Imperial College London

Institute of Clinical Sciences

# Promoter architecture and gene expression dynamics in embryonic development

Dunja Vučenović

October 2019

Submitted in part fulfilment of the requirements for the degree of

Doctor of Philosophy of Imperial College London

# Declaration

I hereby declare that the work presented in this thesis is my own, and that work carried out by others has been acknowledged. To the best of my knowledge, it contains no material which has been accepted for any other degree of any university or other institute of higher learning.

# Abstract

Genes indispensable for proper embryonic development show intricate patterns of expression throughout the time, space and magnitude of their activity. This diversity is enabled by elaborate regulatory mechanisms that guide their expression. They also possess a distinct type of core promoters that enable the integration of all regulatory inputs. However, it is still not clear how is coordination of regulation achieved. The first step towards understanding this process is to characterise dynamics of expression, and core promoter features that process the regulation.

In this thesis, I explored the diversity of spatio-temporal gene expression during zebrafish development. I defined a novel measure of anatomical specificity that defines how precisely an anatomical structure is defined in the Anatomical Ontology system. Using anatomical specificity measure, I quantified gene expression dynamics from mRNA *in situ* hybridisation data. Gene expression divergence from *in situ*s was used to predict expression levels from RNA-seq expression data. This analysis allowed me to propose a measure of gene expression complexity which showed that genes with the highest complexity score are developmental genes, whereas genes with low complexity score are involved in housekeeping functions. Next, I developed a method that reports significantly enriched core promoter elements in a group of genes. Using this method, I compared differences in core promoter composition in active genes expressed in different developmental periods. In addition, this method found groups of genes with a specific core promoter structure that are specified for a biological process. Finally, I used scRNA-seq data from zebrafish development to identify patterns of gene co-expression across different cell clusters. Co-expression suggests that a gene pair possesses a common regulatory programme. I show that genes with the most divergent co-expression patterns across development are developmental genes and that housekeeping genes have least diverse co-expression patterns. I went further to create co-expression networks which allowed me to analyse co-expression patterns into more details.

# Acknowledgements

This thesis would certainly not have been possible without the help, support and guidance of many amazing people. I have been fortunate to be surrounded by a fantastic crew.

Firstly, I would like to thank Boris for giving me the opportunity to join his lab and do this PhD. Thank you for the patient guidance and advice you provided throughout my time in the lab. Being a part of ZENCODE and Lenhard group was a real treat.

My PhD journey was a pleasure mainly because of the amazing members of the Lenhard group. Anja and Piotr, thank you for introducing me to zebrafish genomics. Anja, thanks for caring and for giving me that little extra push when I needed it the most. I immensely enjoyed discussing the bigger picture of our projects. Piotr, thank you for always being there to answer questions regarding biology, sysadmin or where next to travel. Liz and Malcolm, thanks for your help during the initial stages of my PhD. Alex, you were a great deskmate and you set the bar so high! It was great having your around. Nev, you helped me so many times, that I cannot even count. You truly are an inspiration for every young scientist. Thank you for everything. Damir your enthusiasm for research is unbeatable. It was a joy working with you. Leonie, I really admire your determination and organisational skills. One day I might be more like you =) Eleni, thanks for all the laughs and cakes. Radina, I hope you are much more productive now that I am not in the office! Keep up the good work. Finally, James, thanks for all distractions, curry breads and super interesting abstract chats.

Being a part of ZENCODE-ITN was an enriching experience. I thank everyone who took part in the organisation of many meetings and trainings. My fellow Zenkids were a great company and remain to be good friends.

During these four years, I was fortunate to make some very dear friends. Angela and Marco, thank you for all the food and fun that we shared along these years, you really made my life so much more pleasurable. Claudia, I enjoyed our super beautiful walks in the morning so

much that it almost made me a morning person! My S3 family that showed me how our hobbies can have a great impact on others. Thank you Petra, Marieke, Leo, and co. for that eye-opener. Also, thanks to my Illumina folks for all the support in the last months of my writing and for constantly asking if I am already finished.

Najljepše hvala mojoj familiji koja je uvijek tu samnom. Hvala vam na brizi, podršci i poticaju čak i u trenutcima kad niste znali čemu sve to.

Na poslijetku, mojoj prasenčadi doma koji mi svakodnevno uljepšavaju dane. Davore, bez tebe ovo sve ne bi ništa značilo. Hvala ti na svoj podršci i povjerenju. Uvijek i zauvijek.

*Mojim roditeljcima*

# Contents

# List of Figures

# List of Tables

# List of abbreviations

| **Abbreviation** | **Meaning** |
|---|---|
| 2D | Two dimensional |
| 3D | Three dimensional |
| 5’ | Five prime, the asymmetric end of a DNA strand possessing a terminal phosphate group |
| AO | Anatomical Ontology |
| BP | Biological process |
| bp | Base pairs |
| BRE | B recognition element |
| CAGE | Cap analysis of gene expression |
| CC | Cellular component |
| CGI | CpG island |
| ChIP | Chromatin immunoprecipitation |
| CNE | Conserved non-coding element |
| CpA | Cytosine followed by an adenine in 5’ to 3’ direction |
| CpG | Cytosine followed by a guanine in 5’ to 3’ direction |
| CRM | Cis-regulatory module |
| CTSS | CAGE transcription start site |
| DNA | Deoxyribonucleic acid |
| DPE | Downstream promoter element |
| DRE | DNA replication-related element |
| dTSS | Dominant transcription start site |
| EFA | Exploratory factor analysis |
| ENA | European Nucleotide Archive |
| ESC | Embryonic stem cell |
| FANTOM5 | Functional annotation of the mammalian genome project, phase 5 |

| Abbreviation | Meaning |
|---|---|
| GM12878 | B-lymphocyte cell line transformed by Epstein-Barr Virus |
| GO | Gene ontology |
| GTEx | The Genotype-Tissue Expression project |
| CGI | CpG island |
| HCNE | Highly conserved non-coding element |
| INR | Initiator |
| IQ | Interquantile |
| IQR | Interquantile range |
| ISH | In situ hybridisation |
| knn | K-nearest neighbors |
| MF | Molecular function |
| MI | Mutual information |
| MTE | Motif ten element |
| MZT | Maternal-to-zygotic transition |
| NGS | Next generation sequencing |
| PCA | Principal component analysis |
| PIC | Pre-initiation complex |
| PO | Promoter Ontology |
| PolII | RNA polymerase II |
| PWM | Position weight matrix |
| RNA | Ribonucleic acid |
| scRNA-seq | Single-cell RNA sequencing |
| TBP | TATA binding protein |
| TC | Tag cluster |
| TF | Transcription factor |
| TFBS | Transcription factor binding site |
| TFIID | Transcription factor II D |

| **Abbreviation** | **Meaning** |
|---|---|
| TOM | Topological overlap matrix |
| tpm | Tag per million |
| tSNE | T-distributed stochastic neighbour embedding |
| TSS | Transcription start site |
| UCSC | University of California, Santa Cruz |
| UMI | Unique molecular identifier |
| WGCNA | Weighted gene co-expression network analysis |
| ZFIN | The Zebrafish Information Network |
| ZGA | Zygotic genome activation |

# 1 Introduction

## 1.1 Embryonic development

Every animal develops from a fertilised egg cell into a fully grown organism that consists of functionally diverse tissues and cell types, each optimised for their specific function. This stepwise, massively parallel series of events, called embryonic development, is a carefully regulated process mainly determined by the genetic information contained in DNA (Levine and Tjian 2003). Despite the differences observed in phenotypes of adult cells, a large majority of the cells within an organism possess the same copy of DNA. DNA contains sequences for protein-coding genes as well as non-coding regulatory sequences which instruct the timing, location and the magnitude of expression for each gene. The process of converting sequence information encoded in DNA into a functional product is called gene expression. The quantity of functional product produced in space and time is determined by both internal and external signals that act on the transcriptional or post-transcriptional level. Precise regulation of gene expression is a key element enabling proper functioning and development of an organism.

During the first hours of animal development, the expression of zygotic genes is completely silenced. In this period, reprogramming of germ cells takes place and the embryo functions by using proteins and RNAs that were maternally supplied in the oocyte (Vastenhouw, Cao, and Lipshitz 2019). Nevertheless, for the full development of an embryo, activation of the zygotic genome is required. The process in which maternally deposited transcripts are degraded and transcriptional control is transferred to the embryo is called maternal-to-zygotic transition (MZT). The MZT is conserved across metazoans and in many, this transition takes place simultaneously with the mid-blastula transition (Tadros and Lipshitz 2009). Just after fertilisation, embryos undergo rapid cell divisions. During this time, the cell cycle consists only of DNA replication (S phase) and mitosis, while the gap phases, necessary for protein synthesis, do not occur (Yuan et al. 2016). After MZT cell cycle slows down and gap phases (G1 and G2), which allow growth, get introduced. These transitions in the early embryo are preparing it for gastrulation, a process in which cells start migrating and differentiating into germ layers and cells migrate and self-organise to form the

embryo body (Tadros and Lipshitz 2009).

Gastrulation is a key period of animal embryogenesis that shapes the external and internal features of the animal. During gastrulation, seemingly identical, unstructured cells transform into three germ layers - endoderm, mesoderm and ectoderm. Ectodermal cells will remain on the surface of the embryo and will give rise to neural tissues and the epidermis. On the other hand, mesodermal and endodermal cells will move through the embryo and become internalised. The mesoderm will later form muscles, as well as the skeletal and cardiovascular systems, while endoderm will generate the digestive tube and its accessory organs.

The underlying genetic and molecular mechanisms that guide these processes are highly conserved and include a cascade of signalling molecules, signalling pathways and transcription factors. Their action specifies cell fate and guides cell movements (De Robertis et al. 2000). For example, induction of mesoderm and endoderm in all vertebrates is driven by Nodal signalling by ligands from the TGF$\beta$ superfamily (Schier 2004). Spatio-temporal expression of many genes regulates this process. Deciphering regulation of spatio-temporal expression is essential for understanding processes driving embryonic development.

