporter form the donor plasmids., e.g. leading to lens colour change from green to red as indicated in the schematic, allowing a quick screening for successful integration events. Modified from (Roberts et al. 2014). Yellow bar indicates the mir430 and other promoter sequences to be tested by the integration system.

A docking line *Tg(Xla.Cryg-attP:Gal4VP16;14XUAS:Clover)* based on the UAS/Gal4 has been established in our laboratory to improve signal detection of weakly expressed single-copy transgenes (**Figure 6-6**). In these variant lines an active integrated promoter will drive expression of the Gal4-VP16 fusion protein which in turn, once translated, will bind to the upstream activator sequences (UAS) and drive expression of the green fluorescent reporter.

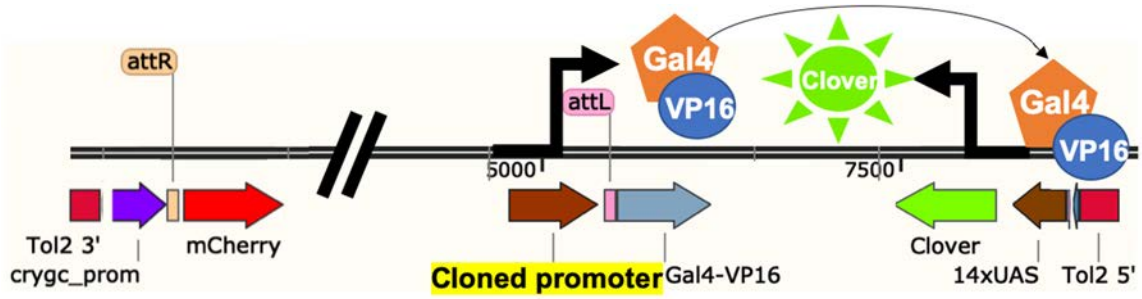


Figure 6-6 transgenic PhiC31 UAS/Gal4 docking lines

Schematic summarising docking site mechanics of the UAS/Gal4 variant PhiC31 docking line. Upon integration, the gamma crystalline promoter (crygc_prom) is displaced by the CRE/promoter module from the donor plasmid, which then will drive expression of the Gal4-VP16 fusion protein instead of the gamma crystalline promoter, once translated it will function as a transcription factor and bind the upstream activator sequences (UAS) to drive expression of the green fluorescent reporter. The mCherry from the donor plasmid is simultaneously placed in front of the gamma crystalline promoter, resulting in lens colour change from green to red allowing screening for successful integration events.

6.2.2 miR-430 promoter does not show activity when integrated in single-copy at an alternative genomic location

In order to explore whether the timing of activation of a minimal promoter element from one of the earliest and highly expressed genes in the zebrafish embryo an experiment was designed to insert the promoter and attached reporter gene outside of its endogenous environment in a stable transgenic line. To ask whether the miR-430 promoter contains autonomous regulatory sequences sufficient to activate the promoter before the main wave of ZGA in a heterologous genomic

environment, a single site integration of the miR-430 minimal promoter element into PhiC31 docking lines were generated. Docking line *Tg(Xla.crygc:attP--Gal4vp16, 14UAS:Clover)UoBL3,* previously generated in our lab (unpublished) was used. The integration site of this line is on Chr17 ~7 kb upstream of *slc16a9a* and ~6kb downstream of *npy4r (*chr17:20640317-20646440, danRer11*)*

Following injection of the miR-430 construct in combination with integrase mRNA to target integration into the docking site of the PhiC31 UAS/Gal4 transgenic docking line, up to 58% (n=71) of injected embryos presented with red lens at day 4, demonstrating a successful integration event, which subsequently were selected and grown to adulthood. From the resulting adult founders (F0) 4 from 35 produced F1 embryos with lens colour (green to red) change. The overall transmission rate was 11.4% (n=35). These F1 generation fish were grown to adulthood. 6 F1 were crossed, and offspring checked for transmission of the integration by analysis of eye colour in larvae. 58% percent of 278 embryos were showing red eye colour indicating transmission of the integration events. In order to determine activity of the stable integrated promoter element, the F1 generation fish were crossed pairwise, and activity of the miR-430 element was probed by detection of Gal4 expression via *in situ* hybridisation analysis of embryos at 256-512 cell and 90%-epiboly-bud stages.

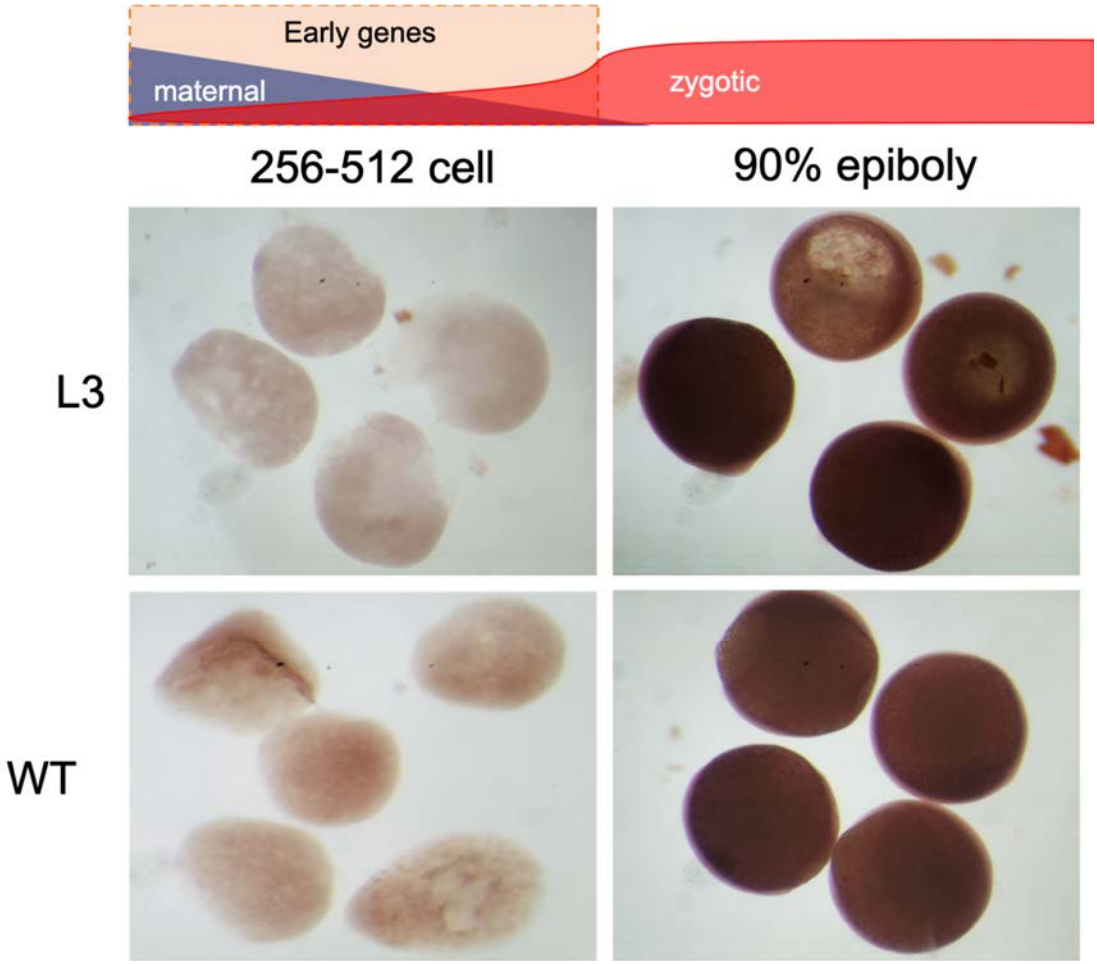


***Figure* 6-7 lack of miR-430 promoter expression in established transgenic lines**

UAS/Gal4 miR-430 promoter stable F1 line and wild type (WT) embryos showed no detectable signal when probed for miR-430 minimal promoter driven Gal4 expression. Pre-MBT stage contained embryos between 256- and 512-cell stages. 90% epiboly stage staining is non-specific inferred by uniformity of staining.

	256-512 cell		30% Epiboly	
Promoter	total	% expressing	total	% expressing
miR-430	78	0	63	0
WT	34	0	37	0

***Table* 6-2 Expression frequency of stable integrated transgenic embryos by WISH**

Table shows cumulative amounts of embryos recorded from WISH against Gal4 expression, with percentage score of those in which reporter expression was detected for miR-430 promoter construct and control non-transgenic / wildtype (WT) embryos.

Upon stable integration into a non-native locus, it was no longer possible to detect any expression from the minimal miR-430 promoter element through WISH analysis (**Figure 6-7** and **Table 6-2**). A functional miR-430 promoter would be expected to drive expression of Gal4-VP16, but apparent lack of Gal4-VP16 in both pre- and post-MBT embryos suggests lack of miR-430 promoter activity when established in this non-native locus, although without a positive control it is not possible to conclude.

6.3 Discussion

In this study, the ability of a promoter sequence of one of the earliest and most highly expressed genes pre-MBT, miR-430 (Heyn et al. 2014), to drive transcription has been demonstrated through transient *in vivo* expression of a reporter on episomal DNA. This result validates, by an independent approach, the conclusions from experiments by (Newport et al. 1982) which demonstrated that genome activation is not due to lack of competence of the early embryo to transcribe, but is instead limited by surplus repressive factors, and therefore depends on the titration of repressive cytoplasmic factors. These factors include histones, that once diluted allow access of basal transcriptional machinery to the core promoter. By injecting excess DNA with promoter capacity, this DNA in the injected embryos may actively compete with repressive factors resulting in premature gene activation (Newport et al. 1982). In line with this observation, all 3 promoters tested in transient transgenic assays showed similar capacity to activate reporter genes before the main wave of genome activation, therefore it is not possible to conclude in these experiments on the ability of the cloned sequences to regulate gene expression before the main wave of ZGA in a genomic integrated environment. Thus, the only conclusion that can be drawn is that the cloned miR-430 sequences carry promoter information, which is interpreted in the embryo similarly to a developmental gene promoter (gata2) and T-A string viral enhancer-promoter sequence (CMV). Notably, all three promoters carry TATA-box and thus it is not clear whether the TATA-box is selectively

required for activation of promoters before the main wave of genome activation, as suggested by the analysis of early acting promoters (Chapter 5).

The transient transgenesis experiments also indicated activation of the miR-430 promoter at prim5 stage, well beyond the developmental stage when this gene is active (end of blastula). This difference is either due to the different chromatin regulation of the episomal and or mosaically integrated miR-430 promoter construct to that of the endogenous miR-430, or it is due to the excessive stability of the fluorescence reporter protein which continues to fluoresce well after transcription from the reporter construct has stopped and the daughter cells and lineages originating from the miR-430 reporter active cells continue to show the reporter activity. It was previously shown that miR-430 locus, alongside the long arm of chromosome four, becomes silent after start of gastrulation and becomes constitutively heterochromatinised (White et al. 2017). It is conceivable that the episomal miR-430 promoter reporter construct does not go through this chromatin repression, either in episomal or random integration sites, which may occur in the injected embryos. To determine whether the miR-430 promoter and locus are autonomously acquiring repressed chromatin state or alternatively, are subjected to repression due to the globally repressive environment of the highly repetitive long arm of chr 4, analysis of chromatin states in stable transgenic lines which carry the miR-430 reporter on heterologous chromosomes must be performed.

The reporter expression analysis of a stable transgenic line carrying the miR-430 promoter linked to the reporter indicated no detectable expression in 141 embryos analysed (**Table 6-2**, **Figure 6-7**). This result is in line with a model that miR-430 is not only regulated by the miR-430 promoter, but also additional factors. These factors may include, i) high copy number of the otherwise weak promoter; ii) additional regulatory elements and transcription factor binding sites not present on the cloned promoter; and iii) a local chromatin environment present in the genomic context of the miR-430 locus that is not then autonomously present on the cloned mir-430 promoter. However, the current experiments are not fully conclusive on the sufficiency of miR-430 promoter as they are based on a single transgenic line and a single integration site in the genome. More repeats are needed with independent transgenic lines and using additional chromosomal locations. Although it is clear from the activity of the γCry promoter driving expression of red fluorescent protein in the lens at day 4 that the integration site is conducive for gene expression, it is not wholly sufficient to conclude that the cloned miR-430 promoter is not sufficient to activate gene expression outside of its endogenous environment until further lines are generated.

It is possible, that the chr4 long arm provides a unique environment amenable to transcription pre-MBT (Howe et al. 2013, White et al. 2017) with a possible dependence on high density of TF binding sites present at the gene cluster (see Chapter 5), as it is already known that miR-430 activation is dependent on maternal stem cell factors Nanog, Pou5f3 and SoxB1 for activation endogenously

(Lee et al. 2013). As such, due to the high density of promoters and binding sites for such factors (**Figure 5-7**) stemming from the multicopy nature of the gene cluster, the miR-430 locus may provide a permissible environment for activation in comparison to a single promoter copy placed into a non-native locus. To test this, a multicopy fragment of miR-430 genes, encompassing a large segment or all of the copies at the locus, should be used in a reporter linked transgene construct (for example by BAC transgenesis) and inserted in a heterologous chromosomal context and analysed for reporter activity. Such an experiment would indicate if multicopy promoters are sufficient to activate pre-MBT gene expression due to enrichment of transcription factor binding sites of the miR-430 promoters.

Another possible explanation for the lack of promoter expression upon single-copy integration into a non-native environment is that other activating, enhancer-like features may be present, either within the underlying triplet structure or in the surroundings of the miR-430 locus, not present on the transgene. For example, repetitive elements, such as LTRs which are suitably enriched on the long arm of chr4 (Howe et al. 2013), have been shown to play a role in genome activation in mammals (Macfarlan et al. 2012, Grow et al. 2015). As a result of the multicopy nature of the locus, each promoter may even act both as a promoter and enhancer, facilitating transcription locally of surrounding clusters. An extended reporter construct containing the current cloned promoter, together with the triplets, will allow transgenic testing of the potential enhancer effect of local sequences around the promoter, while the potential longer-range impact of

sequences from transcribed repeats and other repeat sequences on the miR-430 locus will not be possible with current technologies due to the excessively repeated nature of the long-range environment of the miR-430 locus (**Figure 4-1**).

However, it may be possible to test whether the miR-430 locus environment confers the capacity for pre-MBT gene activation with or without the requirement for TATA box, by insertion of a late zygotic gene with a TATA-less promoter and which typically contain sequences, which are regulated by nucleosome positioning (Haberle et al. 2014) into the miR-430 locus and compared to similar insertion of a promoter which does contain a TATA box. This experiment will become possible upon identification of the correct genomic environment of the miR-430 clusters and identification of single copy sequences suitable for targeted insertion by CRISPR insertion transgenesis (Wierson et al. 2020).

Taken together, it is clear that the miR-430 locus hosts many repeating sequences that possess promoter function, but it is not yet clear whether these sequences are able to operate independently to drive gene expression outside of their endogenous environment.

7 SUMMARY AND CONCLUSIONS

7.1 Identification and characterisation of the promoter of the first genes expressed in the zebrafish embryo

In this thesis I aimed to ask how the first genes get activated in the vertebrate embryo and how promoter level regulation contributes to this process. To get insight into promoter level regulation I chose to identify and characterise the promoter of the miR-430 microRNA gene cluster. This gene cluster has previously been shown to be among the first genes activated in zebrafish embryos (Heyn et al. 2014, Chan et al. 2019, Hadzhiev et al. 2019) and required for the clearance of maternal mRNAs deposited into the egg by the mother, in order to prepare the fertilized egg for transition into zygotic transcription control (Giraldez et al. 2006). Given its key role, miR-430 genes are activated before the global activation of the genome and represent an exciting example for the paradigm of first wave of genome activation. Besides miR-430 a small number of other genes are also activated before global genome activation (Heyn et al. 2014) and the mechanism how they escape the global repressive control characteristic before the mid blastula transition is largely unexplored. The understanding of how they get activated will not only provide insight into the awakening of the genome, but will also contribute to understanding the mechanisms of global expression control characteristic before the mid blastula transition not only in zebrafish, but in most anamniote vertebrates and externally developing non-vertebrate models such as Drosophila. Interestingly, the two waves of genome activation are also

demonstrated in amniotes, which do not go through the mid blastula transition, but are activated already at the two cell stage (Yartseva et al. 2015) In mouse, a first wave of gene activation occurs at the early two cell stage and a main wave of genome activation follows in the late phase of two cell stage (Tadros et al. 2009). These two waves show similarities to that seen in anamniotes and raise the likelihood of parallels not only in the biological reasons for two waves of genome activation but also in their distinct activation mechanisms.

To be able to understand how the first genes get activated as a prerequisite, it is important to consider the mechanisms involved in the main wave of genome activation, which has been studied for many years to a far greater extent (Tadros et al. 2009, Yartseva et al. 2015, Vastenhouw et al. 2019). Several models have been suggested for the main wave of genome activation, the main models listed here (for details see introduction sections): limitation of availability of transcription machinery, excess of repressors, timely chromatin remodelling for providing permissive state, and gradual cell cycle lengthening to allow transcription in emerging G phases, which are lacking in the previous fast cell division cycles of the pre MBT cleavage stages.

My hypothesis was that analysis of miR-430 promoter may help in understanding how any of the above mechanisms may act during the first wave of genome activation and how the first wave genes counter the repressive mechanisms, such as limitation of transcription factors, timing of chromatin remodelling or competition of repressive factors.