### 1.1.1 Developmental genes and their expression regulation

Surprisingly, it was shown that early developmental processes occur by similar principles across the animal kingdom. In addition, there is a relatively small number of developmental pathways that are driving development which are highly conserved across the animal kingdom (Wolpert 1994). A great proportion of genes involved in these developmental pathways encode for transcription factors and signalling molecules. Developmental genes are important for cell fate specification and pattern formation. These developmental genes control patterning and large-scale cell movements to spatially shape the embryo. These genes were identified during genetic screens since embryos with aberrant developmental gene expression would show mutant phenotype (Dickmeis and Muller 2005). In addition, important developmental genes can be discovered by the presence of highly conserved stretches of DNA in their neighbourhood (Bejerano et al. 2004; Sandelin, Bailey, et al. 2004). Surprisingly, both of these methods identified only a limited number of developmental genes.

The first master regulator of MZT identified was Zelda (Liang et al. 2008). Zelda may be considered as a "pioneer" transcription factor (TF) since it is able to bind nucleosomal DNA and in that way locally reduce nucleosome occupancy. This, in turn, enables other TSs to bind enhancers. (Sun et al. 2015). It is active in flies and it is maternally deposited. An hour before MZT, it gets highly translationally upregulated after which it promotes activation of hundreds of genes. Zelda orthologs in mammals, fish or amphibians have not been found, so it is unknown if such an activation mechanism exists in vertebrates. Despite that, two studies found that Nanog, SoxB1 and Pou5f3 work together to activate zygotic genome activation (ZGA) (Lee et al. 2013; Leichsenring et al. 2013). One of the reasons for their identification was the fact that these transcription factors are the most highly translated factors after fertilisation in zebrafish (Lee et al. 2013). Even though they are not related to Zelda, they are, however, homologs of mammalian pluripotency factors NANOG, SOX2 and OCT4 which are well known for their ability to transform differentiated cells to a stem cell like cells (Takahashi and Yamanaka 2016).

After comparing sequences of developmental genes from different organisms, it was found that a great proportion of developmental transcription factor genes are spanned by clusters of highly conserved non-coding elements (HCNEs) (Sandelin, Bailey, et al. 2004; Woolfe et al. 2004). HCNEs are thought to have an indispensable role in animal development by providing extra regulatory inputs for developmental genes. Many of these elements were tested for the ability to drive expression by inserting them into a vector carrying a minimal promoter (Pennacchio et al. 2006; Visel et al. 2007; Ellingsen et al. 2005). Through experiments in transgenic model organisms, it was shown that many of HCNEs are able to induce a segment of overall developmental expression of a developmental gene in the neighbourhood (Nobrega et al. 2003; Pennacchio et al. 2006; Woolfe et al. 2004). Individual were able to only partially reproduce the spatio-temporal expression pattern of their target gene, suggesting that their combined regulatory input is required to fully reconstruct complex expression profiles. These experiments showed that even HCNEs that are far apart from their developmental gene in the linear genome are able to regulate its expression, suggesting that these elements can be long-range regulators. Indeed, model organisms have shown that many HCNEs are developmental enhancers (Calle-Mustienes et al. 2005; Visel et al. 2007; Woolfe et al. 2007).

### 1.1.2 Zebrafish as a model organism for embryonic development

Pattern formation taking place during early development is driven by similar principles across the animal kingdom. There are a small number of developmental pathways active in an organism and they are highly conserved across diverse organisms like humans, mice, zebrafish or fly (Wolpert 1994). This finding was the base of a new scientific discipline, comparative genomics. The findings about developmental processes and disease in model organisms were a foundation for understanding human development and disease. The zebrafish, *Danio rerio*, is an important model organism for understanding development (Zacchigna, Ruiz de Almodovar, and Carmeliet 2008). In the past 30 years, the development of methods such as gene knockdowns and embryo microinjections enabled many studies to analyse specific processes and their genetic underlining in development. Integration of all these studies provides a rich resource to learn more about embryogenesis, the neuronal system and disease.

Zebrafish is a freshwater fish, inhabiting rivers of the Himalayan region of South Asia. The first time zebrafish was used as a model organism was by George Streisinger (University of Oregon) in the 1970s. Zebrafish was much easier to genetically manipulate and maintain than mouse. In the 1990s, after two large genetic screens were published (Mullins et al. 1994; Solnica-Krezel, Schier, and Driever 1994), zebrafish became a common model organism for development. *D. rerio* is preferred due to several key features. The embryo develops rapidly and is translucent for the first two days of embryogenesis. Translucency makes experiments and resulting observation easily achievable. In addition, the embryo develops rapidly. Blastulation lasts for three hours, while gastrulation takes five hours. By 24 hours post fertilisation segmentation is complete, and all primary organ systems are formed. In just four days the embryo goes from zygote to a juvenile fish.

Due to the importance of zebrafish as a model organism, multiple databases have been created to facilitate studies between the community. One such example is the Zebrafish Model Information Network (ZFIN) (Ruzicka et al. 2015), a database that keeps track of all studies performed in zebrafish. It reports investigated genes, periods of their temporal expression, as well as anatomical structures in which gene expression was observed. The expression information is annotated using the Annotation Ontology (AO) system (Ciccarese et al. 2011) - a controlled

hierarchical vocabulary describing anatomical structures of zebrafish and their relationships with other structures. Zebrafish AO system can be used as a reference for anatomical descriptions, and it enables associating homologous structures between anatomical ontologies of different species (Van Slyke et al. 2014).

## 1.2 Spatio-temporal gene expression

After understanding how spatio-temporal regulation of gene expression is fundamental for the development of highly specialised cells, developmental biologists were among the first to recognise its importance. For this reason, numerous methods for identification of gene expression localisation have been developed. *In situ* hybridization (ISH) assays for localisation of expression in the cellular environment are particularly popular (Armant et al. 2013; Thisse and Thisse 2008). In ISH techniques, a labelled RNA or DNA probe can be used to hybridise to complementary sequences in the tissues of interest. Labelled hybrids can be visualised by microscopy. The targeted nucleic acid is retained *in situ*, thereby allowing to visualise the spatial distribution of transcripts in their original state in developing embryos. Performing ISH at different stages of development allows for both the temporal and spatial expression of a gene to be investigated.

A large scale *in situ* hybridization study analysing localisation of 6003 genes in *Drosophila* embryo was performed (Tomancak et al. 2007). A study of similar scope was performed in zebrafish by Thisse et al. (Thisse et al. 2001; Tomancak et al. 2007). The results of these studies showed how, based on their expression pattern, genes can be clustered into distinct groups. Tomancak et al. showed how 34% of all genes show spatially restricted expression. Another 46% of genes are expressed uniformly, while for 19%, expression was not observed. In addition, they have concluded that transcription factor genes are highly enriched in the spatially restricted gene cluster throughout the majority of development. TFs and signalling molecules were not found among uniformly expressed genes. More recently, developmental time course expression has been systematically examined. White *et al.* have generated a high-resolution mRNA expression reference for 23,642 zebrafish genes (White et al. 2017). Again, the relationship between the expression dynamics of a gene and its importance in development was prominent. Genes encoding for key transcription

factors during development showed sharp changes in expression levels in time and space.

An illustrative example of the importance of appropriate gene expression pattern was published by Kellerer et al (Kellerer et al. 2006). In this study, authors deleted transcription factor Sox10 in mouse and replaced it with Sox8, a closely related paralogous gene whose expression patterns were overlapping in some embryonic segments during development. Furthermore, both genes regulate oligodendrocyte differentiation and development of the enteric neuronal system. Despite their similarities, inserted Sox8 was able to rescue the loss of Sox10 in a subset of tissues both genes are active in. Development of sensory neurons and glial cells was not affected, but melanocyte development was defective, and the development of the enteric neuronal system was limited. It is still not clear what is the cause for the observed functional inequivalence, however, these differences could be caused by differential patterns of posttranslational modifications. For example, it was shown that Sox9 and Sox10 proteins are sumoylated on the N-terminal site, while this was not observed in Sox8 (Girard and Goossens 2006). These results clearly showed how developmental genes can have distinct roles in different tissues and that functional equivalence of genes is dependent on the cell type being assayed.

Based on the pattern of gene expression, it is possible to categorize genes into distinct groups:

- The first group consists of genes uniformly expressed throughout an organism. These genes are required for the maintenance of basal cellular functions and are therefore called housekeeping genes. They are essential for the survival of the cell, regardless of cell's specific role in the organism. Housekeeping genes are therefore expressed in all cells in the organism with similar quantities under normal conditions irrespective of cell type, cell stage or external signals. When analysing chromatin structure across the genome, Dixon et al. reported that TSSs of housekeeping genes have a strong localisation preference for TAD boundaries (Dixon et al. 2012). Promoters of these genes are less conserved than those of developmental genes. Their regulation is less complex than developmental genes as well, but it is still unclear how ubiquitous expression is achieved. One hypothesis suggests that housekeeping genes are regulated by a special type of enhancers which are not cell type-specific (Zabidi et al.

2015). Housekeeping genes have important applications as internal controls for computational and experimental analyses. By analysing RNA-seq across many cell types, Eisenberg et al. identified 3804 human housekeeping genes (Eisenberg and Levanon 2013).

- The second group consists of genes whose expression depends on the cell type and often on the developmental period. Such genes are called tissue-specific genes as their expression in a subset of tissues is significantly higher than baseline expression across all tissues. Because of their specification for a tissue, they are often critical for the biological processes unique to those tissues. A deeper understanding of molecular processes in those tissues was obtained after studying features of tissue specific genes (Odom et al. 2004). Nevertheless, a comprehensive list of tissue-specific genes is challenging to generate. Initially, ISH analyses were used to define if a gene is tissue-specific, however, a single experiment can define specificity only for a single gene. Development of high-throughput methods made it possible to create gene expression atlas across multiple tissues. Early studies that used microarrays suffered from limited power to detect tissue specific genes because they were able to analyse a limited number of samples and tissues (Su et al. 2004). Development of RNA-seq enabled more detailed analysis of transcriptomes, hence these studies improved power and sensitivity to define tissue specific genes (Mortazavi et al. 2008). More recently, The Genotype Tissue Expression (GTEx) project generated the largest collection of transcriptomes from 38 different human tissues (GTEx Consortium 2013; Yang et al. 2018). This dataset was then used to define tissue specific genes and understand their gene regulation (Sonawane et al. 2017). In addition, tissue-specific genes are gaining attention as desirable targets for drug development. Selective expression in one or a few tissues means that there is a reduced risk of side effects in comparison to targeting a broadly expressed gene. In 2008 quantitative analysis of microarray expression data showed how tissue-specific genes are twice as likely to become drug targets in comparison to housekeeping genes (Dezso et al. 2008).

- A final group of genes that could be derived from their particular expression pattern belongs to developmental genes. More about specific features of developmental genes can be found in previous sections of this Introduction.

## 1.3 Promoters

Promoters are genomic sites that define where transcription starts. They are fundamental to genome function. Promoter regions span the transcription start site (TSS) and are recognised by proteins of the transcription initiation complex (Figure 1.1A). Promoters position RNA polymerase and its associated general transcription factors, enabling transcription to initiate at the correct site. In addition, integration of all regulatory inputs takes place at the promoters. A rate of transcription is going to be defined based on all TFs and histone modifications at all regulatory elements targeting this gene.

Transcription initiates at a narrow promoter region overlapping the TSS called core promoter. Core promoters encompass 40-100 base pairs (bp) upstream and downstream from TSS. Many core promoters alone are sufficient to drive transcription initiation, but this expression is usually low level and with low diversity among cells (Kadonaga 2012). For full expression repertoire, promoters most likely require additional inputs from chromatin marks or distal regulatory elements.

Metazoan core promoters can harbour DNA sequence patterns that are often found around the binding site of the transcription initiation complex. A schematic representation of a core promoter and its sequence motifs can be found in Figure 1.1B. Some of these motifs are recognised and bound by proteins (TATA-box), while others have no known interactor (such as downstream promoter element (DPE)). Presence of different core promoter elements define promoter architectures which, in turn, determine a promoter's susceptibility to long-range regulation and developmental programs (Haberle and Lenhard 2016).

Promoter structure and activity can be analysed by methods that map TSSs including Cap analysis gene expression (CAGE) (Kodzius et al. 2006) and CapSeq (Gu et al. 2012). These methods select capped mRNA molecules whose 5' ends are later sequenced. This way, we are able to obtain an accurate identification of transcription initiation location genome-wide. Development of high-throughput methods for promoter analysis enabled finding promoter-associated features and their impact on gene expression.

Different types of genes possess promoters with distinct sequence motifs. A characteristic

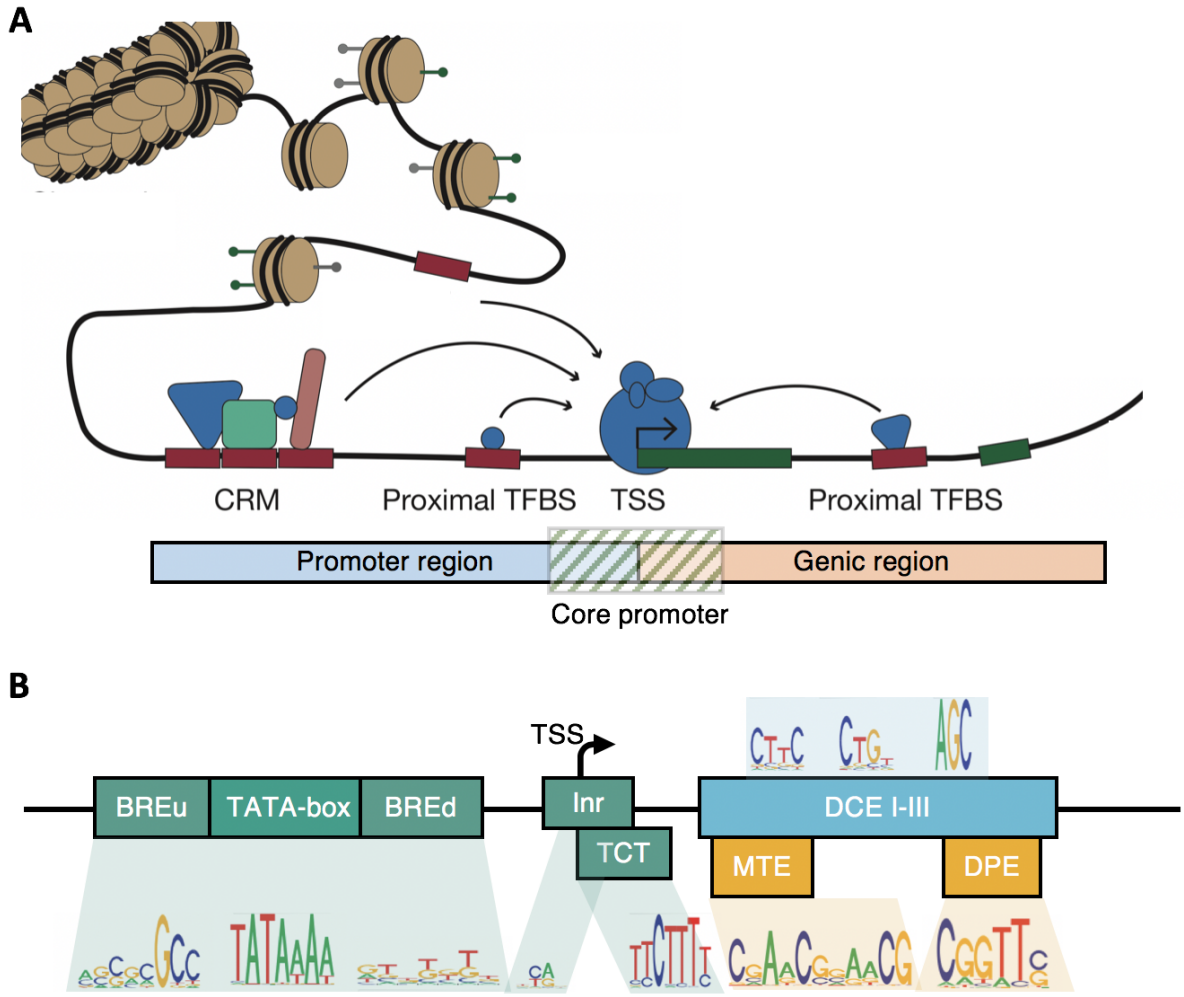


**Figure 1.1: Metazoan promoter and a schema of a core promoter.**
(**A**) The genomic region around promoters. Promoter region adjacent to TSS is called core promoter. Core promoter spans into geneic region. Promoters integrate signals from all regulatory elements to drive appropriate expression. Gene expression is coordinated by the binding of transcription factors to their specific sequence motif (TFBS) either in the proximity of core promoter, or away from it. TFBSs can be concentrated in a region, forming cis-regulatory modules (CRMs). Regulatory inputs on this figure are represented as arrows. (**B**) Schematic representation of the most frequent core promoter elements positioned relative to TSS. Colour of the motif boxes defines if the motif is vertebrate-specific (blue), *Drosophila*-specific (orange) or common to both (green). Sequence motif for each of the elements was obtained from the JASPAR database (Mathelier et al. 2016). Represented Initiator motif is vertebrate-specific. Adapted from (Lenhard, Sandelin, and Carninci 2012).

feature of tissue-specific genes is that their promoter sequence frequently contains a TATA-box motif (Schug et al. 2005). On the other hand, in mammals, ubiquitously expressed genes rarely contain this motif and instead have a high frequency of cytosine followed by guanine (CpG) dinucleotides. Long stretches of CpGs constitute CpG islands (CGIs) (Akalin et al. 2009; Carninci et al. 2006). Further research showed how metazoan RNA polymerase II promoters can be subclassified into at least three promoter classes based on their sequence features and the spread of transcription initiation (Lenhard, Sandelin, and Carninci 2012). Since then, two additional promoter classes have been identified: TCT promoters characteristic of genes coding for translational machinery proteins (Parry et al. 2010) and maternal promoters (Haberle et al. 2014). It is considered that there are more promoter classes not identified at the moment.

By looking at how precise the transcription start position is within the promoter, two promoter classes have been devised. The first group of promoters initiate transcription in a very narrow window with the majority of initiations coming from a single nucleotide and are therefore called sharp or peaked promoters. In contrast to sharp promoters, broad promoters initiate transcription across a wider area where initiation events are dispersed across multiple positions within a 50 - 100 bp range. These initiation patterns are associated with the function of the corresponding gene and are consistent across Metazoa (Haberle et al. 2014; Hoskins et al. 2011). Sharp promoters are consistently found to be enriched in tissue-specific genes, while broad promoters are associated with housekeeping and many developmental genes (Forrest et al. 2014; Rach et al. 2009).

### 1.3.1 Core promoter motifs

#### 1.3.1.1 TATA-box

The idea that promoters contain recurrent motifs was known even before the advent of next-generation sequencing. The first core promoter motif to be discovered is TATA-box (Lifton et al. 1978). The TATA-box motif can be found about 30 bp upstream from the dominant TSS in the sharp promoters (Ponjavic et al. 2006). This motif is specifically bound by TATA-binding protein (TBP), a component of TFIID protein complex that mediates recruitment and the positioning of RNA polymerase II (Louder et al. 2016). It is suggested that TATA-box is one of the main determinants in the choice of dominant TSS for sharp promoters. This core promoter motif is conserved from humans to yeast, but despite that, it is found in only a small subset of all promoters. For example, in flies, only about 10% of promoters contain a TATA-box (Ohler et al. 2002).