It was shown by our lab (Hadzhiev et al. 2019) and others (Chan et al. 2019) that miR-430 is activated in a unique nuclear compartment, characterised by dynamic formation in every cell cycle in addition to an extremely large nuclear body depleted in chromatin. The chromatin features high levels of H3K27a mark which is known to be required for early gene activation (Chan et al. 2019), is rich in Ser2p PolII, the transcriptionally active form of PolII and accumulation of nascent RNA. These observations suggested that genes active during the first wave of genome activation are uniquely organised in a nuclear compartment which may provide clues to their early activation.

My work on the miR-430 promoter cluster has led to several key findings:

1. By multimapping analysis of CAGE-seq datasets I was able to verify the transcription start site of miR-430, which paved the way for identifying promoter features. I was able to identify a minimal promoter sequence which was shown by transgenesis experiments to be able to support transcription in transient transgenic zebrafish embryos.
2. My analysis of the reference genome verified previous suggestions that the miR-430 locus assembly is incomplete and is likely truncated and motivated our initiative for long read sequencing, which was necessary for completing the genomic structure.
3. Long read sequencing by nanopore and PACBIO together allowed generation of a corrected miR-430 locus model, which is dramatically larger (>200kb), than previously appreciated, contains an unprecedented

high density of promoters (>120), which together may explain the unique and large transcription body formation associated with the miR-430 locus.

4. By both unique and multimapping of epigenetic data associated with promoter and enhancers (ATAC-seq, post-translational histone modifications), I was able to demonstrate the temporal sequence of chromatin remodelling towards accessibility at the miR-430 locus and compared that to other gene groups activated in distinct time points before, during and after MBT. These comparisons verify previous observations of H3K27ac enrichment at the locus, suggest an unexpected lack of H3K4me3 (while presence of polycomb repression associated H3K27me3) at miR-430 in comparison to other first wave genes, which are typically bivalent and carry both marks and show a distinctly and widely open chromatin spanning the mir430 cluster.
5. My analysis of transcription factor binding sites verify previous predictions that the stem cell factors Pou5f1 (Pou5f3), Sox2 (Sox19b, SoxB1) and Nanog likely regulate the locus and argue for these factors, which are maternally deposited, to also regulate its activity before the main wave of genome activation.
6. My analysis of the promoter architecture and positionally constrained motifs suggest that miR-430 cluster is regulated by a TATA-box dependent mechanism. In addition, I showed that the first wave of genome activation is generally characterised by TATA box and W-box containing promoters, which is clearly distinct from the main wave of genome activation, mostly

characterised by TATA-less promoter architecture and regulation by nucleosome positioning signals.

The above findings need to be considered when models for the first wave of genome activation are built. Out of the four mechanisms listed above, my findings may have relevance to several of them:

7.2 Promoters of first genes to be activated in the embryo are distinct from other developmental classes

In chapters 4 and 5 it was demonstrated that genes active during zygotic genome activation have a distinct promoter architecture, making them divergent from other genes active post-ZGA, more similar to type 1 genes associated with adult terminally differentiated peripheral tissue (Lenhard et al. 2012, Haberle et al. 2016, Haberle et al. 2018). The promoter profile for first wave genes is highly enriched for TATA element, as was described in Drosophila (Chen et al. 2013), distinct from genes active at the MBT which, in contrast, are enriched for maternal W-box with subsequent broad initiation pattern of transcription. This observation is in contrast to the expectation that these promoter types would be predominately type 3 promoters; CGI enriched with broad pattern of initiation, although they do have some common features associated with genes that are developmentally regulated such as enrichment in both H3K4me3 and H3K27me3 histone modifications, which in ESCs are implicated in creation of bivalent

domains (Bernstein et al. 2006) suggesting these genes are polycomb targets (Mikkelsen et al. 2007).

The TATA-box enrichment in first wave promoters may point at regulation by TBP family proteins. However, it has been demonstrated that TBP is rate limiting and is not yet translated until MBT and unlikely to be present at 64 cell stage as protein to regulate the first wave genes. However, a maternally deposited homolog of TBP, called TBP2 /TRF3 or TBPL2 is known to be present as maternal protein and binds the TATA box (Bartfai et al. 2004) and thus may access the miR-430 promoter. This notion would argue, that among the 4 mechanisms of main wave genome activation, transcription factor accessibility (namely TBPL2) may hold one of the keys for early activation of miR-430 locus and other first wave genes carrying a TATA box.

7.3 Unique multicopy environment may facilitate miR-430 activation as first gene to be expressed in the zebrafish embryo

Divergence of the miR-430 promoter from other developmental genes active in the embryo, characterised by the comparatively low levels of H3K4me3 present at the locus is likely due to a lack of enrichment in CGIs within the core promoter. H3K4me3 is known to facilitate stabilisation of TAF3 and thus supports assembly of the PIC (Vermeulen et al. 2007, Lauberth et al. 2013), yet it is not required at

this promoter which may suggest binding of distinct promoter recognition complexes.

Overall, the epigenetic profiles around the miR-430 promoter: such as a nucleosome free region upstream of the TSS, subsequent H3K4me3 and H3K27ac signal downstream, predicted binding sites of stem cell factor and known regulator of miR-430 expression Nanog (Lee et al. 2013, Leichsenring et al. 2013, Gagnon et al. 2018), transient activity of the miR-430 reporter construct *in vivo* and the substantial amount of CAGE tag TSSs detected are all characteristic features of a functional promoter.

Furthermore, the multimapping results on the revised miR-430 cluster suggest a likely arrangement where we predict that each promoter is marked by promoter-associated post-translational histone modification marks with the key defining feature of the locus seemingly to be in the high density of miR-430 gene units, ~120 copies are present in the reassembled miR-430 cluster. Consequently it is possibly the combination of open chromatin, high marking for active regions (H3K27ac) and the multicopy nature of the cluster that provides a platform for regulatory factors to bind while also having the potential to form a local high concentration of TFs which may lead to overcoming the global transcriptional repression. Indeed Chan et al. (2019) suggest miR-430 is the first gene to be active in the zebrafish embryo, with its unique structure not dissimilar to that described of a super enhancer (SE). It has been suggested that early active genes may be situated in unique environments within the genome, forming local

SEs that facilitate mass utilisation of localized *cis*-regulatory elements (Whyte et al. 2013), promoting higher transcriptional output in regions containing clusters of similar genes. Thus, the miR-430 locus with its high density of promoters may escape the global repressive environment of excess histones (James et al., 2018), which are deposited into the egg by the mother in high concentration, by competing with histones and providing higher affinity environment for binding by rate limiting transcription factors.

Finally, cell cycle lengthening has been suggested as a regulator which allows the main wave of genome activation by permitting the emergence of G-phase in the otherwise extremely fast and synchronous cleavage cell cycles composed of M and S phases. While no direct evidence exists to suggest that the miR-430 locus is replicated distinctly from the rest of the genome, it is conceivable that similar to repeat sequences of the long arm of chr4, which contain the miR-430 locus and are known to be late replicating (Siefert et al. 2017), a distinct and late timing of replication may offer the miR-430 locus more time for transcription in the otherwise extremely fast cell cycles between 64-512 stages, when the first wave of genome activation takes place, providing a unique environment amenable to its early activation.

8 APPENDIX

8.1 Public NGS dataset used in the analyses

Dataset	**BioProject/Accession**	**Article DOI**
H3K4me3 and H3K27me3 ChIP-seq	PRJNA434216	10.1186/s12864-019-5611-7
ATAC-seq	PRJNA395463	10.1101/gr.228833.117
H3K27ac ChIP-seq	PRJNA473799	10.1016/j.molcel.2018.10.017
Nanog ChIP-seq	PRJNA156233	10.1016/j.devcel.2012.01.003
Pou5f1 and Sox2 ChIP-seq	PRJNA171706	10.1126/science.1242527
CAGE-seq	PRJNA169500	10.1101/gr.153692.112
RNA-seq	PRJEB12982	10.7554/eLife.30860.001

SRA Run IDs for all public NGS datasets used in this study

SRA Run ID	Assay Type	Stage
SRR6476749,SRR6476750,SRR6476751	ATAC-seq	64-cell
SRR5860459,SRR5860460,SRR5860461	ATAC-seq	256-cell
SRR5860462,SRR5860463,SRR5860464	ATAC-seq	1k-cell
SRR5860465,SRR5860466,SRR5860467	ATAC-seq	oblong
SRR5860468,SRR5860469,SRR5860470	ATAC-seq	dome
SRR516545,SRR516546	CAGE-seq	Unfertilized egg
SRR516547	CAGE-seq	Fertilized egg
SRR516548	CAGE-seq	64-cell
SRR516549	CAGE-seq	512-cell
SRR516550	CAGE-seq	High
SRR516551	CAGE-seq	Oblong
SRR516553,SRR516552	CAGE-seq	Sphere
SRR516554	CAGE-seq	30%-epiboly
SRR516555,SRR516556	CAGE-seq	Shield
SRR516557,SRR516558	CAGE-seq	14-somites
SRR516559	CAGE-seq	Prim-5
SRR516561,SRR516560	CAGE-seq	Prim-20
SRR6729442,SRR6729443	H3K4me3 ChIP-seq	2-cell
SRR6729444,SRR6729445	H3K4me3 ChIP-seq	16-cell
SRR6729446,SRR6729447	H3K4me3 ChIP-seq	128-cell
SRR6729448,SRR6729449	H3K4me3 ChIP-seq	1k-cell
SRR6729453,SRR6729454	H3K27me3 ChIP-seq	2-cell
SRR6729455,SRR6729456	H3K27me3 ChIP-seq	16-cell
SRR6729458	H3K27me3 ChIP-seq	128-cell
SRR6729459,SRR6729460	H3K27me3 ChIP-seq	1k-cell
SRR7235547,SRR7235548	H3K27ac ChIP-seq	4-cell
SRR7235554,SRR7235555,SRR7235556	H3K27ac ChIP-seq	256-cell
SRR7235529,SRR7235530,SRR7235531	H3K27ac ChIP-seq	Dome
SRR908306	Pou5f1-Flag ChIP	512-cell
SRR530642,SRR530643	Pou5f1-Flag ChIP	30%-epiboly
SRR530641	Sox2 ChIP-seq	30%-epiboly
SRR530644	WCE Input ChIP-seq (Contol for Pou5f1 and Sox2)	30%-epiboly
SRR392588	Nanog-Myc ChIP-seq	Oblong
SRR392589	Nanog-Myc ChIP-seq	Dome
SRR392591	WCE Input ChIP-seq (Contol for Nanog)	Oblong
SRR392592	WCE Input ChIP-seq (Contol for Nanog)	Dome

ERR1442736,ERR1442646,ERR1442920,ERR1442830,ERR1442740, ERR1442650,ERR1442919,ERR1442829,ERR1442739,ERR1442649, ERR1442918,ERR1442828,ERR1442648,ERR1442917,ERR1442827, ERR1442737,ERR1442647,ERR1442916,ERR1442826,ERR1442738	RNA-seq	1-cell
ERR1442807,ERR1442717,ERR1442628,ERR1442897,ERR1442626, ERR1442627,ERR1442896,ERR1442806,ERR1442716,ERR1442718, ERR1442808,ERR1442898,ERR1442629,ERR1442719,ERR1442809, ERR1442899,ERR1442630,ERR1442720,ERR1442810,ERR1442900	RNA-seq	2-cell
ERR1442903,ERR1442634,ERR1442724,ERR1442814,ERR1442904, ERR1442635,ERR1442725,ERR1442815,ERR1442905,ERR1442811, ERR1442813,ERR1442723,ERR1442633,ERR1442902,ERR1442812, ERR1442722,ERR1442632,ERR1442901,ERR1442721,ERR1442631	RNA-seq	128-cell
ERR1442733,ERR1442915,ERR1442825,ERR1442735,ERR1442645 ,ERR1442914,ERR1442824,ERR1442734,ERR1442644,ERR1442913 ,ERR1442823,ERR1442643,ERR1442912,ERR1442822,ERR1442732, ERR1442642,ERR1442911,ERR1442821,ERR1442731,ERR1442641	RNA-seq	1k-cell
ERR1442711,ERR1442621,ERR1442895,ERR1442805,ERR1442715, ERR1442625,ERR1442894,ERR1442804,ERR1442714,ERR1442624, ERR1442893,ERR1442713,ERR1442803,ERR1442801,ERR1442891, ERR1442622,ERR1442712,ERR1442802,ERR1442892,ERR1442623	RNA-seq	Dome
ERR1442835,ERR1442561,ERR1442651,ERR1442745,ERR1442741, ERR1442831,ERR1442562,ERR1442652,ERR1442742,ERR1442832, ERR1442563,ERR1442653,ERR1442743,ERR1442833,ERR1442564, ERR1442654,ERR1442744,ERR1442834,ERR1442565,ERR1442655	RNA-seq	50pc-epiboly
ERR1442783,ERR1442875,ERR1442785,ERR1442695,ERR1442605, ERR1442874,ERR1442784,ERR1442694,ERR1442604,ERR1442873, ERR1442691,ERR1442693,ERR1442603,ERR1442872,ERR1442782, ERR1442692,ERR1442602,ERR1442871,ERR1442781,ERR1442601	RNA-seq	Shield

ERR1442606,ERR1442876,ERR1442609,ERR1442607,ERR1442697, ERR1442787,ERR1442877,ERR1442608,ERR1442698,ERR1442788, ERR1442878,ERR1442696,ERR1442699,ERR1442789,ERR1442879, ERR1442610,ERR1442700,ERR1442790,ERR1442880,ERR1442786	RNA-seq	75pc-epiboly
ERR1442636,ERR1442910,ERR1442820,ERR1442730,ERR1442640, ERR1442909,ERR1442819,ERR1442729,ERR1442639,ERR1442908, ERR1442818,ERR1442728,ERR1442638,ERR1442907,ERR1442817, ERR1442727,ERR1442637,ERR1442906,ERR1442816,ERR1442726	RNA-seq	1-4-somites
ERR1442611,ERR1442705,ERR1442613,ERR1442615,ERR1442884, ERR1442794,ERR1442704,ERR1442614,ERR1442883,ERR1442793, ERR1442703,ERR1442885,ERR1442882,ERR1442792,ERR1442702, ERR1442612,ERR1442881,ERR1442791,ERR1442701,ERR1442795	RNA-seq	14-19-somites
ERR1442616,ERR1442706,ERR1442796,ERR1442886,ERR1442617, ERR1442707,ERR1442797,ERR1442887,ERR1442618,ERR1442708, ERR1442798,ERR1442888,ERR1442619,ERR1442709,ERR1442799, ERR1442889,ERR1442620,ERR1442710,ERR1442800,ERR1442890	RNA-seq	20-25-somites
ERR1442659,ERR1442749,ERR1442839,ERR1442570,ERR1442660, ERR1442750,ERR1442840,ERR1442746,ERR1442566,ERR1442569, ERR1442838,ERR1442748,ERR1442658,ERR1442568,ERR1442837, ERR1442747,ERR1442657,ERR1442567,ERR1442836,ERR1442656	RNA-seq	Prim-5
ERR1442571,ERR1442755,ERR1442845,ERR1442665,ERR1442575, ERR1442844,ERR1442754,ERR1442664,ERR1442574,ERR1442843, ERR1442753,ERR1442663,ERR1442573,ERR1442842,ERR1442752, ERR1442662,ERR1442572,ERR1442841,ERR1442751,ERR1442661	RNA-seq	Prim-15
ERR1442576,ERR1442666,ERR1442760,ERR1442850,ERR1442756, ERR1442846,ERR1442577,ERR1442667,ERR1442757,ERR1442847, ERR1442578,ERR1442668,ERR1442848,ERR1442579,ERR1442669, ERR1442759,ERR1442849,ERR1442580,ERR1442670,ERR1442758	RNA-seq	Prim-25

ERR1442581,ERR1442671,ERR1442761,ERR1442851,ERR1442582, ERR1442672,ERR1442762,ERR1442852,ERR1442583,ERR1442673, ERR1442853,ERR1442584,ERR1442674,ERR1442764,ERR1442854, ERR1442585,ERR1442675,ERR1442765,ERR1442855,ERR1442763	RNA-seq	Long-pec
ERR1442680,ERR1442770,ERR1442590,ERR1442859,ERR1442769, ERR1442679,ERR1442589,ERR1442858,ERR1442768,ERR1442678, ERR1442588,ERR1442857,ERR1442767,ERR1442677,ERR1442587, ERR1442856,ERR1442766,ERR1442676,ERR1442586,ERR1442860	RNA-seq	Protruding-mouth
ERR1442591,ERR1442771,ERR1442681,ERR1442865,ERR1442775, ERR1442685,ERR1442595,ERR1442864,ERR1442774,ERR1442684, ERR1442594,ERR1442773,ERR1442683,ERR1442593,ERR1442862, ERR1442772,ERR1442682,ERR1442592,ERR1442861,ERR1442863	RNA-seq	Day-4
ERR1442868,ERR1442599,ERR1442689,ERR1442779,ERR1442869,ERR1442600,ERR1442690,ERR1442780,ERR1442686,ERR1442870,ERR1442778,ERR1442688,ERR1442598,ERR1442867,ERR1442777,ERR1442687,ERR1442597,ERR1442866,ERR1442776,ERR1442596	RNA-seq	Day-5

8.2 Morpholino oligonucleotides

Name	**Sequence**
miR430_MO1	CACACGCATCTTGTTGTCTGCTGTT
miR430_M02	GCAGCAAGAGAAATCATTGGCAGGC
miR430_MO3	AGCAGATTAAAACCTTGTCCGTTCA
miR430_M01_SENSE	AACAGCAGACAACAAGATGCGTGTG
miR430_M02_SENSE	GCCTGCCAATGATTTCTCTTGCTGC
miR430_M03_SENSE	TGAACGGACAAGGTTTTAATCTGCT