#### 1.3.1.2 Initiator

The TATA-box is not the only positionally restrained core promoter element. One additional motif with a well-defined position is the initiator (Inr) motif (Smale and Baltimore 1989). This motif directly overlaps TSS, but its width is not conserved among organisms. The *Drosophila* Inr is wider and better defined. It consists of several nucleotides surrounding the TSS which serve as a binding site for TFIID sub-components (Louder et al. 2016). The human Inr motif is significantly less information-rich, consisting of only a pyrimidine (C or T) upstream of a purine (A or G). As this dinucleotide overlaps TSS, the purine within Inr is the first transcribed nucleotide. Unlike TATA-box, Initiator is found in a higher proportion of promoters. Since the human version of the motif is less well defined, position weight matrix (PWM) searching for this motif in DNA results with many hits. Due to this, it was found that about 80% of promoters have this motif around their TSS.

#### 1.3.1.3 TCT initiator

Another core promoter element that overlaps the TSS and is found in genes that have specialised for a particular biological process is the TCT initiator element (Hariharan and Perry

1990). This motif is characterised by a cytosine at the first downstream position from the dominant TSS which is surrounded by a stretch of 4-13 pyrimidines. Because of the nucleotide composition, this motif is also called polypyrimidine initiator (Hariharan and Perry 1990). Further analysis in *Drosophila* found that this motif replaces canonical Initiator in almost all ribosomal protein genes. Later, a similar finding was confirmed in at least 48 human ribosomal protein genes (Roepcke et al. 2006). Along with ribosomal genes, some translation elongation and initiation factor proteins also contain the TCT element. The TCT element allows a different type of transcription initiation from the canonical initiation with a purine at the TSS. PolII has a strong preference for a purine at the TSS, and because TCT doesn't have purine in this position, canonical TFIID does not recognise these promoters (Parry et al. 2010). Therefore, it seems that promoters containing the TCT motif recruit a pre-initiation complex (PIC) that is specific for this motif and that is distinct from TFIID. Most likely, this specific PIC helps ribosomal protein genes to be highly expressed.

#### 1.3.1.4 DPE

A subset of *Drosophila* promoters that don't have a TATA-box have a conserved sequence motif in their TSS downstream region. This sequence is found about 35 nucleotides downstream of the TSS and is called downstream promoter element (Burke and Kadonaga 1996). DPE is shown to compensate for the lack of TATA-box by assisting in PolII positioning. Precise positioning of this motif in respect to Inr is thought to be essential for its function (Louder et al. 2016). Although DPE was discovered and characterised in flies, based on promoter similarity it has been suggested that it is present in human promoters too (Burke and Kadonaga 1996). However, it is still not clear which group of human promoters are enriched for this motif, if any (FitzGerald et al. 2006). Since DPE and TATA-box are rarely found together in *Drosophila* promoters, it is suggested that these elements are functional in distinct promoter classes (FitzGerald et al. 2006; Kutach and Kadonaga 2000).

#### 1.3.1.5 BRE elements

Additional core promoter elements identified are the two TFIIB recognition elements. The crystal structure of TFIIB complex with TBP showed that TFIIB binds DNA upstream of

TATA-box in the DNA major groove and in the DNA minor groove downstream of TATA-box (Nikolov et al. 1995). These elements are called BREu and BREd (TFIIB recognition element upstream and downstream) and are directly flanking the TATA-box (Lagrange et al. 1998). *In vitro* experiments showed that BREd elements enhance assembly of the complex between TFIIB and TBP (Nikolov Crystal structure). The BREd element positively affects transcription when the promoter only contains BREd. In contrast to that, when a promoter also contains BREu the presence of BREd has a negative effect on the rate of transcription (Deng and Roberts 2005). Considering BRE elements enhance binding of TBP complex, it would be expected that they are characteristic of TATA-box containing promoters, however, it has been shown that promoters without TATA-box are more likely to contain BRE sequence motif. Therefore, it is not possible to discriminate if BRE elements are genuine core promoter elements found around TATA-box or if they are just sequence artefact surrounding the GC-poor TATA-box region (Sandelin et al. 2007).

#### 1.3.1.6 MTE

In addition to the aforementioned motifs, other core promoter motifs have a well-defined position relative to the TSS. One of them is motif ten element (MTE) that is positioned 18 to 22 nucleotides downstream from the TSS (Lim et al. 2004). MTE requires the presence of canonical Inr motif for its function. It was found in flies, but it is thought to be conserved from *Drosophila* to humans (Lim et al. 2004). Interestingly, promoters with the MTE motif were able to compensate for the loss of transcription upon mutation of DPE or TATA-box (Lim et al. 2004). Another positionally defined core promoter element found in human promoters is the downstream core element (DCE). This motif can be found 10 to 40 nucleotides downstream from the TSS in $\beta$ globin genes. There, it acts as the TATA-box and the Initiator, even in the promoters without the TATA-box.

Computational analysis of a large collection of *Drosophila* promoters found new over-represented motifs that were not precisely positioned in regard to the TSS (Ohler et al. 2002). Computational analysis of a large collection of *Drosophila* promoters found new overrepresented motifs that were not precisely positioned in regard to the TSS (Ohler et al. 2002). Some of these motifs are the DNA replication-related element (DRE) and Ohler motifs 1, 6 and 7. They are usually found in the promoters of housekeeping genes (Figure 1.2B).

#### 1.3.1.7 GC-box

The core promoter elements mentioned above are considered as canonical promoter elements, however, there is a set of promoter motifs that are found further from the TSS and are considered as non-canonical promoter elements. One of these elements is the GC-box. The GC-box is found about 100 nucleotides upstream from the TSS and it consists of a C surrounded by a sequence of Gs. This motif is position-constrained but orientation independent and can be found in multiple copies along with the promoter (Lundin, Nehlin, and Ronne 1994). At this motif, Zinc finger proteins bind to DNA, one of them being the transcription factor Sp1. Binding of Sp1 enhances the assembly of PIC and, in turn, increases expression. Sp1 is widely expressed among eukaryotic cells so having a GC box present was sufficient to increase the expression of that gene.

Sequence motifs are often represented as PWMs (Stormo et al. 1982). These matrices are made out of four rows, representing each of the nucleotides, while the number of columns is dependent on the width of the motif being described. Values in the PWM are log-transformed values of likelihood ratios that a certain nucleotide is going to be found at that position. With PWMs it is possible to scan a sequence of interest and learn if this motif could be found and what is the certainty of finding it. PWMs for many transcription factor binding sites and sequence motifs exist and can be found in curated databases. One of the best-known ones is JASPAR (Sandelin, Alkema, et al. 2004; Mathelier et al. 2016; Khan et al. 2018). JASPAR allows querying motifs based on the organism, sequence features as well as downloading given PWMs. Curated collections of PWMs are also available with one of them being a collection of PWMs of motifs found in PolII core promoters.

Besides sequence motifs that define promoter elements, core promoters can differ in their nucleotide composition. Many mammalian promoters have a greater proportion of CpG dinucleotides than expected which form GCIs (Akalin et al. 2009; Carninci et al. 2006). Promoters that contain long CGIs are depleted of other sequence motifs. The exact mechanism of how CGIs aid transcription is not known. It has been shown that G+C rich sequences favour nucleosome formation in vitro which suggests that mammalian promoters prefer high concentration of nucleosomes (Zhang et al. 2009). CGIs within promoters are often demethylated if a gene is active. Demethylated CGIs are destabilising nucleosomes and in that way attract the PIC (Zhang and Zhu 2012). Conversely,

methylation is thought to promote the stabilisation of the nucleosome by causing the DNA to overwrap around the histone octamer (Lee and Lee 2012; Choy et al. 2010). Furthermore, Collings et al. showed that methylated CpGs form a different rotational orientation where minor groove faces histone octamer which allows DNA to overwrap into a nucleosome (Collings, Waddell, and Anderson 2013).

### 1.3.2 Differential usage of core promoter elements

Ability to change the transcription initiation process allows genes to switch between different modes of regulation and therefore the activity of a gene in different circumstances. A well-known example of a promoter changing core promoter elements used to activate transcription is observed during embryonic development (Haberle et al. 2014). Transcripts expressed in the oocyte are maternally inherited and are transcribed from a different promoter that is specifically used during development (Rach et al. 2009). Genes that show differential promoter usage also utilise different core promoter elements in these promoters. For example, genes that are expressed in the female germline in *Drosophila* are enriched in DNA replication-related element (DRE) and Ohler elements, while TATA-box and Inr are depleted in these promoters. In contrast, the developmental regulators that are embryonically transcribed are enriched for TATA-box and Inr (Down et al. 2007; FitzGerald et al. 2006; Rach et al. 2009).

Genome-wide analysis of the usage of mammalian promoters concluded that genes that use alternative promoters are usually carefully regulated developmental genes, while genes using only one promoter are more often involved in cellular processes across many tissues (Lenhard, Sandelin, and Carninci 2012; Ohler et al. 2010). Differential usage of core promoter elements in embryonic development seems to coincide with differential expression of PIC subunits. These two facts suggest that PIC composition at promoters varies based on the cellular context in which a gene is expressed (Müller and Tora 2009). Finally, sequence motifs found in core promoters influence transcription initiation patterns and are associated with cellular conditions during which they are used (Figure 1.2).