8.3 Position weight matrices used (from JASPAR database)

ID	Name
MA0108.1	TBP/TATA-box
MA0143.3	Sox2
MA0142.1	Pou5f1::Sox2
MA0039.2	Klf4
UN0383.1	Nanog

8.4 First Wave gene list

Expression levels based on EU-RNA-seq and fold change from 128 Cell to 512 Cell from Heyn et al. (2014)

ensGeneID	geneName	128Cell	256Cell	512Cell	foldChange
ENSDARG00000037178	zic2b	0	0.62	10.49	3.52
ENSDARG00000030703	otx2a	0.02	0.54	6.02	2.79
ENSDARG00000036670	aplnrb	1.31	31.24	233.5	6.66
ENSDARG00000061120	slc43a2b	0.28	1.21	5.52	2.35
ENSDARG00000027744	gadd45ba	0.13	3.34	13.67	3.7
ENSDARG00000059327	gata2a	0.11	1.67	12.32	3.59
ENSDARG00000045219	dkk1b	0.44	5.86	16.59	3.61
ENSDARG00000015906	mxtx2	2.98	119.19	496.49	6.96

ENSDARG00000021201	noto	0.18	8.46	48.69	5.4
ENSDARG00000039269	arg2	0.07	0.55	3.28	2
ENSDARG00000008796	her5	0.74	19.46	77.55	5.5
ENSDARG00000075666	tsc22d3	0.05	0.66	5.55	2.65
ENSDARG00000002707	her11	0.61	39.2	180.22	6.82
ENSDARG00000058917	nnr	0.05	1.7	25.12	4.64
ENSDARG00000040955	dharma	0.34	5.31	19.38	3.92
ENSDARG00000006110	chrd	0.01	0.6	4.68	2.49
ENSDARG00000053291	pnrc2	11.14	54.49	174.96	3.86
ENSDARG00000061773	neurl1ab	0.27	1.18	4.29	2.06
ENSDARG00000078552	grhl3	0.57	26.11	236.65	7.24
ENSDARG00000002353	tagapa	0.01	0.61	6.95	2.98

ENSDARG00000044059	lft2	0.34	12.21	57.61	5.45
ENSDARG00000059073	gsc	0.18	2.28	12.6	3.53
ENSDARG00000077712	zgc:113886	10.88	342.99	968.93	6.35
ENSDARG00000055043	depdc7a	0.06	0.96	11.78	3.59
ENSDARG00000012788	foxa3	0.01	0.29	6.81	2.95
ENSDARG00000040503	sb:cb81	0.14	2.75	54.12	5.6
ENSDARG00000022280	brd2a	5.65	54.57	153.83	4.54
ENSDARG00000033182	arl4ab	0.04	1.22	7.14	2.97
ENSDARG00000038792	klf17	2.58	108.16	360.23	6.66
ENSDARG00000040432	klf2b	0.11	6.71	57.53	5.72
ENSDARG00000016253	ralab	0.03	2.79	14.17	3.88
ENSDARG00000038569	fzd8b	0.18	2.7	9.3	3.12

ENSDARG00000025372	admp	0.17	6.14	47.35	5.37
ENSDARG00000018693	cdh2	0.27	1.06	8.06	2.84
ENSDARG00000015611	rasl11b	0.25	4.81	13.17	3.5
ENSDARG00000033539	paics	0.51	1.42	7.92	2.56
ENSDARG00000055678	akap12b	0.03	1.07	22.08	4.48
ENSDARG00000019920	lft1	0.17	6.85	19.8	4.15
ENSDARG00000031683	fosab	0.17	8.68	33.14	4.87
ENSDARG00000094859	si:dkey-11o15.10	0.07	2.09	32.32	4.97
ENSDARG00000057096	ndr1	0.36	7.74	23.39	4.16
ENSDARG00000069105	fgfr4	0.01	0.3	3.56	2.18
ENSDARG00000040141	si:ch73-138e16.5	5.22	153.69	421.93	6.09
ENSDARG00000043126	blf	0.48	40.88	186.5	6.98

ENSDARG00000009881	ier5	0.05	2.57	16.06	4.02
ENSDARG00000076554	cdkn1a	0.1	1.25	4.2	2.24
ENSDARG00000076264	zgc:195170	0	4.8	15.67	4.06
ENSDARG00000043180	gpd1b	0.04	0.96	7.39	3.02
ENSDARG00000087844	plekhn1	0.01	0.46	9.33	3.35
ENSDARG00000002601	irx7	2.11	69.27	276.05	6.48
ENSDARG00000045444	fzd8a	0.05	2.29	12.27	3.66
ENSDARG00000059279	tfap2a	0.2	0.73	8.94	3.05
ENSDARG00000059337	tgif1	3.38	6.56	16.58	2
ENSDARG00000112131	zgc:173552	0.94	4.78	9.72	2.47
ENSDARG00000070914	dusp6	0.19	12.76	59.76	5.67
ENSDARG00000089042	foxb1a	0.03	3.93	19.06	4.28

ENSDARG00000038141	atf4b	0.11	2.43	14.81	3.83
ENSDARG00000009040	si:ch73-141c7.1	0.08	1.89	19.98	4.28
ENSDARG00000045898	si:ch211-152c2.3	0	2.54	59.63	5.92
ENSDARG00000038386	ascl1a	0	1.3	5.94	2.8
ENSDARG00000032765	net1	0.17	3.7	56.81	5.63
ENSDARG00000027249	btg1	0.44	4.36	18.89	3.79
ENSDARG00000016526	gata3	0.34	24.27	71.27	5.76
ENSDARG00000086296	znf1137	0.01	2.1	13.91	3.88
ENSDARG00000041299	zgc:110796	0.66	2.37	10.36	2.78
ENSDARG00000035147	lrwd1	0.23	8.72	33.4	4.8
ENSDARG00000010770	sox19a	0.05	5.25	43.12	5.4
ENSDARG00000055647	ftr82	0.02	0.58	5.31	2.62

ENSDARG00000078004	drl	0.02	2.26	17.21	4.16
ENSDARG00000020893	slc25a55a	0.28	12.13	67.34	5.74
ENSDARG00000077473	mych	0.67	4.3	32.78	4.34
ENSDARG00000040606	tfap2c	0.04	1.19	4.64	2.44
ENSDARG00000058179	ackr3b	0.01	0.92	23.15	4.57
ENSDARG00000088835	efs	0.04	2.44	9.07	3.28
ENSDARG00000006487	anp32a	2.68	9.8	39.7	3.47
ENSDARG00000042485	foxd5	0.21	16.56	78.5	6.04
ENSDARG00000076857	her3	0.1	1.15	9.28	3.22
ENSDARG00000026845	rhoaa	0.11	0.87	4.66	2.36
ENSDARG00000060004	fzd7a	0.06	0.46	5.21	2.54
ENSDARG00000056561	asb11	0.18	3.91	56.85	5.62

8.5 Main Wave gene list

ensGeneID	geneName
ENSDARG00000059751	abtb2a
ENSDARG00000004094	acsbg2
ENSDARG00000056250	add3b
ENSDARG00000075819	arhgef2
ENSDARG00000116732	CABZ01031870.1
ENSDARG00000069185	celsr1a
ENSDARG00000009544	cldnb
ENSDARG00000010032	cmtm7
ENSDARG00000030449	crabp2b
ENSDARG00000011701	ctsll
ENSDARG00000000943	dzip1
ENSDARG00000008986	e2f7
ENSDARG00000099960	elovl1a
ENSDARG00000003313	eps8l1b
ENSDARG00000011245	esrp1
ENSDARG00000075564	fam13a
ENSDARG00000003399	fgf8a
ENSDARG00000067524	fut9b
ENSDARG00000017821	gata5
ENSDARG00000103589	gata6
ENSDARG00000099329	h2afy

ENSDARG00000006615	hnf1ba
ENSDARG00000054842	hsd17b14
ENSDARG00000020541	ism1
ENSDARG00000013056	itpk1a
ENSDARG00000104665	kirrel3l
ENSDARG00000019365	krt99
ENSDARG00000079790	LO018102.1
ENSDARG00000062552	lpar6a
ENSDARG00000006409	mapk12b
ENSDARG00000086815	MYADM
ENSDARG00000094965	nfil3-5
ENSDARG00000044685	nr0b2a
ENSDARG00000078256	ovol1b
ENSDARG00000033537	p4ha1a
ENSDARG00000023181	pcp4l1
ENSDARG00000022807	pitpnc1b
ENSDARG00000036194	pitx2
ENSDARG00000051861	pkp3a
ENSDARG00000057575	pnp4a
ENSDARG00000062986	pnpla7a
ENSDARG00000012044	polr3gla
ENSDARG00000058800	rab25a
ENSDARG00000045628	rab34a

NR_120315.1	wu:fc47e11
ENSDARG00000099970	CR383676.1
ENSDARG00000012248	rgma
ENSDARG00000099746	rgs3a
ENSDARG00000006677	ripk1l
ENSDARG00000004218	rnd1b
ENSDARG00000088143	sema4gb
ENSDARG00000103000	si:ch211-107p11.3
ENSDARG00000078138	si:ch211-202h22.8
ENSDARG00000100842	si:ch211-223a21.4
ENSDARG00000054723	si:ch211-242b18.1
ENSDARG00000101829	si:dkey-16b10.2
ENSDARG00000103471	si:dkey-199m13.4
ENSDARG00000079403	si:dkey-204l11.1
ENSDARG00000098293	si:dkey-27i16.2
ENSDARG00000100479	si:dkey-43f9.4
ENSDARG00000101320	si:dkey-50i6.5
ENSDARG00000106749	si:dkey-7i4.13
ENSDARG00000097091	si:dkey-7j22.2
ENSDARG00000092141	si:dkeyp-100a1.6
ENSDARG00000103522	si:zfos-943e10.1
ENSDARG00000005463	slc30a1a
ENSDARG00000018149	slc38a4

ENSDARG00000098954	stard13b
ENSDARG00000035694	stm
ENSDARG00000038428	sulf1
ENSDARG00000033466	tagln2
ENSDARG00000025920	tiam1b
ENSDARG00000087804	tnfrsf11a
ENSDARG00000043442	zgc:153665
ENSDARG00000002509	zgc:153911
ENSDARG00000089875	zgc:173705
ENSDARG00000079035	zgc:174573
ENSDARG00000071497	zic3
ENSDARG00000074009	znf1059
ENSDARG00000101245	znf1063

8.6 Constitutive gene list

ensGeneID	geneName
ENSDARG00000014727	acox1
ENSDARG00000070076	actr2b
ENSDARG00000041199	alg12
ENSDARG00000009484	arf1
ENSDARG00000018190	asna1
ENSDARG00000036956	asxl1
ENSDARG00000014531	atg4a
ENSDARG00000068601	atmin
ENSDARG00000036135	bbox1
ENSDARG00000094551	BX855614.1
ENSDARG00000058869	CABZ01040556.1
ENSDARG00000013705	ccm2
ENSDARG00000020201	cdc73
ENSDARG00000076659	cdca7b
ENSDARG00000061357	chst6
ENSDARG00000007135	cnot3a
ENSDARG00000043658	cxadr
ENSDARG00000016016	dbnlb
ENSDARG00000041137	dhrs13a.3
ENSDARG00000010745	dnaja2b
ENSDARG00000095623	dpm2

ENSDARG00000013931	eif3m
ENSDARG00000070046	eny2
ENSDARG00000071449	fam234a
ENSDARG00000038024	flr
ENSDARG00000028327	gmfb
ENSDARG00000053326	gna11b
ENSDARG00000005840	got2b
ENSDARG00000019236	gsr
ENSDARG00000016038	hacd3
ENSDARG00000060594	hadhab
ENSDARG00000067741	itpkcb
ENSDARG00000009493	MAP3K13
ENSDARG00000002659	mapre1b
ENSDARG00000059581	mbd5
ENSDARG00000014148	med22
ENSDARG00000006253	mier3a
ENSDARG00000079067	mms19
ENSDARG00000024381	msto1
ENSDARG00000063690	nrm
ENSDARG00000044565	ola1
ENSDARG00000031981	pcbd1
ENSDARG00000018902	pes
ENSDARG00000028322	pex14

ENSDARG00000035043	pfdn5
ENSDARG00000010773	pnpla6
ENSDARG00000045175	pofut2
ENSDARG00000057151	prep
ENSDARG00000076718	prkra
ENSDARG00000057026	ran
ENSDARG00000073665	rasa1b
ENSDARG00000078966	rbm15b
ENSDARG00000002020	rictorb
ENSDARG00000042751	riok3
ENSDARG00000036929	rnf139
ENSDARG00000034291	rpl37
ENSDARG00000010487	sae1
ENSDARG00000045885	scaf11
ENSDARG00000000779	scfd1
ENSDARG00000078267	si:dkey-181m9.8
ENSDARG00000051836	si:dkeyp-19e1.3
ENSDARG00000093656	si:dkeyp-2e4.6
ENSDARG00000015158	slc50a1
ENSDARG00000061446	smap2
ENSDARG00000039989	snrnp48
ENSDARG00000010563	spopla
ENSDARG00000023309	spryd4

ENSDARG00000017762	srrt
ENSDARG00000003769	tada2b
ENSDARG00000040727	tfb1m
ENSDARG00000056628	tmem170a
ENSDARG00000038239	tnpo2
ENSDARG00000004114	tti1
ENSDARG00000060480	ubr1
ENSDARG00000018192	ubr5
ENSDARG00000044362	unc119b
ENSDARG00000032327	usp36
ENSDARG00000036338	vps11
ENSDARG00000061257	zgc:153240
ENSDARG00000037652	zgc:56409

8.7 Post gastrulation gene list

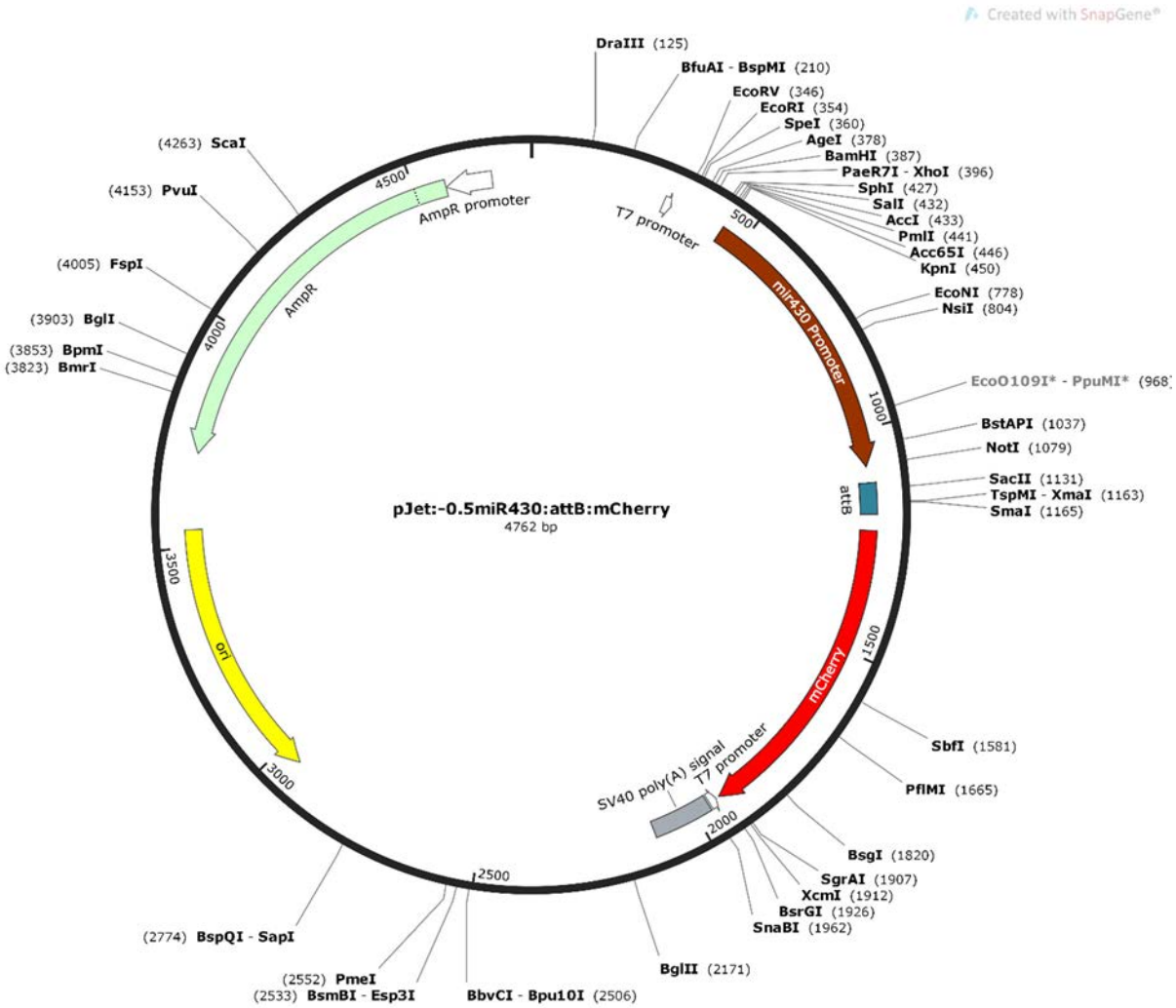
ensGeneID	geneName
ENSDARG00000088524	acot17
ENSDARG00000009196	anxa3a
ENSDARG00000010472	atp1a2a
ENSDARG00000035005	bag2
ENSDARG00000102742	bmpr1bb
ENSDARG00000097984	BX323067.1
ENSDARG00000094614	BX323458.2
ENSDARG00000092881	BX649294.1
ENSDARG00000040917	cbfb
ENSDARG00000052783	cdc42ep3
ENSDARG00000102456	cfhl4
ENSDARG00000014106	cfl2
ENSDARG00000000472	cntn2
ENSDARG00000005762	col14a1a
ENSDARG00000041925	cryba2b
ENSDARG00000078865	cxxc5a
ENSDARG00000021488	cyth1b
ENSDARG00000035602	dao.1
ENSDARG00000044982	dhrs3a
ENSDARG00000004232	dlb

ENSDARG00000088717	ecrg4b
ENSDARG00000021539	ephb2b
ENSDARG00000020741	fga
ENSDARG00000079355	flrt2
ENSDARG00000055398	foxc1b
ENSDARG00000102626	frem2b
ENSDARG00000097057	fxyd6l
ENSDARG00000018967	gabbr1a
ENSDARG00000056831	gng2
ENSDARG00000004621	gpm6ab
ENSDARG00000062020	gse1
ENSDARG00000035735	gsx1
ENSDARG00000004904	gys2
ENSDARG00000008305	hand2
ENSDARG00000024160	hao1
ENSDARG00000059231	hephl1a
ENSDARG00000007941	hmx4
ENSDARG00000010630	hoxb6a
ENSDARG00000076919	igdcc4
ENSDARG00000042282	itga6a
ENSDARG00000013168	jag1b
ENSDARG00000053509	kazald3
ENSDARG00000031222	lhx2b

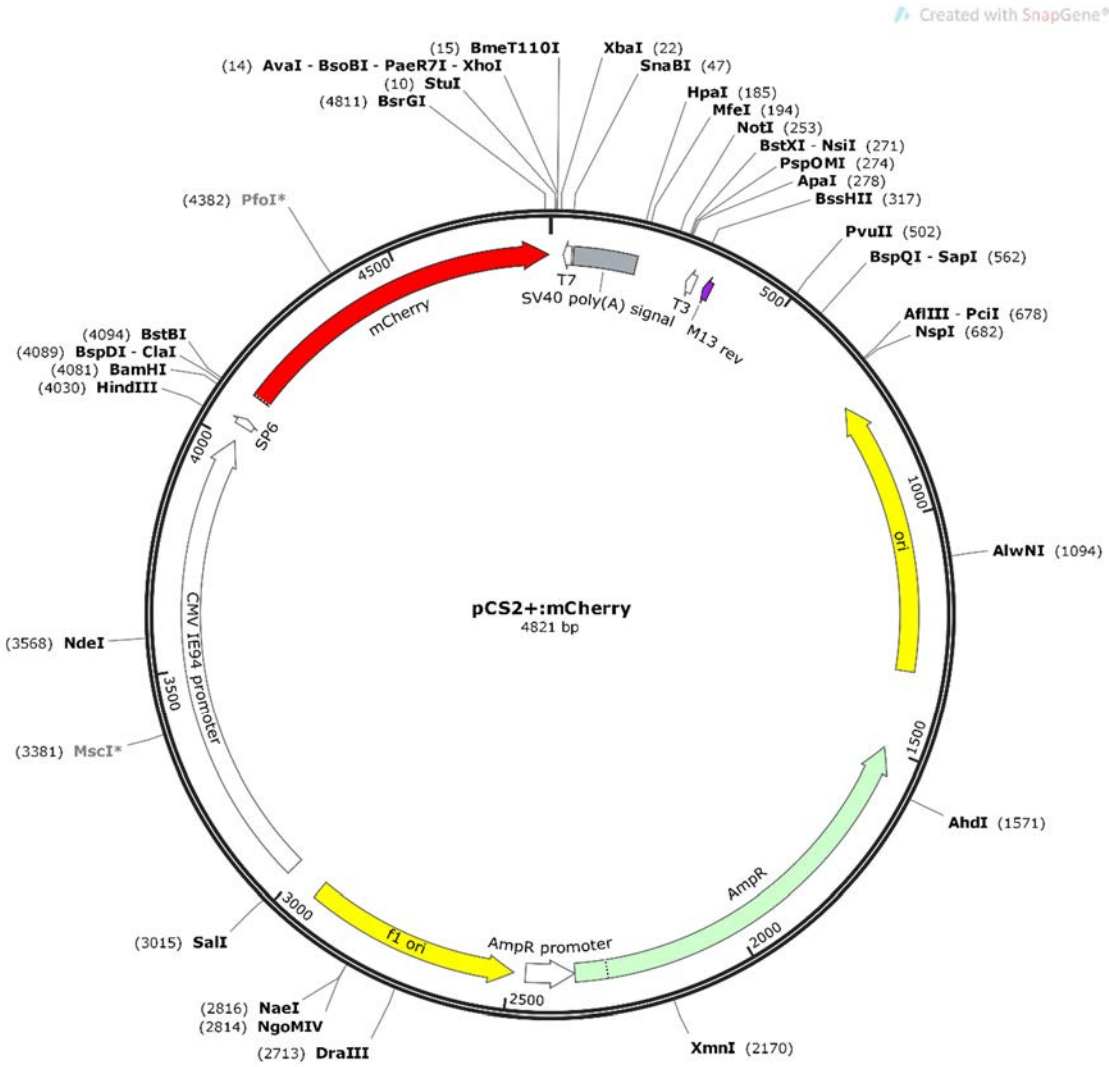
ENSDARG00000052154	lix1
ENSDARG00000015890	mafa
ENSDARG00000074030	myt1a
ENSDARG00000027930	naprt
ENSDARG00000003469	neurod4
ENSDARG00000096546	nrp2a
ENSDARG00000059043	ntf3
ENSDARG00000070818	pax7b
ENSDARG00000018178	pgm2
ENSDARG00000060184	pkn1a
ENSDARG00000062590	pleca
ENSDARG00000039577	ptk2bb
ENSDARG00000092638	ptprua
ENSDARG00000098733	pxdn
ENSDARG00000098645	rgn
ENSDARG00000023797	ryr1b
ENSDARG00000035175	schip1
ENSDARG00000061471	sema5ba
ENSDARG00000075648	sema5bb
ENSDARG00000061383	serpinf2b
ENSDARG00000097523	si:ch73-364h19.1

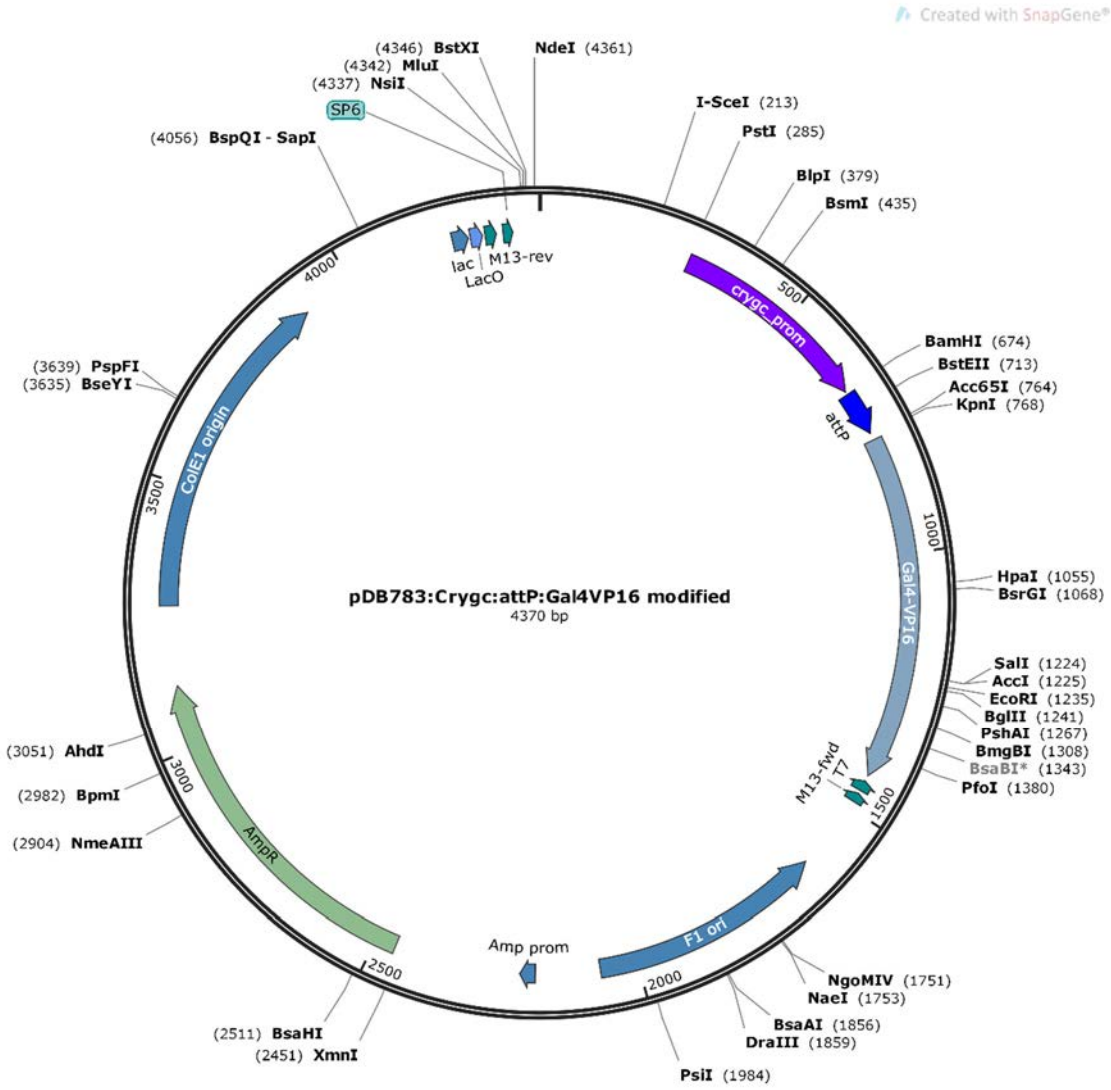
ENSDARG00000100712	si:dkey-19b23.12
ENSDARG00000056519	si:dkey-280e21.3
ENSDARG00000040920	si:dkey-49n23.1
ENSDARG00000000241	slc40a1
ENSDARG00000093064	SMIM18
ENSDARG00000107526	smkr1
ENSDARG00000069866	sox1a
ENSDARG00000074644	spock1
ENSDARG00000101452	SUSD2
ENSDARG00000018025	tbx2a
ENSDARG00000040606	tfap2c
ENSDARG00000100055	tox2
ENSDARG00000086033	tspan5a
ENSDARG00000014796	wnt11r
ENSDARG00000062338	zeb2a
ENSDARG00000093498	zgc:194281

8.8 Plasmid maps




Created with SnapGene®
pDB896:-0.8gata2:mCherry
5554 bases
(5529) BstXI
(5525) MluI
(5520) NsiI
HindIII (198)
EcoRV (206)
SalI - SgrDI (210)
AccI (211)
BamHI (237)
AscI (275)
AsiSI (286)
FseI (295)
BsgI (348)
(5122) PciI
SP6 promoter
M13 rev
TspMI - XmaI (687)
SmaI (689)
(4822) PspFI
(4818) BseYI
(4713) AlwNI
5000
ori
GATA2
1000
(4234) AhdI
(4168) BsaI
(4165) BpmI
(4087) NmeAIII
PspXI (1359)
NcoI (1365)
4000
AmpR
mCherry
Bsu36I (1632)
BbsI (1790)
2000
SV40 poly(A) signal
AmpR promoter
(3634) XmnI
BsrGI (2070)
NotI (2082)
BspDI - ClaI (2092)
T7 promoter
f1 ori
3000
HpaI (2234)
SwaI (2353)
AflII (2567)
PspOMI (2614)
ApaI (2618)
(3042) DraIII
(3039) BsaAI
M13 fwd


Created with SnapGene®
(15) BmeT110I
XbaI (22)
(14) AvaI - BsoBI - PaeR7I - XhoI
SnaBI (47)
(10) StuI
HpaI (185)
(4811) BsrGI
MfeI (194)
NotI (253)
BstXI - NsiI (271)
PspOMI (274)
ApaI (278)
BssHII (317)
(4382) PfoI*
PvuII (502)
BspQI - SapI (562)
4500
T7
SV40 poly(A) signal
T3
M13 rev
500
AflIII - PciI (678)
NspI (682)
mCherry
(4094) BstBI
(4089) BspDI - ClaI
(4081) BamHI
(4030) HindIII
SP6
4000
1000
ori
AlwNI (1094)
pCS2+:mCherry
4821 bp
(3568) NdeI
CMV IE94 promoter
3500
1500
(3381) MscI*
AhdI (1571)
AmpR
3000
f1 ori
AmpR promoter
2000
(3015) SalI
2500
(2816) NaeI
(2814) NgoMIV
(2713) DraIII
XmnI (2170)


Created with SnapGene®
(4346) BstXI
NdeI (4361)
(4342) MluI
(4337) NsiI
SP6
I-SceI (213)
PstI (285)
(4056) BspQI - SapI
BlpI (379)
BsmI (435)
lac
M13-rev
LacO
4000
500
crygc_prom
BamHI (674)
BstEII (713)
Acc65I (764)
KpnI (768)
attP
(3639) PspFI
(3635) BseYI
ColE1 origin
3500
1000
Gal4-VP16
HpaI (1055)
BsrGI (1068)
pDB783:Crygc:attP:Gal4VP16 modified
4370 bp
SalI (1224)
AccI (1225)
EcoRI (1235)
BglII (1241)
PshAI (1267)
BmgBI (1308)
BsaBI* (1343)
PfoI (1380)
M13-fwd
T7
1500
(3051) AhdI
3000
(2982) BpmI
(2904) NmeAIII
AmpR
Amp prom
F1 ori
2500
2000
NgoMIV (1751)
NaeI (1753)
BsaAI (1856)
DraIII (1859)
(2511) BsaHI
(2451) XmnI
PsiI (1984)

9 REFERENCES

Abe, K.-I., S. Funaya, D. Tsukioka, M. Kawamura, Y. Suzuki, M. G. Suzuki, R. M. Schultz and F. Aoki (2018). "Minor zygotic gene activation is essential for mouse preimplantation development." Proceedings of the National Academy of Sciences of the United States of America **115**(29): E6780-E6788.

Adamson, E. D. and H. R. Woodland (1974). "Histone synthesis in early amphibian development: Histone and DNA syntheses are not co-ordinated." Journal of Molecular Biology **88**(2): 263-285.

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

Akhtar, W. and G. J. Veenstra (2009). "TBP2 is a substitute for TBP in Xenopus oocyte transcription." BMC Biol **7**: 45.

Alberts, B. (2015). Molecular biology of the cell. New York, NY, Garland Science, Taylor and Francis Group.

Albright, S. R. and R. Tjian (2000). "TAFs revisited: more data reveal new twists and confirm old ideas." Gene **242**(1): 1-13.

Almouzni, G., M. Mechali and A. P. Wolffe (1991). "Transcription complex disruption caused by a transition in chromatin structure." Mol Cell Biol **11**(2): 655-665.

Almouzni, G. and A. P. Wolffe (1995). "Constraints on transcriptional activator function contribute to transcriptional quiescence during early Xenopus embryogenesis." The EMBO Journal **14**(8): 1752-1765.

Ambros, V. (2004). "The functions of animal microRNAs." Nature **431**(7006): 350-355.

Amodeo, A. A., D. Jukam, A. F. Straight and J. M. Skotheim (2015). "Histone titration against the genome sets the DNA-to-cytoplasm threshold for the Xenopus midblastula transition." Proc Natl Acad Sci U S A **112**(10): E1086-1095.

Anderson, K. V. and J. A. Lengyel (1980). "Changing rates of histone mRNA synthesis and turnover in Drosophila embryos." Cell **21**(3): 717-727.

Andersson, R. (2015). "Promoter or enhancer, what's the difference? Deconstruction of established distinctions and presentation of a unifying model." Bioessays **37**(3): 314-323.

Bailey, T. L. and P. Machanick (2012). "Inferring direct DNA binding from ChIP-seq." Nucleic Acids Research **40**(17): e128-e128.

Bannister, A. J., R. Schneider, F. A. Myers, A. W. Thorne, C. Crane-Robinson and T. Kouzarides (2005). "Spatial Distribution of Di- and Tri-methyl Lysine 36 of Histone H3 at Active Genes." Journal of Biological Chemistry **280**(18): 17732-17736.

Baroux, C., D. Autran, C. S. Gillmor, D. Grimanelli and U. Grossniklaus (2008). "The maternal to zygotic transition in animals and plants." Cold Spring Harb Symp Quant Biol **73**: 89-100.

Bartel, D. P. (2004). "MicroRNAs: genomics, biogenesis, mechanism, and function." Cell **116**(2): 281-297.

Bartfai, R., C. Balduf, T. Hilton, Y. Rathmann, Y. Hadzhiev, L. Tora, L. Orban and F. Muller (2004). "TBP2, a vertebrate-specific member of the TBP family, is required in embryonic development of zebrafish." Curr Biol **14**(7): 593-598.

Basehoar, A. D., S. J. Zanton and B. F. Pugh (2004). "Identification and distinct regulation of yeast TATA box-containing genes." Cell **116**(5): 699-709.

Bashirullah, A., S. R. Halsell, R. L. Cooperstock, M. Kloc, A. Karaiskakis, W. W. Fisher, W. L. Fu, J. K. Hamilton, L. D. Etkin and H. D. Lipshitz (1999). "Joint action of two RNA degradation pathways controls the timing of maternal transcript elimination at the midblastula transition in Drosophila melanogaster." Embo Journal **18**(9): 2610-2620.