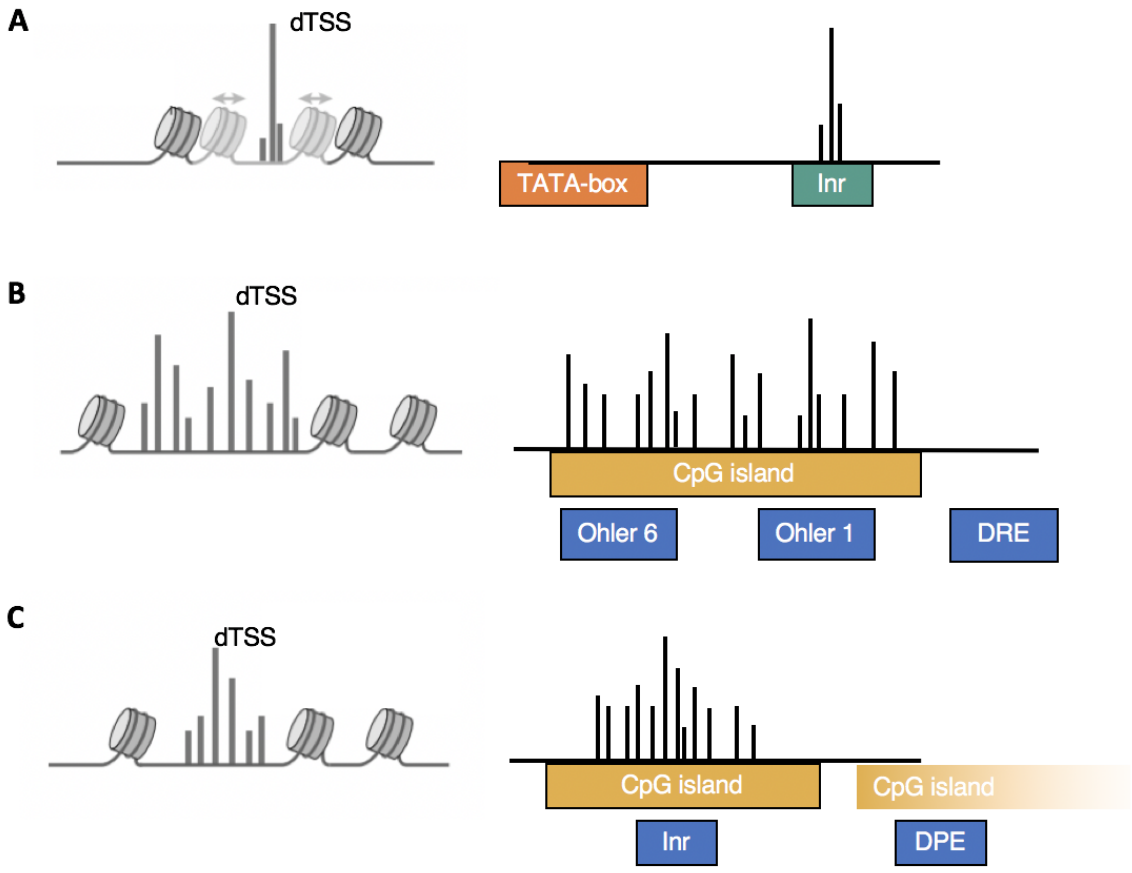


**Figure 1.2: Three main types of core promoters.**
For each of these promoter types, the left illustration shows the width of initiation and nucleosome positioning, and the right illustration represents characteristic core promoter elements. (**A**) Core promoters with precise - narrow initiation and imprecisely positioned nucleosomes. Core motifs like TATA-box and Inr help in the positioning of RNA polymerase. These core promoter motifs are precisely positioned relative to dTSS. These promoters are characteristic of tissue-specific genes. (**B**) Core promoter group with broad - dispersed transcription initiation and precisely positioned nucleosomes. In mammals, they overlap CpG islands, whereas, in *Drosophila*, DRE and Ohler elements are enriched in these core promoters. Ohler elements 1 and 6 are overrepresented sequence motifs found in *Drosophila* core promoter elements that are unconfirmed as functional promoter elements (Ohler et al. 2002). These promoters are characteristic of housekeeping genes. (**C**) Core promoters with broad - dispersed initiation and precisely positioned nucleosomes. They are overlapping long individual CpG islands or multiple short ones. In flies, these promoters often contain DPE and Inr elements. These promoters are characteristic of developmentally important genes. Adapted from (Haberle and Stark 2018).

### 1.3.3 Three main types of core promoters

When first comparing a limited set of promoters, it was suggested that there are two classes of promoters, ones containing TATA-box and the rest without it (Smale 1997). Later, with the advent of NGS methods, this classification was extended. Genome-wide analyses suggested that promoters without a TATA-box often have high GC content and oftentimes overlap CGIs (Juven-Gershon et al. 2008). Based on the above-mentioned promoter features like sequence motifs, nucleotide composition and initiation pattern along with chromatin configuration, three main types of Metazoan core promoters have been suggested (Lenhard, Sandelin, and Carninci 2012) (reviewed in (Haberle and Stark 2018)).

- The first class of promoters (initially called TATA-containing) are promoters where most initiation starts from a single position ('sharp' promoters) (Figure 1.2A). These promoters have low GC content in their core promoter region and show imprecisely positioned nucleosomes. Many promoters from this class have a TATA-box at a fixed distance from the TSS and their key regulatory elements are also in close proximity to the TSS (Lenhard, Sandelin, and Carninci 2012). TATA-box promoters are usually active in differentiated cells and are found in genes with tissue-specific function. In addition to sequence motifs, these promoters will have histone marks characteristic of active transcription (histone H3 Lys 4 trimethylation (H3K4me3) and H3 Lys 27 acetylation (H3K27ac)).

- The second class of promoters are found in genes that are expressed at constant levels in all adult cells, also known as 'housekeeping' genes (Figure 1.2B). These genes show broad transcription initiation pattern with many dispersed nucleotides at which transcription can initiate. The TSS is found in a nucleosome depleted region that is surrounded by precisely positioned nucleosomes (Rach et al. 2011). In *Drosophila*, these promoters have a specific set of core promoter elements including DRE and Ohler elements (Rach et al. 2009), while in mammals these promoters will overlap CGIs (Carninci et al. 2006).

- Finally, the third group of promoters was found in key developmental genes involved in pattern formation (Figure 1.2C). In mammals, these promoters show features similar to housekeeping

promoters. Unlike housekeeping promoters, developmental promoters are associated with wide individual CGIs or a larger number of short CGIs that often extend into the gene body (Lenhard, Sandelin, and Carninci 2012). In *Drosophila*, these promoters were found to have an Inr element frequently accompanied by DPE (Kutach and Kadonaga 2000). When looking at the epigenetic features of these promoters, they are bivalently marked with activating mark H3K4me3 and repressive mark H3K27me3 in ES cells which allows them to be activated in specific cells (Bernstein et al. 2006).

## 1.4 Application of single-cell sequencing methods in embryonic development

Embryonic development is the process where a single cell forms a new organism with multiple different tissues and cell types. To characterise the full repertoire of cells and to understand processes like lineage commitment and differentiation, high throughput technologies were needed to investigate transcriptomes of individual cells. RNA-seq and ChIP-seq technologies have been used for more than 15 years to study transcriptomes and epigenomes in development, however, because these methods require high amounts of starting material, they were applied to populations of cells or even whole embryos. By studying expression features of a large number of cells, the information obtained from such an experiment is averaged out to all assayed cells. Due to this, rare cell types remain poorly understood. Recent applications of single-cell sequencing methods are starting to alleviate this problem (Tang et al. 2009).

Single-cell RNA sequencing (scRNA-seq) technology is transforming our understanding of biological processes. It has already impacted many areas of biology with developmental biology being among the most influenced areas. The ability to analyse transcriptomes of individual cells is enabling us to understand the composition of tissues and the interactions between cells within individual tissues, but also to better understand inner processes within the cell. Single-cell sequencing has been an invaluable method for identifying genes that directly influence cell fate decisions towards specific lineages (Teles et al. 2013).

In short, the procedure for generating scRNA-seq transcriptomes is as follows. The tissue of

interest is dissociated, and individual cells are isolated. Cells are then lysed, and cDNA is produced so that at the ends of sequences, a unique barcode for that cell is added. Next comes the library preparation and sequencing. Finally, the reads containing the unique barcodes are aligned to the reference genome and quantified for expression analysis. By quantifying the number of reads with each barcode that are mapped, we are able to retrieve expression information for individual cells.

Until recently, studies using scRNA-seq were analysing fewer than a thousand single cells (Kolodziejczyk et al. 2015). Although this was a great achievement, these projects were still limited in understanding cell heterogeneity or lineage commitment decisions. With the advancements of scRNA-seq technology cells from whole tissues and organisms are being analysed. Currently, one of the largest collections of cells assayed in a scRNA-seq experiment consists of over 1.3 million cells isolated from the brains of embryonic mice (Genomics 2017). This single experiment was able to detect major neuronal and non-neuronal cells, including rare neurons, without additional enrichment protocols. Detailed transcriptome analysis of prenatal mice brain motivated creation of the human cell atlas by Human Cell Atlas Consortium that aims to investigate over 10 billion transcriptomes of individual cells from the human body (Regev et al. 2017). More recently, Cao et al. assayed transcriptome of about 2 million cells extracted from 61 mice embryos during organogenesis. This resulted in the creation of the 'mouse organogenesis cell atlas' which provides a global, in-depth view of organogenesis (Cao et al. 2019).

A particularly useful application of scRNA-seq for development stems from the ability to analyse cells at various developmental stages in a single experiment. This way it is possible to reconstruct embryonic signalling pathways and construct trajectories of cell differentiation. An early study of mouse development assayed transcriptome differences between zygote and 2-cell stage embryos. They found that in the initial cell division, transcriptomes are unevenly distributed to daughter cells. These initial differences in transcript quantities get even more pronounced later when the zygotic genome activates (Piras, Tomita, and Selvarajoo 2014; Shi et al. 2015). At the four-cell stage differences in transcriptomes cause cells to commit to become a part of pluripotent inner cell mass or extra-embryonic trophoectoderm (Shi et al. 2015). Last year, large scale scRNA-seq studies analysed organogenesis in mouse and zebrafish. More than 20,000 cells from E8.25 embryos

were analysed which enabled identifying 20 major cell types (Ibarra-Soria et al. 2018). The first day of zebrafish embryogenesis was studied across seven timepoints by analysing more than 90,000 cells (Wagner et al. 2018). Here authors recapitulated transition from pluripotent blastomeres to a large number of cell types. Another important finding from this paper was that cells, during differentiation, do not invariantly follow lineage commitment pathway, as some cells that were initially identified to differentiate towards the neural crest cell type, later became cells of pharyngeal arches. With the development of bigger and more detailed single-cell atlases of development, we will be able to get a deeper understanding of developmental processes at a finer scale.

## 1.5 Aims of this thesis

Achieving coordinated gene expression in appropriate spatio-temporal conditions is a crucial mechanism for development and life of multicellular organisms. Genes that have a specific spatio-temporal gene expression profile have to be tightly regulated to ensure this specificity. Examples of highly regulated genes can be found in genes that regulate cell-cycle stages, genes that are activated in response to external stimuli or genes that regulate development. This thesis focuses on characterising spatio-temporal gene expression dynamics during embryonic development. I investigated if indeed it is the case that genes with the most complex expression repertoire are key developmental genes. I went further to characterise their co-expression partners and their promoter structure. In the following bullet-points, I will summarise the aims of the individual chapters:

- In Chapter 2 I developed a method that annotates promoters using core promoter elements and performs enrichment analysis of these features for a group of promoters. For this analysis, the correct positioning of the dominant TSS is crucial since the distances of core promoter elements are dependent on it. I have used publicly available CAGE-seq datasets for both human and zebrafish to create a reference of promoter locations and core promoter elements that they contain. This method can then be used to identify novel promoter classes, based on the enrichment of promoter features, or to test if a gene group has specific promoter features that could suggest the way these genes are regulated. I have used this method in the subsequent chapters to analyse promoter features of genes with complex spatio-temporal expression patterns.

- In Chapter 3 I analysed ZFIN's mRNA *in situ* hybridisation data across zebrafish development with the aim to explore the complexity of gene expression patterns. Using this information I was able to define how tissue- and stage-specific is the expression of each gene. Since mRNA *in situ* hybridisation data only defines a spatial component of expression, but not quantities of expression, I utilised RNA-seq expression data generated from the whole embryo which does not have a spatial component but can be used to estimate how uniform the expression level is throughout development. By combining these two methods of interrogating gene expression, I have defined a new coefficient of gene expression complexity. This coefficient was later used

to investigate if developmental genes are indeed most complexly expressed genes in zebrafish and if housekeeping genes are the least complex.

- Chapter 4 is focused on gene co-expression expression patterns that can be defined from a scRNA-seq experiment. Here, I have analysed data from slightly fewer than 40,000 cells spanning nine hours of zebrafish development. Key developmental genes are expressed in multiple tissues across the embryo, controlling the expression of downstream genes and committing cells to their lineages. I was interested in whether these genes have a similar pattern of activity across different tissues, or if their expression is context-specific. In addition, I analysed the pair-wise similarity of co-expression across different cell groups. To do this, I clustered genes based on their expression profile, and within these clusters, I have identified modules of co-expressed genes. The knowledge that two genes are co-expressed across different tissues could be a strong indication that they share a segment of their regulatory programme, which can assist in understanding mechanisms and interaction of gene regulation in development.

- In Chapter 5 I discuss the results of this thesis and propose ideas for future work.

# 2 Promoter Ontology

## 2.1 Introduction

Precise gene expression has a central role in embryonic development and metabolism. An aberrant expression has been shown to cause disorders and disease. Gene expression is regulated by a plethora of transcription factors binding at promoters, or by distal regulatory elements that are contacting promoters. One of the key events of gene expression is the recruitment of RNA polymerase II by general transcription factors and their assembly into a preinitiation complex. Formation of preinitiation complex occurs at core promoters. Core promoters are DNA regions found approximately 50 bp around the transcription initiation site. They are sufficient to initiate transcription, but in most cases, they provide a basal level of expression. Further refinements to gene expression profile are established by contacts with distal regulatory elements that specifically increase expression, or by chromatin repression. All these regulatory signals are integrated at core promoters.

Core promoters are highly diverse in terms of general transcription factors that are required to drive expression, but also their sequence composition varies. The advances in Next-generation sequencing technologies have enabled mapping of endogenous transcription initiation at single-nucleotide resolution. This led to the discovery of at least three main classes of core promoters that are found in genes with specific expression patterns (Carninci et al. 2005). These promoter classes were devised based on the transcription initiation pattern, sequence composition and corresponding motifs. The first group of genes initiates transcription from a narrow region and contains key regulatory elements close to TSS. These genes are mainly active in a tissue specific manner. The second group of genes is expressed in housekeeping manner and their initiation regions are wide with individual CpG islands overlapping core promoters. Finally, the third group of genes is alike housekeeping promoters, with the difference in histone mark composition in the vicinity of core

promoters and the fact that they have larger or multiple CpG islands that can extend into the gene body. These genes have been identified to have an important role in development.

In this chapter, I explore the specificity of core promoters for the gene function. I have analysed core promoter characteristics of various gene groups clustered based on common function or gene product with the aim to find if three main promoter classes can be further subdivided. In addition, I have analysed if orthologous genes retain their promoter features in different organisms. To do this, I developed a method that that annotates promoters using core promoter features and performs enrichment analysis of these features for a group of promoters compared to all active genes in that sample. This method enables identification of significantly enriched promoter features for the analysed group of genes and comparison across different gene groups. I present a use case for this method by using it along with gene ontology to characterise differentially expressed genes identified by differences in transcriptomes in pancreatic cell types.

## 2.2 Methods

### 2.2.1 Defining promoters

To identify promoters and a position from which the majority of transcription starts I have used FANTOM5 CAGE data for human samples (Forrest et al. 2014). FANTOM5 is the fifth phase of a collaborative project aiming to map the majority of human promoters across different primary cells, cell lines and tissues. As a part of the project, they have isolated RNA and created CAGE libraries originating from every major human organ. The human dataset comprises more than 200 primary cell types and over 200 cancer cell lines. This dataset provides a great potential to study core promoters of different tissues. To present the functionality of the developed method, I have focused on a small subset of samples from diverse, non-cancerous, cell types. I used the following samples: GM12878 cell line, ESCs, liver and muscle tissue sample.

Zebrafish samples were obtained from (Nepal et al. 2013), a study profiling promoters across 12 stages of zebrafish development, from unfertilised oocyte to Prim25 stage.

For both organisms, CAGE transcription start sites (CTSS) were obtained in a BED format where each transcription initiation event is represented by one entry. Single CTSSs were clustered with "distclu" clustering method into tag clusters (TCs), bigger genomic regions from which transcription initiates. A frequency of CTSSs along TCs gives information about promoter shape and structure. In order to get a robust set of TCs, we excluded all TCs supported by less than 1 tpm. In addition, TCs consisting of a single nt position that had less than 5 tpm were omitted too.

Gene annotation for hg19 and danRer10 genomes were obtained from BioMart (Smedley et al. 2015). Gene coordinates for protein coding genes were overlapped with CAGE tag clusters, and dominant CTSS was used as a centre of a promoter region. In cases where multiple TCs overlapped a single BioMart defined promoter, only the most highly expressed cluster was used. Promoters were defined as a one kilobase region centred on the dominant CTSS.

### 2.2.2 Scanning for PolII promoter motifs

The collection of position frequency matrices describing DNA sequence patterns found in RNA Polymerase II (Pol II) promoters were obtained from JASPAR2016 (Mathelier et al. 2016; Tan 2015). It consists of 13 known DNA motives experimentally described as elements of RNA Polymerase II core promoters. Two of these motives are Drosophila specific (DPE and MTE) and were omitted from downstream analyses. Transcription factor binding sites were predicted by PWM matrix scan of all promoters with 80% identity threshold. To perform sequence scan I used the TFBS tools R package (Tan and Lenhard 2016).

### 2.2.3 Classification of PWM-based features

Scanning sequences with PWM matrices returns numerous hits that pass the 80% threshold. For each hit, location and score are reported. Sequence scanning with PWMs yields many false positive hits. To separate false positive PWM scan hits, I devised two coefficients that, for each sequence hit, provide information about hit's location accuracy and similarity to true PWM. A combination of these two coefficients informs about the overall accuracy of a sequence scan hit.

To devise the coefficient that describes location accuracy of a hit, I have estimated a

probability of finding a hit for each position in the promoter. To do this, I calculated the positional frequency of all reported hits from all promoters in the sample. Finally, the frequency of hits on each position was normalised by the frequency at the nucleotide with the most hits. Coefficient obtained by normalising position of a hit is going to penalise positions with a low number of hits, that come from noisy regions. Additionally, the PWM identity scores from sequence scans were normalised after assigning all hits above the $95^{\text{th}}$ percentile to the $95^{\text{th}}$ percentile. This way the hit with the highest identity score had a value of 1 and the lowest-scoring hit had a value of 0. New coefficients of location probability and sequence identity allow us to examine the likelihood of a predicted hit to be a true positive. Coefficients were binned, and bins were used to plot frequency heatmaps and to test the correlation between location and a score of a hit.

The normalised location of a TF scan hit and the identity score for PWM match were combined into a single coefficient. Combined coefficient, for each scan hit, defines how precisely positioned was the hit in regards to all hits from all active promoters, and how similar to the original PWM is this region. Based on the distribution of all reported hits in the population, I defined a threshold that a hit needed to satisfy to be considered a true hit. Distribution of combined coefficients for all hits is a right-skewed distribution, with most hits having very low scores because they either were not correctly positioned, or hit sequence was not highly similar to true PWM. Due to the majority of hits having a low score, a threshold was defined as the inflection point of the empirical cumulative distribution function of cumulative scores for all hits. To calculate the inflection point, I used R package `inflection`. In cases when inflection point could not be defined or was higher than 0.9, I used $80^{\text{th}}$ percentile of all scores as a threshold.

### 2.2.4 Promoter annotation of range-based features - CpG islands and bidirectionality

Coordinates of CpG islands for hg19 were extracted from UCSC Table Browser and overlapped with predefined promoter sequences of human samples. Zebrafish CpG islands were predicted using a sliding window approach with all regions satisfying specified observed/expected ratio (data created by Christopher Previti, a previous group member). Bidirectional promoters were

determined by calculating the distance between dTSSs for all pairs of genes in human and zebrafish genome. For pairs of genes which were in head to head orientation and whose distance between dominant TSSs would be less than 1000 nt, both genes would be assigned as bidirectional.