Beagan, J. A. and J. E. Phillips-Cremins (2020). "On the existence and functionality of topologically associating domains." Nature Genetics **52**(1): 8-16.

Belting, H. G., G. Hauptmann, D. Meyer, S. Abdelliah-Seyfried, A. Chitnis, C. Eschbach, I. Soll, C. Thisse, B. Thisse, K. B. Artinger, K. Lunde and W. Driever (2001). "spiel ohne grenzen/pou2 is required during establishment of the zebrafish midbrain-hindbrain boundary organizer." Development **128**(21): 4165-4176.

Belting, H. G., B. Wendik, K. Lunde, M. Leichsenring, R. Mossner, W. Driever and D. Onichtchouk (2011). "Pou5f1 contributes to dorsoventral patterning by positive regulation of vox and modulation of fgf8a expression." Dev Biol **356**(2): 323-336.

Bernstein, B. E., A. Meissner and E. S. Lander (2007). "The Mammalian Epigenome." Cell **128**(4): 669-681.

Bernstein, B. E., T. S. Mikkelsen, X. Xie, M. Kamal, D. J. Huebert, J. Cuff, B. Fry, A. Meissner, M. Wernig, K. Plath, R. Jaenisch, A. Wagschal, R. Feil, S. L. Schreiber and E. S. Lander (2006). "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells." Cell **125**(2): 315-326.

Bertrand, E., P. Chartrand, M. Schaefer, S. M. Shenoy, R. H. Singer and R. M. Long (1998). "Localization of ASH1 mRNA particles in living yeast." Mol Cell **2**(4): 437-445.

Bird, A. (2002). "DNA methylation patterns and epigenetic memory." Genes & Development **16**(1): 6-21.

Blythe, S. A. and E. F. Wieschaus (2015). "Zygotic genome activation triggers the DNA replication checkpoint at the midblastula transition." Cell **160**(6): 1169-1181.

Blythe, S. A. and E. F. Wieschaus (2016). "Establishment and maintenance of heritable chromatin structure during early Drosophila embryogenesis." eLife **5**: e20148.

Bogenhagen, D. F. (2012). "Mitochondrial DNA nucleoid structure." Biochim Biophys Acta **1819**(9-10): 914-920.

Boija, A., I. A. Klein, B. R. Sabari, A. Dall'Agnese, E. L. Coffey, A. V. Zamudio, C. H. Li, K. Shrinivas, J. C. Manteiga, N. M. Hannett, B. J. Abraham, L. K. Afeyan, Y. E. Guo, J. K. Rimel, C. B. Fant, J. Schuijers, T. I. Lee, D. J. Taatjes and R. A. Young (2018). "Transcription Factors Activate Genes through the Phase-Separation Capacity of Their Activation Domains." Cell **175**(7): 1842-1855.e1816.

Bonev, B. and G. Cavalli (2016). "Organization and function of the 3D genome." Nature Reviews Genetics **17**(11): 661-678.

Bothma, J. P., H. G. Garcia, E. Esposito, G. Schlissel, T. Gregor and M. Levine (2014). "Dynamic regulation of eve stripe 2 expression reveals transcriptional bursts in living Drosophila embryos." Proc Natl Acad Sci U S A **111**(29): 10598-10603.

Bouhedda, F., A. Autour and M. Ryckelynck (2018). "Light-Up RNA Aptamers and Their Cognate Fluorogens: From Their Development to Their Applications." International Journal of Molecular Sciences **19**(1).

Boyer, L. A., T. I. Lee, M. F. Cole, S. E. Johnstone, S. S. Levine, J. P. Zucker, M. G. Guenther, R. M. Kumar, H. L. Murray, R. G. Jenner, D. K. Gifford, D. A. Melton, R. Jaenisch and R. A. Young (2005). "Core transcriptional regulatory circuitry in human embryonic stem cells." Cell **122**(6): 947-956.

Brand, A. H., L. Breeden, J. Abraham, R. Sternglanz and K. Nasmyth (1985). "Characterization of a “silencer” in yeast: A DNA sequence with properties opposite to those of a transcriptional enhancer." Cell **41**(1): 41-48.

Brandeis, M., D. Frank, I. Keshet, Z. Siegfried, M. Mendelsohn, A. Names, V. Temper, A. Razin and H. Cedar (1994). "Spl elements protect a CpG island from de novo methylation." Nature **371**(6496): 435-438.

Brangwynne, C. P., T. J. Mitchison and A. A. Hyman (2011). "Active liquid-like behavior of nucleoli determines their size and shape in Xenopus laevis oocytes." Proceedings of the National Academy of Sciences **108**(11): 4334.

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

Buratowski, S., S. Hahn, P. A. Sharp and L. Guarente (1988). "Function of a yeast TATA element-binding protein in a mammalian transcription system." Nature **334**(6177): 37-42.

Burgess-Beusse, B., C. Farrell, M. Gaszner, M. Litt, V. Mutskov, F. Recillas-Targa, M. Simpson, A. West and G. Felsenfeld (2002). "The insulation of genes from external enhancers and silencing chromatin." Proc Natl Acad Sci U S A **99 Suppl 4**(suppl 4): 16433-16437.

Burke, T. W. and J. T. Kadonaga (1996). "Drosophila TFIID binds to a conserved downstream basal promoter element that is present in many TATA-box-deficient promoters." Genes & Development **10**(6): 711-724.

Bushati, N., A. Stark, J. Brennecke and S. M. Cohen (2008). "Temporal reciprocity of miRNAs and their targets during the maternal-to-zygotic transition in Drosophila." Curr Biol **18**(7): 501-506.

Buxbaum, A. R., G. Haimovich and R. H. Singer (2015). "In the right place at the right time: visualizing and understanding mRNA localization." Nat Rev Mol Cell Biol **16**(2): 95-109.

Campos, E. I. and D. Reinberg (2009). "Histones: Annotating Chromatin." Annual Review of Genetics **43**(1): 559-599.

Carninci, P., A. Sandelin, B. Lenhard, S. Katayama, K. Shimokawa, J. Ponjavic, C. A. Semple, M. S. Taylor, P. G. Engstrom, M. C. Frith, A. 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 (2006). "Genome-wide analysis of mammalian promoter architecture and evolution." Nat Genet **38**(6): 626-635.

Cavallini, B., J. Huet, J.-L. Plassat, A. Sentenac, J.-M. Egly and P. Chambon (1988). "A yeast activity can substitute for the HeLa cell TATA box factor." Nature **334**(6177): 77-80.

Chan, S. H., Y. Tang, L. Miao, H. Darwich-Codore, C. E. Vejnar, J. D. Beaudoin, D. Musaev, J. P. Fernandez, M. D. J. Benitez, A. A. Bazzini, M. A. Moreno-Mateos and A. J. Giraldez (2019). "Brd4 and P300 Confer Transcriptional Competency during Zygotic Genome Activation." Dev Cell **49**(6): 867-881 e868.

Chen, H., L. C. Einstein, S. C. Little and M. C. Good (2019). "Spatiotemporal Patterning of Zygotic Genome Activation in a Model Vertebrate Embryo." Dev Cell **49**(6): 852-866 e857.

Chen, K., J. Johnston, W. Shao, S. Meier, C. Staber and J. Zeitlinger (2013). "A global change in RNA polymerase II pausing during the Drosophila midblastula transition." eLife **2**: e00861-e00861.

Chen, P. Y., H. Manninga, K. Slanchev, M. C. Chien, J. J. Russo, J. Y. Ju, R. Sheridan, B. John, D. S. Marks, D. Gaidatzis, C. Sander, M. Zavolan and T. Tuschl (2005). "The developmental miRNA profiles of zebrafish as determined by small RNA cloning." Genes & Development **19**(11): 1288-1293.

Chen, T. and S. Y. Dent (2014). "Chromatin modifiers and remodellers: regulators of cellular differentiation." Nat Rev Genet **15**(2): 93-106.

Cho, W.-K., J.-H. Spille, M. Hecht, C. Lee, C. Li, V. Grube and I. I. Cisse (2018). "Mediator and RNA polymerase II clusters associate in transcription-dependent condensates." Science (New York, N.Y.) **361**(6400): 412-415.

Chong, S., C. Dugast-Darzacq, Z. Liu, P. Dong, G. M. Dailey, C. Cattoglio, A. Heckert, S. Banala, L. Lavis, X. Darzacq and R. Tjian (2018). "Imaging dynamic and selective low-complexity domain interactions that control gene transcription." Science (New York, N.Y.) **361**(6400): eaar2555.

Cisse, I. I., I. Izeddin, S. Z. Causse, L. Boudarene, A. Senecal, L. Muresan, C. Dugast-Darzacq, B. Hajj, M. Dahan and X. Darzacq (2013). "Real-Time Dynamics of RNA Polymerase II Clustering in Live Human Cells." Science **341**(6146): 664.

Collart, C., G. E. Allen, C. R. Bradshaw, J. C. Smith and P. Zegerman (2013). "Titration of Four Replication Factors Is Essential for the Xenopus laevis Midblastula Transition." Science **341**(6148): 893.

Conic, S., D. Desplancq, A. Ferrand, V. Fischer, V. Heyer, B. Reina San Martin, J. Pontabry, M. Oulad-Abdelghani, K. Babu N, G. D. Wright, N. Molina, E. Weiss and L. Tora (2018). "Imaging of native transcription factors and histone phosphorylation at high resolution in live cells." Journal of Cell Biology **217**(4): 1537-1552.

Core, L. J., J. J. Waterfall and J. T. Lis (2008). "Nascent RNA Sequencing Reveals Widespread Pausing and Divergent Initiation at Human Promoters." Science **322**(5909): 1845-1848.

Corrigan, A. M., E. Tunnacliffe, D. Cannon and J. R. Chubb (2016). "A continuum model of transcriptional bursting." Elife **5**: e13051.

Cramer, P. (2002). "Multisubunit RNA polymerases." Curr Opin Struct Biol **12**(1): 89-97.

Cremer, T., C. Cremer and P. Lichter (2014). "Recollections of a scientific journey published in human genetics: from chromosome territories to interphase cytogenetics and comparative genome hybridization." Human Genetics **133**(4): 403-416.

Crowley, T. E., T. Hoey, J.-K. Liu, Y. N. Jan, L. Y. Jan and R. Tjian (1993). "A new factor related to TATA-binding protein has highly restricted expression patterns in Drosophila." Nature **361**(6412): 557-561.

Cui, K. and K. Zhao (2012). "Genome-wide approaches to determining nucleosome occupancy in metazoans using MNase-Seq." Methods in molecular biology (Clifton, N.J.) **833**: 413-419.

Cvetesic, N., H. G. Leitch, M. Borkowska, F. Müller, P. Carninci, P. Hajkova and B. Lenhard (2018). "SLIC-CAGE: high-resolution transcription start site mapping using nanogram-levels of total RNA." Genome Research **28**(12): 1943-1956.

Danino, Y. M., D. Even, D. Ideses and T. Juven-Gershon (2015). "The core promoter: At the heart of gene expression." Biochim Biophys Acta **1849**(8): 1116-1131.

Dantonel, J.-C., J.-M. Wurtz, O. Poch, D. Moras and L. Tora (1999). "The TBP-like factor: an alternative transcription factor in Metazoa?" Trends in Biochemical Sciences **24**(9): 335-339.

De Iaco, A., E. Planet, A. Coluccio, S. Verp, J. Duc and D. Trono (2017). "DUX-family transcription factors regulate zygotic genome activation in placental mammals." Nat Genet **49**(6): 941-945.

De Iaco, A., S. Verp, S. Offner, D. Grun and D. Trono (2020). "DUX is a non-essential synchronizer of zygotic genome activation." Development **147**(2).

De Renzis, S., O. Elemento, S. Tavazoie and E. F. Wieschaus (2007). "Unmasking activation of the zygotic genome using chromosomal deletions in the Drosophila embryo." PLoS Biol **5**(5): e117.

Deaton, A. M. and A. Bird (2011). "CpG islands and the regulation of transcription." Genes & Development **25**(10): 1010-1022.

Dekker, J. (2008). "Gene regulation in the third dimension." Science **319**(5871): 1793-1794.

Dekker, J., K. Rippe, M. Dekker and N. Kleckner (2002). "Capturing Chromosome Conformation." Science **295**(5558): 1306.

Deng, W., J. Lee, H. Wang, J. Miller, A. Reik, Philip D. Gregory, A. Dean and Gerd A. Blobel (2012). "Controlling Long-Range Genomic Interactions at a Native Locus by Targeted Tethering of a Looping Factor." Cell **149**(6): 1233-1244.

Deng, W., Jeremy W. Rupon, I. Krivega, L. Breda, I. Motta, Kristen S. Jahn, A. Reik, Philip D. Gregory, S. Rivella, A. Dean and Gerd A. Blobel (2014). "Reactivation of Developmentally Silenced Globin Genes by Forced Chromatin Looping." Cell **158**(4): 849-860.

Deng, W. S. and S. G. E. Roberts (2005). "A core promoter element downstream of the TATA box that is recognized by TFIIB." Genes & Development **19**(20): 2418-2423.

Dobin, A., C. A. Davis, F. Schlesinger, J. Drenkow, C. Zaleski, S. Jha, P. Batut, M. Chaisson and T. R. Gingeras (2013). "STAR: ultrafast universal RNA-seq aligner." Bioinformatics **29**(1): 15-21.

Dostie, J., 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 (2006). "Chromosome Conformation Capture Carbon Copy (5C): A massively parallel solution for mapping interactions between genomic elements." Genome Research **16**(10): 1299-1309.

Edgar, B. A. and G. Schubiger (1986). "Parameters controlling transcriptional activation during early Drosophila development." Cell **44**(6): 871-877.

Fabian, M. R., N. Sonenberg and W. Filipowicz (2010). "Regulation of mRNA translation and stability by microRNAs." Annu Rev Biochem **79**(1545-4509): 351-379.

Fejes-Toth, K., V. Sotirova, R. Sachidanandam, G. Assaf, G. J. Hannon, P. Kapranov, S. Foissac, A. T. Willingham, R. Duttagupta, E. Dumais, T. R. Gingeras, E. T. P. Affymetrix/Cold Spring Harbor Laboratory, L. Cold Spring Harbor and Affymetrix (2009). "Post-transcriptional processing generates a diversity of 5′-modified long and short RNAs." Nature **457**(7232): 1028-1032.

Ferg, M., R. Sanges, J. Gehrig, J. Kiss, M. Bauer, A. Lovas, M. Szabo, L. Yang, U. Straehle, M. J. Pankratz, F. Olasz, E. Stupka and F. Muller (2007). "The TATA-binding protein regulates maternal mRNA degradation and differential zygotic transcription in zebrafish." EMBO J **26**(17): 3945-3956.

Foe, V. E. and B. M. Alberts (1983). "Studies of nuclear and cytoplasmic behaviour during the five mitotic cycles that precede gastrulation in Drosophila embryogenesis." Journal of Cell Science **61**(1): 31.

Foo, S. M., Y. Sun, B. Lim, R. Ziukaite, K. O'Brien, C. Y. Nien, N. Kirov, S. Y. Shvartsman and C. A. Rushlow (2014). "Zelda potentiates morphogen activity by increasing chromatin accessibility." Curr Biol **24**(12): 1341-1346.

Frith, M. C., M. C. Li and Z. Weng (2003). "Cluster-Buster: finding dense clusters of motifs in DNA sequences." Nucleic Acids Research **31**(13): 3666-3668.

Gagnon, J. A., K. Obbad and A. F. Schier (2018). "The primary role of zebrafish nanog is in extra-embryonic tissue." Development **145**(1): dev147793.

Garcia, J. F. and R. Parker (2015). "MS2 coat proteins bound to yeast mRNAs block 5′ to 3′ degradation and trap mRNA decay products: implications for the localization of mRNAs by MS2-MCP system." RNA **21**(8): 1393-1395.

Gardiner-Garden, M. and M. Frommer (1987). "CpG Islands in vertebrate genomes." Journal of Molecular Biology **196**(2): 261-282.

Gazdag, E., U. G. Jacobi, I. van Kruijsbergen, D. L. Weeks and G. J. C. Veenstra (2016). "Activation of a T-box-Otx2-Gsc gene network independent of TBP and TBP-related factors." Development (Cambridge, England) **143**(8): 1340-1350.

Gazdag, E., A. Rajkovic, M. E. Torres-Padilla and L. Tora (2007). "Analysis of TATA-binding protein 2 (TBP2) and TBP expression suggests different roles for the two proteins in regulation of gene expression during oogenesis and early mouse development." Reproduction **134**(1): 51-62.

Gazdag, E., A. Santenard, C. Ziegler-Birling, G. Altobelli, O. Poch, L. Tora and M.-E. Torres-Padilla (2009). "TBP2 is essential for germ cell development by regulating transcription and chromatin condensation in the oocyte." Genes & development **23**(18): 2210-2223.

Gentsch, G. E., N. D. L. Owens and J. C. Smith (2019). "The Spatiotemporal Control of Zygotic Genome Activation." iScience **16**: 485-498.

Gentsch, G. E., T. Spruce, N. D. L. Owens and J. C. Smith (2019). "Maternal pluripotency factors initiate extensive chromatin remodelling to predefine first response to inductive signals." Nature Communications **10**(1): 4269.

Gerhart, J. C. (1980). Mechanisms Regulating Pattern Formation in the Amphibian Egg and Early Embryo. Biological Regulation and Development. R. Goldberger, Springer US. **2:** 133-316.

Gershenzon, N. I. and I. P. Ioshikhes (2005). "Synergy of human Pol II core promoter elements revealed by statistical sequence analysis." Bioinformatics **21**(8): 1295-1300.