### 2.2.5 CAGE features

The following features extracted from CAGE-seq data were used to annotate promoters of all active genes: interquantile (IQ) width, expression value of the whole CAGE tag cluster and expression of the dominant TSS. Promoter width is defined as an interquantile width of CAGE signal at each locus and is calculated by using the CAGEr package (Haberle et al. 2015). A total sum of CAGE signal at a locus was considered as the expression value of the whole cluster and expression level of dominant TSS was defined as the amount of expression at the most highly expressed nucleotide in a CAGE cluster.

### 2.2.6 Annotation enrichment analysis

After annotating all promoters with PWM, CAGE and range-based features, I have created a promoter feature matrix that for each gene describes core promoter features that a gene contains. Presence of PWM and range-based core promoter features are binary (for example, a promoter either overlaps a feature or it does not). To test their annotation enrichment of features in a sample compared to the whole population, a hypergeometric test was performed. Hypergeometric cumulative distribution function used to calculate significance can be found in Equation 1.

$$P_{enrich}(x, N, n, m) = 1 - \sum_{i=0}^{x-1} \frac{\begin{pmatrix} m \\ i \end{pmatrix} \begin{pmatrix} N-m \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{1}$$

At the same time depletion was calculated by calculating lower tail of cumulative hypergeometric function Equation 2.

$$P_{depl}(x, N, n, m) = \sum_{i=0}^{x} \frac{\begin{pmatrix} m \\ i \end{pmatrix} \begin{pmatrix} N-m \\ n-i \end{pmatrix}}{\begin{pmatrix} N \\ n \end{pmatrix}} \tag{2}$$

In both functions, $x$ denotes the number of features present in a sample of $n$ genes, drawn from a population of $N$ active genes containing $m$ promoter features.

Unless an explicit background set is provided, all genes active in provided CAGE sample are used as background for enrichment analysis.

Results of enrichment analysis are provided as a table with all significantly enriched features, and graphically. Three graphs are provided and are by default saved into a pdf file. Those graphs present results of the over-enrichment analysis, analysis of significantly depleted promoter features and distribution of CAGE features for the sample and the background. Figures generated for annotation enrichment test contain these parameters: ratio and number of genes in a sample containing a given feature as well as the significance of enrichment test. R packages ggplot2 (Wickham 2006) and the ggthemes were used to generate graphs.

### 2.2.7 RNA-seq analysis

This analysis was performed as an example of the usage of the Promoter Ontology (PO). In the original study, authors used Gene Ontology to annotate gene functions while by using Promoter Ontology would enable additional insights into groups of genes they obtained. To obtain identical results, I have analysed data just like the original publication did. Data used for RNA-seq analysis presented in this chapter was published in (Tarifeño-Saldivia et al. 2017) and was obtained from the European Nucleotide Archive (ENA) under the accession number PRJEB10140. Data contains pair-end RNA sequences from the pancreatic acinar, alpha, beta and delta cells from adult zebrafish. All datasets were quality checked by FASTQ. The trimmomatic was used to remove sequencing adaptors and trim low quality reads (Bolger, Lohse, and Usadel 2014). Reads were then mapped to Zv9 genome build using Tophat v.2.0.9 (Trapnell et al. 2012) with slightly modified default

parameters (–segment-length 18, –min-intron-length 30). Gene expression tables were generated from the mapped reads by using HTSeq count algorithm (Anders, Pyl, and Huber 2015). To find a characteristic gene expression profile of each of these cells I used R package DESeq2. DESeq2 normalises all expression values by using variance stabilisation transformation. To create endocrine specific expression sample, just like the authors of the study - (Tarifeño-Saldivia et al. 2017), I merged RNA-seq data from alpha, beta and delta cells. To identify differentially expressed genes in pancreatic cell types, I used DESeq2. DESeq2 uses Wald test to test significantly different genes after the shrinkage estimation for each of the samples. Final p-values are corrected for multiple testing by using Benjamini-Hochberg method (Benjamini and Hochberg 1995). Genes whose corrected p-value was less than 0.05 and whose fold change was greater than 4 were considered differentially expressed and characteristic for that cell type.

The characteristic function of all differentially expressed genes was determined by doing a Gene Ontology (GO) analysis using the clusterProfiler R package (Yu et al. 2012). For GO analysis, the background used were all active genes in the sample and p-value threshold of 0.01. To determine the characteristic promoter architecture of genes characteristic for each cell type, and to compare if they are regulated differently, I used Promoter Ontology with zebrafish Prim-25 (48 hpf) sample. Again, background used were all genes active in the CAGE sample. Since the stage of CAGE sample did not match original datasets, I did not analyse CAGE based features in downstream promoter analysis.

### 2.2.8 Gene ontology analysis

The complete GO controlled vocabulary and associations between genes and GO terms were obtained from GO consortium by using the GO.db R package (Carlson et al. 2015). This package enables the user to extract all child and offspring GO terms, where offspring terms would be children terms and the children terms of all subsequent children terms. For the downstream analysis, I used GO terms that contain at least 5 genes active in the human muscle tissue sample. For each of the remaining GO terms, I used Promoter Ontology to test their core promoter composition of genes annotated to these GO terms. In these analyses, I used all genes active in muscle as background set of genes. For each of the three ontologies, I calculated the proportion of GO terms that show

significant enrichment (p-value $< 0.05$) for core promoter features. To test if related GO terms have more similar promoter structures, I compared the distributions of core promoter feature occurrences between child gene groups and randomly sampled GO gene groups. To be able to compare the amount of significantly different promoter features, I created a random set of GO gene group pairs. For each GO parental term, I sampled 100 unrelated GO terms from the same ontology, whose size was in a range of sizes of the real child GO gene groups. Distributions of promoter feature proportions were compared between true child GO gene groups and randomly sampled ones and tested for significance using Kolmogorov-Smirnov test.

Finally, to assess if all child terms inherit similar core promoter features or some child terms are significantly different, I analysed if occurrences of promoter features in parent and child are coming from the same distribution. This was done by using a hypergeometric test where parental occurrences of promoter features were used as a background. To obtain genes associated with the parent and child term in this comparison, I have extracted all offspring GO gene groups and combined them together in one group. This analysis was done for all child-parent pairs. All significantly different proportions were reported. For parent-child pairs in which there was a significant difference in their promoter structure, I also compared the difference in their tissue specificity distribution and GO semantic similarity to compare if these factors could cause the change. The coefficient of tissue specificity was calculated using the data from Illumina's Human BodyMap 2.0 that contains RNA-seq expression values of 16 human tissue samples (Hishiki et al. 2000). For each gene, $\tau$ index was calculated using the Equation 3:

$$\tau = \frac{\sum_{i=1}^{N} (1 - x_i)}{N - 1} \tag{3}$$

where $N$ represents the number of tissues analysed (in my case 16) and $x_{\mathrm{i}}$ is the expression value from a single tissue normalized by the maximal expression of that gene.

Semantic similarity between all GO terms was calculated using Wang method. The Wang method of calculating semantic similarity is a graph-based method that uses the topology of GO graph. This analysis was done using the R package GOSemSim (Yu et al. 2010).

## 2.3 Results

### 2.3.1 A pipeline to annotate promoter features

To better understand regulatory features shared by a set of genes, I developed a core promoter enrichment analysis method, Promoter Ontology. Promoter Ontology has a dual mode of action:

- Firstly, it provides functionality to obtain promoters and annotate them with core promoter features provided by a user. Annotation of promoters is a step that is done only once per sample. All annotations are saved in a feature table. Feature table is a dataset in which each row presents a gene and columns are promoter features.
- Once Feature table is created, Promoter Ontology performs overrepresentation analysis of promoter features for a query set of genes. Promoter features used for annotations of promoters by default are: core promoter motives are represented as PWMs, obtained from JASPAR PolII collection; CpG islands, bidirectionality and CAGE obtained features: promoter width, the total expression of the promoter region, expression contribution coming from the dTSS.

Users are able to add additional features for annotation either by providing a PWM matrix of a motif, or a genomic range.

A precise position of some TF motives relative to the TSS is very important for core promoters to properly affect gene expression. RNA-seq based methods of defining promoter start site are not accurate enough to help distinguishing functional hits at the correct spacing. To address this problem, we have used CAGE-seq defined dominant TSS position (Kodzius et al. 2006). This way, after scanning promoter sequences for TF motifs, we are able to finely filter for those PWM hits that are within the functional range from TSS. Annotation is done by scanning promoter sequences with core promoter PWMs. The score of each PWM hit that was above 80% similarity was normalised to the maximal score in the population of all active promoters so that all scores are in the range from 0 to 1. In addition, position location was normalised by counting the location of all hits on each nucleotide in the promoter region. The position that had the highest count of hits

was used to normalise all other positions. In combination, these coefficients provide a more precise way of selecting true PWM hits, as it provides a way to select hits that are correctly positioned and of a high score. Once all motif-based promoter features are annotated, range-based features like CpG islands and bidirectionality are annotated by overlapping them with previously defined set of promoters.

Promoter Ontology algorithm is described in the Methods section and a graphical workflow is presented in Figure 2.1.

The output of overrepresentation test consists of a table and graphical representation of all the details of the test. The table informs, for each promoter feature, the significance level of enrichment, number and ratio of promoters having the tested feature.

To make Promoter Ontology methodology more accessible to the community, I have made an R package with a function for each of the steps of the pipeline. PromoterOntology package can be accessed at ('PromoterOntology Github Repository', 2019).

I have used Promoter Ontology to annotate promoters from several human tissues and a cell line, and promoters active during zebrafish development. In addition, to present the overrepresentation test, we have used it to explore differences in promoter composition of four different types of human-zebrafish orthologs.