Geyer, P. K. and V. G. Corces (1992). "DNA position-specific repression of transcription by a Drosophila zinc finger protein." Genes & Development **6**(10): 1865-1873.

Giraldez, A. J., R. M. Cinalli, M. E. Glasner, A. J. Enright, J. M. Thomson, S. Baskerville, S. M. Hammond, D. P. Bartel and A. F. Schier (2005). "MicroRNAs regulate brain morphogenesis in zebrafish." Science **308**(5723): 833-838.

Giraldez, A. J., Y. Mishima, J. Rihel, R. J. Grocock, S. Van Dongen, K. Inoue, A. J. Enright and A. F. Schier (2006). "Zebrafish MiR-430 promotes deadenylation and clearance of maternal mRNAs." Science **312**(5770): 75-79.

Goldberg, M. L. (1979). "Ph.D. Thesis, Stanford University, Stanford, CA, U.S.A."

Goll, M. G. and T. H. Bestor (2005). "EUKARYOTIC CYTOSINE METHYLTRANSFERASES." Annual Review of Biochemistry **74**(1): 481-514.

Goodrich, J. A. and R. Tjian (2010). "Unexpected roles for core promoter recognition factors in cell-type-specific transcription and gene regulation." Nat Rev Genet **11**(8): 549-558.

Grow, E. J., R. A. Flynn, S. L. Chavez, N. L. Bayless, M. Wossidlo, D. J. Wesche, L. Martin, C. B. Ware, C. A. Blish, H. Y. Chang, R. A. Pera and J. Wysocka (2015). "Intrinsic retroviral reactivation in human preimplantation embryos and pluripotent cells." Nature **522**(7555): 221-225.

Gu, W., H.-C. Lee, D. Chaves, Elaine M. Youngman, Gregory J. Pazour, D. Conte and Craig C. Mello (2012). "CapSeq and CIP-TAP Identify Pol II Start Sites and Reveal Capped Small RNAs as C. elegans piRNA Precursors." Cell **151**(7): 1488-1500.

Guenther, M. G., S. S. Levine, L. A. Boyer, R. Jaenisch and R. A. Young (2007). "A Chromatin Landmark and Transcription Initiation at Most Promoters in Human Cells." Cell **130**(1): 77-88.

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

Haberle, V. and B. Lenhard (2016). "Promoter architectures and developmental gene regulation." Semin Cell Dev Biol **57**: 11-23.

Haberle, V., 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 (2014). "Two independent transcription initiation codes overlap on vertebrate core promoters." Nature **507**(7492): 381-385.

Haberle, V. and A. Stark (2018). "Eukaryotic core promoters and the functional basis of transcription initiation." Nat Rev Mol Cell Biol **19**(10): 621-637.

Hadzhiev, Y., I. Miguel-Escalada, D. Balciunas and F. Müller (2016). Testing of Cis-Regulatory Elements by Targeted Transgene Integration in Zebrafish Using PhiC31 Integrase. Zebrafish: Methods and Protocols. K. Kawakami, E. E. Patton and M. Orger. New York, NY, Springer New York**:** 81-91.

Hadzhiev, Y., H. K. Qureshi, L. Wheatley, L. Cooper, A. Jasiulewicz, H. Van Nguyen, J. W. Wragg, D. Poovathumkadavil, S. Conic, S. Bajan, A. Sik, G. Hutvagner, L. Tora, A. Gambus, J. S. Fossey and F. Muller (2019). "A cell cycle-coordinated Polymerase II transcription compartment encompasses gene expression before global genome activation." Nat Commun **10**(1): 691.

Hahn, S., S. Buratowski, P. A. Sharp and L. Guarente (1989). "Isolation of the gene encoding the yeast TATA binding protein TFIID: A gene identical to the SPT15 suppressor of Ty element insertions." Cell **58**(6): 1173-1181.

Hahn, S., S. Buratowski, P. A. Sharp and L. Guarente (1989). "Yeast TATA-binding protein TFIID binds to TATA elements with both consensus and nonconsensus DNA sequences." Proceedings of the National Academy of Sciences **86**(15): 5718.

Hamatani, T., M. G. Carter, A. A. Sharov and M. S. Ko (2004). "Dynamics of global gene expression changes during mouse preimplantation development." Dev Cell **6**(1): 117-131.

Hansen, S. K., S. Takada, R. H. Jacobson, J. T. Lis and R. Tjian (1997). "Transcription Properties of a Cell Type–Specific TATA-Binding Protein, TRF." Cell **91**(1): 71-83.

Hara, K. (1977). "The cleavage pattern of the axolotl egg studied by cinematography and cell counting." Wilhelm Roux's Archives of Developmental Biology **181**(1): 73-87.

Harrison, M. M., X. Y. Li, T. Kaplan, M. R. Botchan and M. B. Eisen (2011). "Zelda binding in the early Drosophila melanogaster embryo marks regions subsequently activated at the maternal-to-zygotic transition." PLoS Genet **7**(10): e1002266.

Hart, D. O., T. Raha, N. D. Lawson and M. R. Green (2007). "Initiation of zebrafish haematopoiesis by the TATA-box-binding protein-related factor Trf3." Nature **450**(7172): 1082-1085.

Harvey, S. A., I. Sealy, R. Kettleborough, F. Fenyes, R. White, D. Stemple and J. C. Smith (2013). "Identification of the zebrafish maternal and paternal transcriptomes." Development **140**(13): 2703-2710.

Heintzman, N. D., R. K. Stuart, G. Hon, Y. Fu, C. W. Ching, R. D. Hawkins, L. O. Barrera, S. Van Calcar, C. Qu, K. A. Ching, W. Wang, Z. Weng, R. D. Green, G. E. Crawford and B. Ren (2007). "Distinct and predictive chromatin signatures of transcriptional promoters and enhancers in the human genome." Nature Genetics **39**(3): 311-318.

Henikoff, J. G., J. A. Belsky, K. Krassovsky, D. M. MacAlpine and S. Henikoff (2011). "Epigenome characterization at single base-pair resolution." Proceedings of the National Academy of Sciences **108**(45): 18318.

Hernandez, N. (1993). "TBP, a universal eukaryotic transcription factor?" Genes & Development **7**(7b): 1291-1308.

Heyn, P., M. Kircher, A. Dahl, J. Kelso, P. Tomancak, A. T. Kalinka and K. M. Neugebauer (2014). "The earliest transcribed zygotic genes are short, newly evolved, and different across species." Cell Rep **6**(2): 285-292.

Hilbert, L., Y. Sato, H. Kimura, F. Jülicher, A. Honigmann, V. Zaburdaev and N. L. Vastenhouw (2018). "Transcription organizes euchromatin similar to an active microemulsion." bioRxiv: 234112.

Hirose, F., M. Yamaguchi, H. Handa, Y. Inomata and A. Matsukage (1993). "Novel 8-base pair sequence (Drosophila DNA replication-related element) and specific binding factor involved in the expression of Drosophila genes for DNA polymerase alpha and proliferating cell nuclear antigen." Journal of Biological Chemistry **268**(3): 2092-2099.

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

Hoffmann, S., C. Otto, S. Kurtz, C. M. Sharma, P. Khaitovich, J. Vogel, P. F. Stadler and J. Hackermüller (2009). "Fast Mapping of Short Sequences with Mismatches, Insertions and Deletions Using Index Structures." PLOS Computational Biology **5**(9): e1000502.

Holmes, M. C. and R. Tjian (2000). "Promoter-Selective Properties of the TBP-Related Factor TRF1." Science **288**(5467): 867.

Hong, L., G. P. Schroth, H. R. Matthews, P. Yau and E. M. Bradbury (1993). "Studies of the DNA binding properties of histone H4 amino terminus. Thermal denaturation studies reveal that acetylation markedly reduces the binding constant of the H4 "tail" to DNA." Journal of Biological Chemistry **268**(1): 305-314.

Houbaviy, H. B., M. F. Murray and P. A. Sharp (2003). "Embryonic stem cell-specific MicroRNAs." Dev Cell **5**(2): 351-358.

Howe, K., M. D. Clark, C. F. Torroja, J. Torrance, C. Berthelot, M. Muffato, J. E. Collins, S. Humphray, K. McLaren, L. Matthews, S. McLaren, I. Sealy, M. Caccamo, C. Churcher, C. Scott, J. C. Barrett, R. Koch, G.-J. Rauch, S. White, W. Chow, B. Kilian, L. T. Quintais, J. A. Guerra-Assunção, Y. Zhou, Y. Gu, J. Yen, J.-H. Vogel, T. Eyre, S. Redmond, R. Banerjee, J. Chi, B. Fu, E. Langley, S. F. Maguire, G. K. Laird, D. Lloyd, E. Kenyon, S. Donaldson, H. Sehra, J. Almeida-King, J. Loveland, S. Trevanion, M. Jones, M. Quail, D. Willey, A. Hunt, J. Burton, S. Sims, K. McLay, B. Plumb, J. Davis, C. Clee, K. Oliver, R. Clark, C. Riddle, D. Elliott, G. Threadgold, G. Harden, D. Ware, S. Begum, B. Mortimore, G. Kerry, P. Heath, B. Phillimore, A. Tracey, N. Corby, M. Dunn, C. Johnson, J. Wood, S. Clark, S. Pelan, G. Griffiths, M. Smith, R. Glithero, P. Howden, N. Barker, C. Lloyd, C. Stevens, J. Harley, K. Holt, G. Panagiotidis, J. Lovell, H. Beasley, C. Henderson, D. Gordon, K. Auger, D. Wright, J. Collins, C. Raisen, L. Dyer, K. Leung, L. Robertson, K. Ambridge, D. Leongamornlert, S. McGuire, R. Gilderthorp, C. Griffiths, D. Manthravadi, S. Nichol, G. Barker, S. Whitehead, M. Kay, J. Brown, C. Murnane, E. Gray, M. Humphries, N. Sycamore, D. Barker, D. Saunders, J. Wallis, A. Babbage, S. Hammond, M. Mashreghi-Mohammadi, L. Barr, S. Martin, P. Wray, A. Ellington, N. Matthews, M. Ellwood, R. Woodmansey, G. Clark, J. D. Cooper, A. Tromans, D. Grafham, C. Skuce, R. Pandian, R. Andrews, E. Harrison, A. Kimberley, J. Garnett, N. Fosker, R. Hall, P. Garner, D. Kelly, C. Bird, S. Palmer, I. Gehring, A. Berger, C. M. Dooley, Z. Ersan-Ürün, C. Eser, H. Geiger, M. Geisler, L. Karotki, A. Kirn, J. Konantz, M. Konantz, M. Oberländer, S. Rudolph-Geiger, M. Teucke, C. Lanz, G. Raddatz, K. Osoegawa, B. Zhu, A. Rapp, S. Widaa, C. Langford, F. Yang, S. C. Schuster, N. P. Carter, J. Harrow, Z. Ning, J. Herrero, S. M. J. Searle, A. Enright, R. Geisler, R. H. A. Plasterk, C. Lee, M. Westerfield, P. J. de Jong, L. I. Zon, J. H. Postlethwait, C. Nüsslein-Volhard, T. J. P. Hubbard, H. R. Crollius, J. Rogers and D. L. Stemple (2013). "The zebrafish reference genome sequence and its relationship to the human genome." Nature **496**(7446): 498-503.

Howe, K., P. H. Schiffer, J. Zielinski, T. Wiehe, G. K. Laird, J. C. Marioni, O. Soylemez, F. Kondrashov and M. Leptin (2016). "Structure and evolutionary history of a large family of NLR proteins in the zebrafish." Open Biology **6**(4): 160009.

Hug, C. B., A. G. Grimaldi, K. Kruse and J. M. Vaquerizas (2017). "Chromatin Architecture Emerges during Zygotic Genome Activation Independent of Transcription." Cell **169**(2): 216-228.e219.

Hughes, T. R. (2011). A handbook of transcription factors. Dordrecht, Springer.

Jacobi, U. G., R. C. Akkers, E. S. Pierson, D. L. Weeks, J. M. Dagle and G. J. C. Veenstra (2007). "TBP paralogs accommodate metazoan- and vertebrate-specific developmental gene regulation." The EMBO journal **26**(17): 3900-3909.

Jaenisch, R., D. Jähner, P. Nobis, I. Simon, J. Löhler, K. Harbers and D. Grotkopp (1981). "Chromosomal position and activation of retroviral genomes inserted into the germ line of mice." Cell **24**(2): 519-529.

Jallow, Z., U. G. Jacobi, D. L. Weeks, I. B. Dawid and G. J. C. Veenstra (2004). "Specialized and redundant roles of TBP and a vertebrate-specific TBP paralog in embryonic gene regulation in Xenopus." Proceedings of the National Academy of Sciences of the United States of America **101**(37): 13525-13530.

Jin, V. X., G. A. Singer, F. J. Agosto-Perez, S. Liyanarachchi and R. V. Davuluri (2006). "Genome-wide analysis of core promoter elements from conserved human and mouse orthologous pairs." BMC Bioinformatics **7**(1): 114.

Joseph, S. R., M. Pálfy, L. Hilbert, M. Kumar, J. Karschau, V. Zaburdaev, A. Shevchenko and N. L. Vastenhouw (2017). "Competition between histone and transcription factor binding regulates the onset of transcription in zebrafish embryos." eLife **6**: e23326.

Kadonaga, J. T. (2002). "The DPE, a core promoter element for transcription by RNA polymerase II." Exp Mol Med **34**(4): 259-264.

Kane, D. A. and C. B. Kimmel (1993). "The zebrafish midblastula transition." Development **119**(2): 447-456.

Kellum, R. and P. Schedl (1992). "A group of scs elements function as domain boundaries in an enhancer-blocking assay." Molecular and Cellular Biology **12**(5): 2424.

Kent, W. J. (2002). "BLAT--the BLAST-like alignment tool." Genome Res **12**(4): 656-664.

Kimmel, C. B., W. W. Ballard, S. R. Kimmel, B. Ullmann and T. F. Schilling (1995). "Stages of embryonic development of the zebrafish." Dev Dyn **203**(3): 253-310.

Klein, D. C. and S. J. Hainer (2020). "Genomic methods in profiling DNA accessibility and factor localization." Chromosome Research **28**(1): 69-85.

Klemm, S. L., Z. Shipony and W. J. Greenleaf (2019). "Chromatin accessibility and the regulatory epigenome." Nature Reviews Genetics **20**(4): 207-220.

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

Kotkamp, K., R. Mössner, A. Allen, D. Onichtchouk and W. Driever (2014). "A Pou5f1/Oct4 dependent Klf2a, Klf2b, and Klf17 regulatory sub-network contributes to EVL and ectoderm development during zebrafish embryogenesis." Developmental Biology **385**(2): 433-447.

Kouzarides, T. (2007). "Chromatin Modifications and Their Function." Cell **128**(4): 693-705.

Kruesi, W. S., L. J. Core, C. T. Waters, J. T. Lis and B. J. Meyer (2013). "Condensin controls recruitment of RNA polymerase II to achieve nematode X-chromosome dosage compensation." Elife **2**: e00808.

Kwasnieski, J. C., T. L. Orr-Weaver and D. P. Bartel (2019). "Early genome activation in Drosophila is extensive with an initial tendency for aborted transcripts and retained introns." Genome Research.

Lagrange, T., A. N. Kapanidis, H. Tang, D. Reinberg and R. H. Ebright (1998). "New core promoter element in RNA polymerase II-dependent transcription: sequence-specific DNA binding by transcription factor IIB." Genes Dev **12**(1): 34-44.

Langenberger, D., J. Bermudez-Santana C Fau - Hertel, S. Hertel J Fau - Hoffmann, P. Hoffmann S Fau - Khaitovich, P. F. Khaitovich P Fau - Stadler and P. F. Stadler (2009). "Evidence for human microRNA-offset RNAs in small RNA sequencing data." Bioinformatics **25**(18): 2298-2301.

Langmead, B. and S. L. Salzberg (2012). "Fast gapped-read alignment with Bowtie 2." Nat Methods **9**(4): 357-359.

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

Lau, N. C., L. P. Lim, E. G. Weinstein and D. P. Bartel (2001). "An abundant class of tiny RNAs with probable regulatory roles in Caenorhabditis elegans." Science **294**(5543): 858-862.

Lauberth, S. M., T. Nakayama, X. Wu, A. L. Ferris, Z. Tang, S. H. Hughes and R. G. Roeder (2013). "H3K4me3 interactions with TAF3 regulate preinitiation complex assembly and selective gene activation." Cell **152**(5): 1021-1036.

Lee, M. T., A. R. Bonneau and A. J. Giraldez (2014). "Zygotic genome activation during the maternal-to-zygotic transition." Annu Rev Cell Dev Biol **30**: 581-613.

Lee, M. T., A. R. Bonneau, C. M. Takacs, A. A. Bazzini, K. R. DiVito, E. S. Fleming and A. J. Giraldez (2013). "Nanog, Pou5f1 and SoxB1 activate zygotic gene expression during the maternal-to-zygotic transition." Nature **503**(7476): 360-364.

Leenen, F. A., S. Vernocchi, O. E. Hunewald, S. Schmitz, A. M. Molitor, C. P. Muller and J. D. Turner (2016). "Where does transcription start? 5'-RACE adapted to next-generation sequencing." Nucleic Acids Res **44**(6): 2628-2645.

Leichsenring, M., J. Maes, R. Mossner, W. Driever and D. Onichtchouk (2013). "Pou5f1 transcription factor controls zygotic gene activation in vertebrates." Science **341**(6149): 1005-1009.

Lenhard, B., A. Sandelin and P. Carninci (2012). "Metazoan promoters: emerging characteristics and insights into transcriptional regulation." Nature Reviews Genetics **13**(4): 233-245.

Lettice, L. A., 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 (2003). "A long-range Shh enhancer regulates expression in the developing limb and fin and is associated with preaxial polydactyly." Human Molecular Genetics **12**(14): 1725-1735.

Levsky, J. M. and R. H. Singer (2003). "Fluorescence in situ hybridization: past, present and future." Journal of Cell Science **116**(14): 2833.

Li, H., B. Handsaker, A. Wysoker, T. Fennell, J. Ruan, N. Homer, G. Marth, G. Abecasis, R. Durbin and S. Genome Project Data Processing (2009). "The Sequence Alignment/Map format and SAMtools." Bioinformatics **25**(16): 2078-2079.

Li, X. Y., M. M. Harrison, J. E. Villalta, T. Kaplan and M. B. Eisen (2014). "Establishment of regions of genomic activity during the Drosophila maternal to zygotic transition." Elife **3**: e03737.

Liang, G., J. C. Y. Lin, V. Wei, C. Yoo, J. C. Cheng, C. T. Nguyen, D. J. Weisenberger, G. Egger, D. Takai, F. A. Gonzales and P. A. Jones (2004). "Distinct localization of histone H3 acetylation and H3-K4 methylation to the transcription start sites in the human genome." PNAS **101**(19): 7357.

Lieberman-Aiden, E., N. L. van Berkum, L. Williams, M. Imakaev, T. Ragoczy, A. Telling, I. Amit, B. R. Lajoie, P. J. Sabo, M. O. Dorschner, R. Sandstrom, B. Bernstein, M. A. Bender, M. Groudine, A. Gnirke, J. Stamatoyannopoulos, L. A. Mirny, E. S. Lander and J. Dekker (2009). "Comprehensive Mapping of Long-Range Interactions Reveals Folding Principles of the Human Genome." Science **326**(5950): 289.

Lim, C. Y., B. Santoso, T. Boulay, E. Dong, U. Ohler and J. T. Kadonaga (2004). "The MTE, a new core promoter element for transcription by RNA polymerase II." Genes Dev **18**(13): 1606-1617.

Lindeman, Leif C., Ingrid S. Andersen, Andrew H. Reiner, N. Li, H. Aanes, O. Østrup, C. Winata, S. Mathavan, F. Müller, P. Aleström and P. Collas (2011). "Prepatterning of Developmental Gene Expression by Modified Histones before Zygotic Genome Activation." Developmental Cell **21**(6): 993-1004.

Litt, M. D., M. Simpson, F. Recillas-Targa, M. N. Prioleau and G. Felsenfeld (2001). "Transitions in histone acetylation reveal boundaries of three separately regulated neighboring loci." The EMBO journal **20**(9): 2224-2235.

Loh, Y. H., Q. Wu, J. L. Chew, V. B. Vega, W. Zhang, X. Chen, G. Bourque, J. George, B. Leong, J. Liu, K. Y. Wong, K. W. Sung, C. W. Lee, X. D. Zhao, K. P. Chiu, L. Lipovich, V. A. Kuznetsov, P. Robson, L. W. Stanton, C. L. Wei, Y. Ruan, B. Lim and H. H. Ng (2006). "The Oct4 and Nanog transcription network regulates pluripotency in mouse embryonic stem cells." Nat Genet **38**(4): 431-440.

Long, H. K., D. Sims, A. Heger, N. P. Blackledge, C. Kutter, M. L. Wright, F. Grutzner, D. T. Odom, R. Patient, C. P. Ponting and R. J. Klose (2013). "Epigenetic conservation at gene regulatory elements revealed by non-methylated DNA profiling in seven vertebrates." Elife **2**: e00348.

Lorch, Y., J. W. LaPointe and R. D. Kornberg (1987). "Nucleosomes inhibit the initiation of transcription but allow chain elongation with the displacement of histones." Cell **49**(2): 203-210.

Loubiere, V., A.-M. Martinez and G. Cavalli (2019). "Cell Fate and Developmental Regulation Dynamics by Polycomb Proteins and 3D Genome Architecture." BioEssays **41**(3): 1800222.

Lund, E., M. Liu, R. S. Hartley, M. D. Sheets and J. E. Dahlberg (2009). "Deadenylation of maternal mRNAs mediated by miR-427 in Xenopus laevis embryos." RNA **15**(12): 2351-2363.

Lunde, K., H. G. Belting and W. Driever (2004). "Zebrafish pou5f1/pou2, homolog of mammalian Oct4, functions in the endoderm specification cascade." Curr Biol **14**(1): 48-55.

Macfarlan, T. S., W. D. Gifford, S. Driscoll, K. Lettieri, H. M. Rowe, D. Bonanomi, A. Firth, O. Singer, D. Trono and S. L. Pfaff (2012). "Embryonic stem cell potency fluctuates with endogenous retrovirus activity." Nature **487**(7405): 57-63.

Machyna, M., P. Heyn and K. M. Neugebauer (2013). "Cajal bodies: where form meets function." WIREs RNA **4**(1): 17-34.

Maeno, M., P. E. Mead, C. Kelley, R. H. Xu, H. F. Kung, A. Suzuki, N. Ueno and L. I. Zon (1996). "The role of BMP-4 and GATA-2 in the induction and differentiation of hematopoietic mesoderm in Xenopus laevis." Blood **88**(6): 1965-1972.

Maldonado, E. (1999). "Transcriptional Functions of a New Mammalian TATA-binding Protein-related Factor." Journal of Biological Chemistry **274**(19): 12963-12966.

Mannack, L. V. J. C., S. Eising and A. Rentmeister (2016). "Current techniques for visualizing RNA in cells." F1000Research **5**: F1000 Faculty Rev-1775.

Martinez Arias, A. and B. Steventon (2018). "On the nature and function of organizers." Development **145**(5): dev159525.

Mathavan, S., S. G. P. Lee, A. Mak, L. D. Miller, K. R. K. Murthy, K. R. Govindarajan, Y. Tong, Y. L. Wu, S. H. Lam, H. Yang, Y. Ruan, V. Korzh, Z. Gong, E. T. Liu and T. Lufkin (2005). "Transcriptome analysis of zebrafish embryogenesis using microarrays." PLoS genetics **1**(2): 260-276.

McDaniel, S. L., T. J. Gibson, K. N. Schulz, M. Fernandez Garcia, M. Nevil, S. U. Jain, P. W. Lewis, K. S. Zaret and M. M. Harrison (2019). "Continued Activity of the Pioneer Factor Zelda Is Required to Drive Zygotic Genome Activation." Molecular Cell **74**(1): 185-195.e184.

McSwiggen, D. T., A. S. Hansen, S. S. Teves, H. Marie-Nelly, Y. Hao, A. B. Heckert, K. K. Umemoto, C. Dugast-Darzacq, R. Tjian and X. Darzacq (2019). "Evidence for DNA-mediated nuclear compartmentalization distinct from phase separation." eLife **8**: e47098.

Meier, M., J. Grant, A. Dowdle, A. Thomas, J. Gerton, P. Collas, J. M. Sullivan and J. A. Horsfield (2018). "Cohesin facilitates zygotic genome activation in zebrafish." Development **145**(1): dev156521.

Mikkelsen, T. S., M. Ku, D. B. Jaffe, B. Issac, E. Lieberman, G. Giannoukos, P. Alvarez, W. Brockman, T.-K. Kim, R. P. Koche, W. Lee, E. Mendenhall, A. O'Donovan, A. Presser, C. Russ, X. Xie, A. Meissner, M. Wernig, R. Jaenisch, C. Nusbaum, E. S. Lander and B. E. Bernstein (2007). "Genome-wide maps of chromatin state in pluripotent and lineage-committed cells." Nature **448**(7153): 553-560.

Mir, M., W. Bickmore, E. E. M. Furlong and G. Narlikar (2019). "Chromatin topology, condensates and gene regulation: shifting paradigms or just a phase?" Development **146**(19): dev182766.

Moore, P. A., J. Ozer, M. Salunek, G. Jan, D. Zerby, S. Campbell and P. M. Lieberman (1999). "A human TATA binding protein-related protein with altered DNA binding specificity inhibits transcription from multiple promoters and activators." Molecular and cellular biology **19**(11): 7610-7620.

Morris, S. A., B. A. Rao B Fau - Garcia, S. B. Garcia Ba Fau - Hake, R. L. Hake Sb Fau - Diaz, J. Diaz Rl Fau - Shabanowitz, D. F. Shabanowitz J Fau - Hunt, C. D. Hunt Df Fau - Allis, J. D. Allis Cd Fau - Lieb, B. D. Lieb Jd Fau - Strahl and B. D. Strahl (2007). "Identification of histone H3 lysine 36 acetylation as a highly conserved histone modification." J Biol Chem **9**(10): 7632-7640.

Murata, M., M. Nishiyori-Sueki H Fau - Kojima-Ishiyama, P. Kojima-Ishiyama M Fau - Carninci, Y. Carninci P Fau - Hayashizaki, M. Hayashizaki Y Fau - Itoh and M. Itoh (2014). "Detecting expressed genes using CAGE." Methods Mol Biol. **1164**(1940-6029 (Electronic)): 67-85.

Nelles, D. A., M. Y. Fang, M. R. O'Connell, J. L. Xu, S. J. Markmiller, J. A. Doudna and G. W. Yeo (2016). "Programmable RNA Tracking in Live Cells with CRISPR/Cas9." Cell **165**(2): 488-496.

Nepal, C., M. Coolen, Y. Hadzhiev, D. Cussigh, P. Mydel, V. M. Steen, P. Carninci, J. B. Andersen, L. Bally-Cuif, F. Müller and B. Lenhard (2015). "Transcriptional, post-transcriptional and chromatin-associated regulation of pri-miRNAs, pre-miRNAs and moRNAs." Nucleic Acids Research **44**(7): 3070-3081.

Nepal, C., Y. Hadzhiev, P. Balwierz, E. Tarifeño-Saldivia, R. Cardenas, J. W. Wragg, A.-M. Suzuki, P. Carninci, B. Peers, B. Lenhard, J. B. Andersen and F. Müller (2020). "Dual-initiation promoters with intertwined canonical and TCT/TOP transcription start sites diversify transcript processing." Nature Communications **11**(1): 168.

Nepal, C., Y. Hadzhiev, C. Previti, V. Haberle, N. Li, H. Takahashi, A. M. Suzuki, Y. Sheng, R. F. Abdelhamid, S. Anand, J. Gehrig, A. Akalin, C. E. Kockx, A. A. van der Sloot, W. F. van Ijcken, O. Armant, S. Rastegar, C. Watson, U. Strahle, E. Stupka, P. Carninci, B. Lenhard and F. Muller (2013). "Dynamic regulation of the transcription initiation landscape at single nucleotide resolution during vertebrate embryogenesis." Genome Res **23**(11): 1938-1950.

Newport, J. and M. Kirschner (1982). "A major developmental transition in early Xenopus embryos: I. characterization and timing of cellular changes at the midblastula stage." Cell **30**(3): 675-686.

Newport, J. and M. Kirschner (1982). "A major developmental transition in early xenopus embryos: II. control of the onset of transcription." Cell **30**(3): 687-696.

Nobrega, M. A., I. Ovcharenko, V. Afzal and E. M. Rubin (2003). "Scanning human gene deserts for long-range enhancers." Science **302**(5644): 413.

O'Farrell, P. H., J. Stumpff and T. T. Su (2004). "Embryonic cleavage cycles: how is a mouse like a fly?" Current biology : CB **14**(1): R35-R45.

Ogbourne, S. and T. M. Antalis (1998). "Transcriptional control and the role of silencers in transcriptional regulation in eukaryotes." The Biochemical journal **331 (Pt 1)**(Pt 1): 1-14.

Ohbayashi, T., Y. Makino and T.-a. Tamura (1999). "Identification of a mouse TBP-like protein (TLP) distantly related to the Drosophila TBP-related factor." Nucleic Acids Research **27**(3): 750-755.

Ohbayashi, T., M. Shimada, T. Nakadai, T. Wada, H. Handa and T. Tamura (2003). "Vertebrate TBP-like protein (TLP/TRF2/TLF) stimulates TATA-less terminal deoxynucleotidyl transferase promoters in a transient reporter assay, and TFIIA-binding capacity of TLP is required for this function." Nucleic acids research **31**(8): 2127-2133.

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

Okuda, Y., E. Ogura, H. Kondoh and Y. Kamachi (2010). "B1 SOX coordinate cell specification with patterning and morphogenesis in the early zebrafish embryo." PLoS Genet **6**(5): e1000936.

Okuda, Y., H. Yoda, M. Uchikawa, M. Furutani-Seiki, H. Takeda, H. Kondoh and Y. Kamachi (2006). "Comparative genomic and expression analysis of group B1 sox genes in zebrafish indicates their diversification during vertebrate evolution." Dev Dyn **235**(3): 811-825.

Olson, M. O. J., K. Hingorani and A. Szebeni (2002). Conventional and nonconventional roles of the nucleolus. International Review of Cytology. K. W. Jeon, Academic Press. **219:** 199-266.

Onichtchouk, D., F. Geier, B. Polok, D. M. Messerschmidt, R. Mossner, B. Wendik, S. Song, V. Taylor, J. Timmer and W. Driever (2010). "Zebrafish Pou5f1-dependent transcriptional networks in temporal control of early development." Mol Syst Biol **6**(1): 354.

Orkin, S. H. and K. Hochedlinger (2011). "Chromatin connections to pluripotency and cellular reprogramming." Cell **145**(6): 835-850.

Otto, C., P. F. Stadler and S. Hoffmann (2014). "Lacking alignments? The next-generation sequencing mapper segemehl revisited." Bioinformatics **30**(13): 1837-1843.

Owens, Nick D. L., Ira L. Blitz, Maura A. Lane, I. Patrushev, John D. Overton, Michael J. Gilchrist, Ken W. Y. Cho and Mustafa K. Khokha (2016). "Measuring Absolute RNA Copy Numbers at High Temporal Resolution Reveals Transcriptome Kinetics in Development." Cell Reports **14**(3): 632-647.

Pálfy, M., S. R. Joseph and N. L. Vastenhouw (2017). "The timing of zygotic genome activation." Current Opinion in Genetics & Development **43**: 53-60.

Pálfy, M., G. Schulze, E. Valen and N. L. Vastenhouw (2020). "Chromatin accessibility established by Pou5f3, Sox19b and Nanog primes genes for activity during zebrafish genome activation." PLOS Genetics **16**(1): e1008546.

Park, P. J. (2009). "ChIP–seq: advantages and challenges of a maturing technology." Nature Reviews Genetics **10**(10): 669-680.

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

Pauli, A., E. Valen, M. F. Lin, M. Garber, N. L. Vastenhouw, J. Z. Levin, L. Fan, A. Sandelin, J. L. Rinn, A. Regev and A. F. Schier (2012). "Systematic identification of long noncoding RNAs expressed during zebrafish embryogenesis." Genome Res **22**(3): 577-591.

Perry, R. P. (2005). "The architecture of mammalian ribosomal protein promoters." BMC Evolutionary Biology **5**(1): 15.

Persengiev, S. P., X. Zhu, B. L. Dixit, G. A. Maston, E. L. W. Kittler and M. R. Green (2003). "TRF3, a TATA-box-binding protein-related factor, is vertebrate-specific and widely expressed." Proceedings of the National Academy of Sciences of the United States of America **100**(25): 14887-14891.

Pombo, A. and N. Dillon (2015). "Three-dimensional genome architecture: players and mechanisms." Nature Reviews Molecular Cell Biology **16**(4): 245-257.

Ponjavic, J., B. Lenhard, C. Kai, J. Kawai, P. Carninci, Y. Hayashizaki and A. Sandelin (2006). "Transcriptional and structural impact of TATA-initiation site spacing in mammalian core promoters." Genome Biol **7**(8): R78.

Poulain, S., S. Kato, O. Arnaud, J. E. Morlighem, M. Suzuki, C. Plessy and M. Harbers (2017). "NanoCAGE: A Method for the Analysis of Coding and Noncoding 5'-Capped Transcriptomes." Methods Mol Biol. **1543**(1940-6029 (Electronic)): 57-109.

Prioleau, M. N., J. Huet, A. Sentenac and M. Mechali (1994). "Competition between Chromatin and Transcription Complex Assembly Regulates Gene-Expression during Early Development." Cell **77**(3): 439-449.

Qin, Y., L. K. Kong, C. Poirier, C. Truong, P. A. Overbeek and C. E. Bishop (2004). "Long-range activation of Sox9 in Odd Sex (Ods) mice." Hum Mol Genet **13**(12): 1213-1218.

Rabenstein, M. D., S. Zhou, J. T. Lis and R. Tjian (1999). "TATA box-binding protein (TBP)-related factor 2 (TRF2), a third member of the TBP family." Proceedings of the National Academy of Sciences of the United States of America **96**(9): 4791-4796.

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

Reynolds, G. A., S. K. Basu, T. F. Osborne, D. J. Chin, G. Gil, M. S. Brown, J. L. Goldstein and K. L. Luskey (1984). "HMG CoA reductase: A negatively regulated gene with unusual promoter and 5′ untranslated regions." Cell **38**(1): 275-285.

Riethoven, J.-J. M. (2010). Regulatory Regions in DNA: Promoters, Enhancers, Silencers, and Insulators. Computational Biology of Transcription Factor Binding. I. Ladunga. Totowa, NJ, Humana Press**:** 33-42.

Roberts, J. A., I. Miguel-Escalada, K. J. Slovik, K. T. Walsh, Y. Hadzhiev, R. Sanges, E. Stupka, E. K. Marsh, J. Balciuniene, D. Balciunas and F. Muller (2014). "Targeted transgene integration overcomes variability of position effects in zebrafish." Development **141**(3): 715-724.

Roeder, R. G. (1991). "The complexities of eukaryotic transcription initiation: regulation of preinitiation complex assembly." Trends Biochem Sci **16**(11): 402-408.

Roeder, R. G. (1996). "The role of general initiation factors in transcription by RNA polymerase II." Trends Biochem Sci **21**(9): 327-335.

Roepcke, S., D. Zhi, M. Vingron and P. F. Arndt (2006). "Identification of highly specific localized sequence motifs in human ribosomal protein gene promoters." Gene **365**: 48-56.

Rose, N. R. and R. J. Klose (2014). "Understanding the relationship between DNA methylation and histone lysine methylation." Biochimica et Biophysica Acta (BBA) - Gene Regulatory Mechanisms **1839**(12): 1362-1372.

Rossant, J., L. M. J. Nutter and M. Gertsenstein (2011). "Engineering the embryo." Proceedings of the National Academy of Sciences of the United States of America **108**(19): 7659-7660.

Sabari, B. R., A. Dall'Agnese, A. Boija, I. A. Klein, E. L. Coffey, K. Shrinivas, B. J. Abraham, N. M. Hannett, A. V. Zamudio, J. C. Manteiga, C. H. Li, Y. E. Guo, D. S. Day, J. Schuijers, E. Vasile, S. Malik, D. Hnisz, T. I. Lee, I. I. Cisse, R. G. Roeder, P. A. Sharp, A. K. Chakraborty and R. A. Young (2018). "Coactivator condensation at super-enhancers links phase separation and gene control." Science (New York, N.Y.) **361**(6400): eaar3958.

Salimullah, M., S. Mizuho, C. Plessy and P. Carninci (2011). "NanoCAGE: A High-Resolution Technique to Discover and Interrogate Cell Transcriptomes." Cold Spring Harbor Protocols **2011**(1): pdb.prot5559.

Sanyal, A., B. R. Lajoie, G. Jain and J. Dekker (2012). "The long-range interaction landscape of gene promoters." Nature **489**(7414): 109-113.

Sato, Y., L. Hilbert, H. Oda, Y. Wan, J. M. Heddleston, T. L. Chew, V. Zaburdaev, P. Keller, T. Lionnet, N. Vastenhouw and H. Kimura (2019). "Histone H3K27 acetylation precedes active transcription during zebrafish zygotic genome activation as revealed by live-cell analysis." Development **146**(19).

Satoh, N. (1977). "Metachronous Cleavage and Initiation of Gastrulation in Amphibian Embryos." Development Growth & Differentiation **19**(2): 111-117.

Saxonov, S., P. Berg and D. L. Brutlag (2006). "A genome-wide analysis of CpG dinucleotides in the human genome distinguishes two distinct classes of promoters." Proceedings of the National Academy of Sciences **103**(5): 1412.

Scerbo, P., F. Girardot, C. Vivien, G. V. Markov, G. Luxardi, B. Demeneix, L. Kodjabachian and L. Coen (2012). "Ventx Factors Function as Nanog-Like Guardians of Developmental Potential in Xenopus." PLOS ONE **7**(5): e36855.

Schulz, K. N., E. R. Bondra, A. Moshe, J. E. Villalta, J. D. Lieb, T. Kaplan, D. J. McKay and M. M. Harrison (2015). "Zelda is differentially required for chromatin accessibility, transcription factor binding, and gene expression in the early Drosophila embryo." Genome Research **25**(11): 1715-1726.

Schulz, K. N. and M. M. Harrison (2019). "Mechanisms regulating zygotic genome activation." Nature Reviews Genetics **20**(4): 221-234.

Shi, W., D. Hendrix, M. Levine and B. Haley (2009). "A distinct class of small RNAs arises from pre-miRNA–proximal regions in a simple chordate." Nature Structural & Molecular Biology **16**(2): 183-189.

Shih, J. and S. E. Fraser (1996). "Characterizing the zebrafish organizer: microsurgical analysis at the early-shield stage." Development **122**(4): 1313.

Shindo, Y. and A. A. Amodeo (2019). "Dynamics of Free and Chromatin-Bound Histone H3 during Early Embryogenesis." Current Biology **29**(2): 359-366.e354.

Shiraki, T., 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 (2003). "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.

Shmakov, S., O. O. Abudayyeh, K. S. Makarova, Y. I. Wolf, J. S. Gootenberg, E. Semenova, L. Minakhin, J. Joung, S. Konermann, K. Severinov, F. Zhang and E. V. Koonin (2015). "Discovery and Functional Characterization of Diverse Class 2 CRISPR-Cas Systems." Molecular cell **60**(3): 385-397.

Shrinivas, K., B. R. Sabari, E. L. Coffey, I. A. Klein, A. Boija, A. V. Zamudio, J. Schuijers, N. M. Hannett, P. A. Sharp, R. A. Young and A. K. Chakraborty (2019). "Enhancer Features that Drive Formation of Transcriptional Condensates." Molecular Cell **75**(3): 549-561.e547.

Siefert, J. C., C. Georgescu, J. D. Wren, A. Koren and C. L. Sansam (2017). "DNA replication timing during development anticipates transcriptional programs and parallels enhancer activation." Genome Res **27**(8): 1406-1416.

Sievers, F. and D. G. Higgins (2014). Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences. Multiple Sequence Alignment Methods. D. J. Russell. Totowa, NJ, Humana Press**:** 105-116.

Sigal, Y. M., R. Zhou and X. Zhuang (2018). "Visualizing and discovering cellular structures with super-resolution microscopy." Science **361**(6405): 880.

Signoret, J. and J. Lefresne (1971). "Contribution à l'étude de la segmentation de l'œuf d'axolotl. I-Définition de la transition blastuléenne. ." Ann. Embryol. Morph. **4**: 113-123.

Solomon, M. J., P. L. Larsen and A. Varshavsky (1988). "Mapping proteinDNA interactions in vivo with formaldehyde: Evidence that histone H4 is retained on a highly transcribed gene." Cell **53**(6): 937-947.

Song, Y., R. A. Marmion, J. O. Park, D. Biswas, J. D. Rabinowitz and S. Y. Shvartsman (2017). "Dynamic Control of dNTP Synthesis in Early Embryos." Developmental Cell **42**(3): 301-308.e303.

Soufi, A., Meilin F. Garcia, A. Jaroszewicz, N. Osman, M. Pellegrini and Kenneth S. Zaret (2015). "Pioneer Transcription Factors Target Partial DNA Motifs on Nucleosomes to Initiate Reprogramming." Cell **161**(3): 555-568.

Spemann, H. and H. Mangold (1924). "über Induktion von Embryonalanlagen durch Implantation artfremder Organisatoren." Archiv für mikroskopische Anatomie und Entwicklungsmechanik **100**(3): 599-638.

Suh, M. R., Y. Lee, J. Y. Kim, S. K. Kim, S. H. Moon, J. Y. Lee, K. Y. Cha, H. M. Chung, H. S. Yoon, S. Y. Moon, V. N. Kim and K. S. Kim (2004). "Human embryonic stem cells express a unique set of microRNAs." Dev Biol **270**(2): 488-498.

Sun, F.-L. and S. C. R. Elgin (1999). "Putting Boundaries on Silence." Cell **99**(5): 459-462.

Sun, Y. J., C. Y. Nien, K. Chen, H. Y. Liu, J. Johnston, J. Zeitlinger and C. Rushlow (2015). "Zelda overcomes the high intrinsic nucleosome barrier at enhancers during Drosophila zygotic genome activation." Genome Research **25**(11): 1703-1714.

Suzuki, M. M. and A. Bird (2008). "DNA methylation landscapes: provocative insights from epigenomics." Nature Reviews Genetics **9**(6): 465-476.

Symmons, O., V. V. Uslu, T. Tsujimura, S. Ruf, S. Nassari, W. Schwarzer, L. Ettwiller and F. Spitz (2014). "Functional and topological characteristics of mammalian regulatory domains." Genome Res **24**(3): 390-400.

Tadros, W. and H. D. Lipshitz (2009). "The maternal-to-zygotic transition: a play in two acts." Development **136**(18): 3033-3042.

Takahashi, H., S. Kato, M. Murata and P. Carninci (2012). "CAGE (cap analysis of gene expression): a protocol for the detection of promoter and transcriptional networks." Methods Mol Biol **786**: 181-200.

Takahashi, K. and S. Yamanaka (2006). "Induction of pluripotent stem cells from mouse embryonic and adult fibroblast cultures by defined factors." Cell **126**(4): 663-676.

Tang, F., M. Kaneda, D. O'Carroll, P. Hajkova, S. C. Barton, Y. A. Sun, C. Lee, A. Tarakhovsky, K. Lao and M. A. Surani (2007). "Maternal microRNAs are essential for mouse zygotic development." Genes Dev **21**(6): 644-648.

Tani, S., R. Kusakabe, K. Naruse, H. Sakamoto and K. Inoue (2010). "Genomic organization and embryonic expression of miR-430 in medaka (Oryzias latipes): insights into the post-transcriptional gene regulation in early development." Gene **449**(1-2): 41-49.

Tolhuis, B., R. J. Palstra, E. Splinter, F. Grosveld and W. de Laat (2002). "Looping and interaction between hypersensitive sites in the active beta-globin locus." Mol Cell **10**(6): 1453-1465.

Treangen, T. J. and S. L. Salzberg (2012). "Repetitive DNA and next-generation sequencing: computational challenges and solutions." Nature Reviews Genetics **13**(1): 36-46.

Trinklein, N. D., S. F. Aldred, S. J. Hartman, D. I. Schroeder, R. P. Otillar and R. M. Myers (2004). "An Abundance of Bidirectional Promoters in the Human Genome." Genome Research **14**(1): 62-66.

Tyagi, S. and F. R. Kramer (1996). "Molecular Beacons: Probes that Fluoresce upon Hybridization." Nature Biotechnology **14**(3): 303-308.

Vastenhouw, N. L., W. X. Cao and H. D. Lipshitz (2019). "The maternal-to-zygotic transition revisited." Development **146**(11).

Veenstra, G. J., O. H. Destrée and A. P. Wolffe (1999). "Translation of maternal TATA-binding protein mRNA potentiates basal but not activated transcription in Xenopus embryos at the midblastula transition." Molecular and cellular biology **19**(12): 7972-7982.

Veil, M., L. Y. Yampolsky, B. Gruning and D. Onichtchouk (2019). "Pou5f3, SoxB1, and Nanog remodel chromatin on high nucleosome affinity regions at zygotic genome activation." Genome Res **29**(3): 383-395.

Vermeulen, M., K. W. Mulder, S. Denissov, W. W. Pijnappel, F. M. van Schaik, R. A. Varier, M. P. Baltissen, H. G. Stunnenberg, M. Mann and H. T. Timmers (2007). "Selective anchoring of TFIID to nucleosomes by trimethylation of histone H3 lysine 4." Cell **131**(1): 58-69.

Wang, Y., R. C. Jensen and W. E. Stumph (1996). "Role of TATA box sequence and orientation in determining RNA polymerase II/III transcription specificity." Nucleic Acids Res **24**(15): 3100-3106.

Watanabe, T., A. Takeda, K. Mise, T. Okuno, T. Suzuki, N. Minami and H. Imai (2005). "Stage-specific expression of microRNAs during Xenopus development." FEBS Lett **579**(2): 318-324.

Weil, P. A., D. S. Luse, J. Segall and R. G. Roeder (1979). "Selective and Accurate Initiation of Transcription at the Ad2 Major Late Promotor in a Soluble System Dependent on Purified Rna Polymerase-Ii and DNA." Cell **18**(2): 469-484.

White, R. J., J. E. Collins, I. M. Sealy, N. Wali, C. M. Dooley, Z. Digby, D. L. Stemple, D. N. Murphy, K. Billis, T. Hourlier, A. Fullgrabe, M. P. Davis, A. J. Enright and E. M. Busch-Nentwich (2017). "A high-resolution mRNA expression time course of embryonic development in zebrafish." Elife **6**: e30860.

Whyte, Warren A., David A. Orlando, D. Hnisz, Brian J. Abraham, Charles Y. Lin, Michael H. Kagey, Peter B. Rahl, Tong I. Lee and Richard A. Young (2013). "Master Transcription Factors and Mediator Establish Super-Enhancers at Key Cell Identity Genes." Cell **153**(2): 307-319.

Wieczorek, E., M. Brand, X. Jacq and L. Tora (1998). "Function of TAFII-containing complex without TBP in transcription by RNA polymerase II." Nature **393**(6681): 187-191.

Wiekowski, M., M. Miranda and M. L. DePamphilis (1993). "Requirements for promoter activity in mouse oocytes and embryos distinguish paternal pronuclei from maternal and zygotic nuclei." Dev Biol **159**(1): 366-378.

Wierson, W. A., J. M. Welker, M. P. Almeida, C. M. Mann, D. A. Webster, M. E. Torrie, T. J. Weiss, S. Kambakam, M. K. Vollbrecht, M. Lan, K. C. McKeighan, J. Levey, Z. Ming, A. Wehmeier, C. S. Mikelson, J. A. Haltom, K. M. Kwan, C.-B. Chien, D. Balciunas, S.

C. Ekker, K. J. Clark, B. R. Webber, B. S. Moriarity, S. L. Solin, D. F. Carlson, D. L. Dobbs, M. McGrail and J. Essner (2020). "Efficient targeted integration directed by short homology in zebrafish and mammalian cells." eLife **9**: e53968.

Williamson, I., L. Kane, P. S. Devenney, I. M. Flyamer, E. Anderson, F. Kilanowski, R. E. Hill, W. A. Bickmore and L. A. Lettice (2019). "Developmentally regulated Shh expression is robust to TAD perturbations." Development **146**(19): dev179523.

Woodcock, C. L. (2006). "Chromatin architecture." Current Opinion in Structural Biology **16**(2): 213-220.

Woodland, H. R. and E. D. Adamson (1977). "The synthesis and storage of histones during the oogenesis of Xenopus laevis." Developmental Biology **57**(1): 118-135.

Wragg, J. and F. Müller (2016). Chapter Six - Transcriptional Regulation During Zygotic Genome Activation in Zebrafish and Other Anamniote Embryos. Advances in Genetics. N. S. Foulkes, Academic Press. **95:** 161-194.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman and L. A. Romano (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol **20**(9): 1377-1419.

Xu, C., Z. P. Fan, P. Muller, R. Fogley, A. DiBiase, E. Trompouki, J. Unternaehrer, F. Xiong, I. Torregroza, T. Evans, S. G. Megason, G. Q. Daley, A. F. Schier, R. A. Young and L. I. Zon (2012). "Nanog-like regulates endoderm formation through the Mxtx2-Nodal pathway." Dev Cell **22**(3): 625-638.

Yamashita, R., Y. Suzuki, S. Sugano and K. Nakai (2005). "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.

Yang, L.-Z., Y. Wang, S.-Q. Li, R.-W. Yao, P.-F. Luan, H. Wu, G. G. Carmichael and L.-L. Chen (2019). "Dynamic Imaging of RNA in Living Cells by CRISPR-Cas13 Systems." Molecular Cell **76**(6): 981-997.e987.

Yartseva, V. and A. J. Giraldez (2015). "The Maternal-to-Zygotic Transition During Vertebrate Development: A Model for Reprogramming." Curr Top Dev Biol **113**: 191-232.

Zaret, K. S. and J. S. Carroll (2011). "Pioneer transcription factors: establishing competence for gene expression." Genes Dev **25**(21): 2227-2241.

Zeng, F. and R. M. Schultz (2005). "RNA transcript profiling during zygotic gene activation in the preimplantation mouse embryo." Developmental Biology **283**(1): 40-57.

Zhao, X. and W. Herr (2002). "A Regulated Two-Step Mechanism of TBP Binding to DNA: A Solvent-Exposed Surface of TBP Inhibits TATA Box Recognition." Cell **108**(5): 615-627.

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

Zheng, H. and W. Xie (2019). "The role of 3D genome organization in development and cell differentiation." Nature Reviews Molecular Cell Biology **20**(9): 535-550.

Zhu, W., X. Xu, X. Wang and J. Liu (2019). "Reprogramming histone modification patterns to coordinate gene expression in early zebrafish embryos." BMC Genomics **20**(1): 248.

LIST OF PUBLICATIONS

Zaucker A, A. Nagorska, P. Kumari, N. Hecker, Y. Wang, S. Huang, **L. Cooper**, L. Sivashanmugam, S. VijayKumar, J. Brosens, J. Gorodkin, K. Sampath (2018). “Translational co-regulation of a ligand and inhibitor by a conserved RNA element.” Nucleic Acids Res. **46**(1):104-119.

Hadzhiev, Y., H. K. Qureshi, L. Wheatley**, L. Cooper**, A. Jasiulewicz, H. Van Nguyen, J. W. Wragg, D. Poovathumkadavil, S. Conic, S. Bajan, A. Sik, G. Hutvàgner, L. Tora, A. Gambus, J. S. Fossey and F. Müller (2019). "A cell cycle-coordinated Polymerase II transcription compartment encompasses gene expression before global genome activation." Nature Communications **10**(1): 691.

Westhoff, J. H., P. J. Steenbergen, L. S. V. Thomas, J. Heigwer, T. Bruckner, **L. Cooper**, B. Tönshoff, G. F. Hoffmann and J. Gehrig. “In vivo high-content screening in zebrafish for developmental nephrotoxicity of approved drugs.” Frontiers in Cell and Developmental Biology (unpublished)