#### 2.3.1.1 Human FANTOM5 dataset

**Table 2.1:** Frequency of promoter features in human samples.

| | ESC | GM12878 | Liver | Muscle |
|---|---|---|---|---|
| **INR** | 42.46 | 42.29 | 43.73 | 43.13 |
| **GC-box** | 63.92 | 63.57 | 63.3 | 64.18 |
| **CCAAT-box** | 14.79 | 13.72 | 11.75 | 11.88 |
| **BREu** | 48.64 | 46.59 | 45.18 | 44.64 |
| **BREd** | 7.1 | 5.51 | 4.85 | 7.07 |
| **DCE_S_I** | 26.71 | 23.44 | 25.42 | 23.86 |

| | ESC | GM12878 | Liver | Muscle |
|---|---|---|---|---|
| **DCE_S_II** | 39.68 | 38.67 | 38.91 | 38.5 |
| **DCE_S_III** | 41.64 | 41.22 | 41.21 | 40.66 |
| **XCPE1** | 79.2 | 77.05 | 77.47 | 77.77 |
| **TATA-box** | 7.65 | 7.1 | 9.86 | 9.96 |
| **MED-1** | 88.81 | 86.05 | 89.27 | 88.48 |
| **CpG island** | 89.83 | 87.85 | 84.2 | 84.53 |
| **bidirectionality** | 8.65 | 8.57 | 8.49 | 8.51 |
| **sharp** | 34.87 | 29.85 | 38.36 | 38.48 |

Human datasets were obtained from FANTOM5 consortium. To characterise the core promoter structure of genes active in embryonic stem cells (ESC) and GM12878 cell line, liver and muscle tissue samples, I annotated these promoters with Promoter Ontology method. In addition, I explored the variation across proportions obtained in these samples. A large discrepancy could suggest that core promoter annotation is not consistent. Summary statistics of core promoter annotations in each of these samples can be found in Table 2.1. Importantly, the proportions of core promoter features identified at each of the samples are consistent. Some variation is observed in the case of TATA-box, where ESC and GM12878 cell line samples have about 7.5%, while in tissue samples there are about 10% of TATA-box containing promoters. Since Promoter Ontology annotates only promoters of active genes, it is expected that tissue samples have more tissue-specific genes active, therefore more TATA-box containing genes. This is consistent with the observation that tissue samples have about 38% of sharp promoters whereas ESC and GM12878 contain 35% and 30% respectively. Many human promoters overlap CpG islands, namely, tissue samples contain about 84% of CpG island containing promoters and cell lines have an even higher percentage. All samples contain about 43% promoters with INR motif.

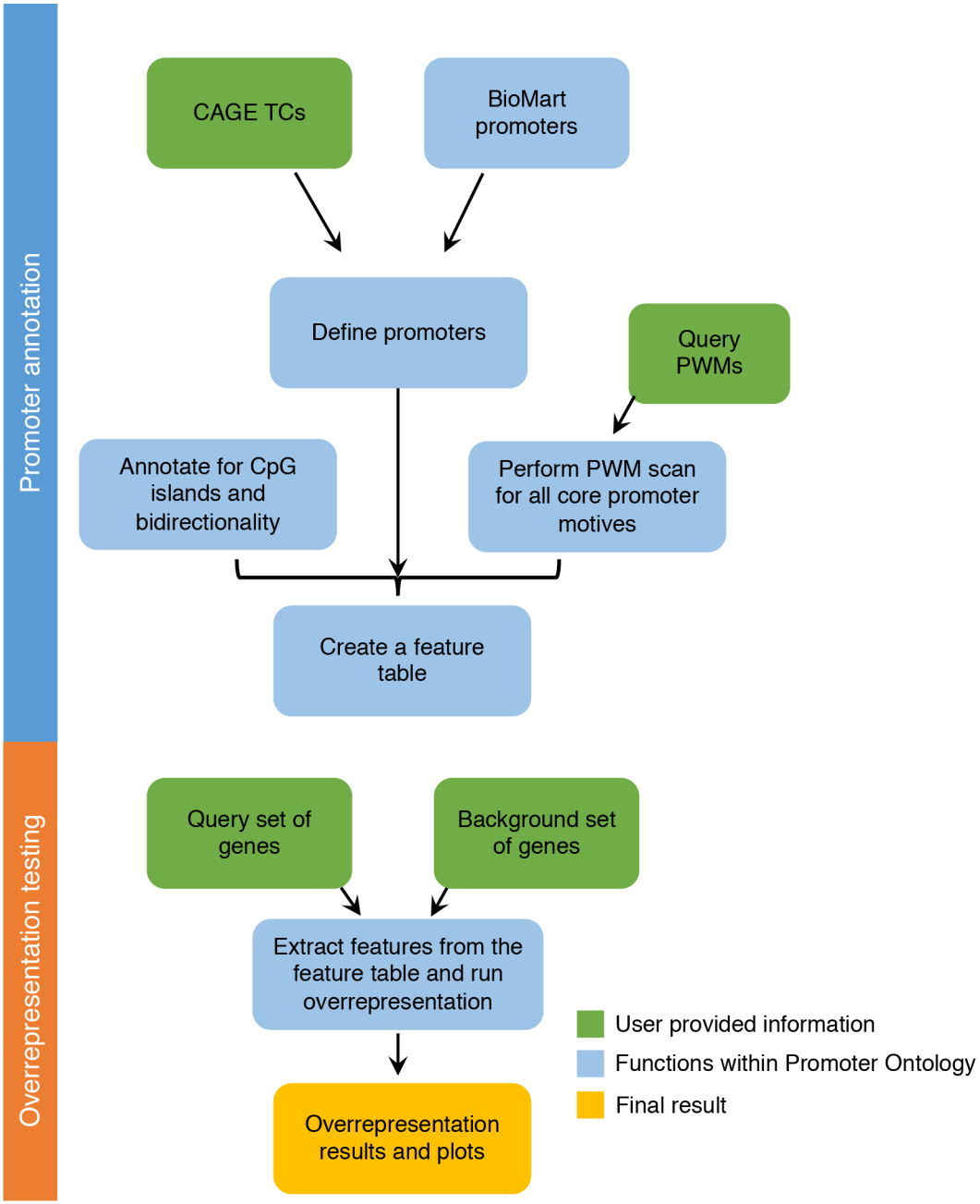


**Figure 2.1: A schema of the Promoter Ontology pipeline.**
The pipeline consists of two parts denoted by a bar on the left. The first functionality of the pipeline is to annotate promoters with specified features. To do that, a user is required to provide identified CAGE tag clusters. With the help of BioMart gene annotations, promoter regions, which are centred on CAGE dominant TSS, are defined. Promoter regions are then annotated with features provided by a user, most commonly those will be a list of PWMs of interest, coordinates of CpG islands and gene names of bidirectional genes. All annotations are concatenated in a feature table. Finally, with promoter annotations in place, overrepresentation analysis can be performed. It is required to provide a query set of gene names and to define a background of genes for the analysis (in a situation where a user fails to provide background, all active genes from a provided sample will be used as a background). With the information from the feature table, a hypergeometric test is performed and results are reported in a table and a graph format.

#### 2.3.1.2 Zebrafish dataset

**Table 2.2:** Frequency of promoter features in zebrafish samples.

| | 64 cell | 30% epiboly | Prim-5 | Prim-25 |
|---|---|---|---|---|
| **INR** | 68.51 | 72.86 | 72.2 | 72.03 |
| **GC-box** | 22.18 | 26.35 | 27.57 | 28.11 |
| **CCAAT-box** | 22 | 19.6 | 16.53 | 17.8 |
| **BREu** | 9.1 | 8.27 | 15.05 | 15.35 |
| **BREd** | 5.49 | 12.71 | 9.13 | 7.84 |
| **DCE_S_I** | 14.34 | 17.55 | 15.55 | 15.33 |
| **DCE_S_II** | 20.18 | 18.1 | 22.87 | 23.15 |
| **DCE_S_III** | 32.88 | 34.27 | 28.97 | 28.27 |
| **XCPE1** | 0.26 | 0.08 | 13.77 | 13.75 |
| **TATA-box** | 8.87 | 8.39 | 11.97 | 11.78 |
| **MED-1** | 18.1 | 64.38 | 25.24 | 26.3 |
| **CpG island** | 53.14 | 52.77 | 50.43 | 50.32 |
| **bidirectionality** | 1.2 | 1.18 | 1.12 | 1.14 |
| **sharp** | 44.67 | 39.55 | 39.21 | 39.66 |

Zebrafish CAGE datasets were obtained from (Nepal et al. 2013), a study that analysed transcription initiation landscape during 12 stages of zebrafish development. Using this dataset, I have annotated promoters of active genes for all available stages of development. General statistics of annotated core promoters for four representative stages can be found in Table 2.2.

Obtained proportions suggest that later developmental stages (after MZT) contain a greater proportion of TATA-box containing genes (about 12%) than early, pre-MZT, stages (about 8%). This result can be explained by the fact that later in the development there are more tissue-specific genes active which increase the proportion of TATA-box containing genes. Additionally, genes active in later developmental stages overlap fewer CpG islands and they overlap fewer CCAAT-box motifs.

There is a striking difference in the frequency of XCPE1 motif. Early developmental stages contain less than 1% of this motif, while in later stages it was identified in almost 14% promoters. To further characterise if this difference in proportions is biological, or it is the artefact of motif scoring, I checked raw values of PWM scan. In the early stages of development, a group of 7 active genes obtained a very high identity score with PWM which caused all other motif hits, when normalising, to score much lower.} MED-1 motif is very variable in this dataset. It was identified in 18% of promoters in the 64-cell stage, while in 30% epiboly, its frequency raises to 64%. Although MED-1 is required for proper development and MED-1 knock-out mice, for example, die at embryonic day 11.5 (Ito et al. 2000), this cannot explain such difference in frequencies. Since Promoter Ontology evaluates all hits in the population of promoters and based on the group of the highest-scoring hits defines thresholds for calling a hit true-positive. In 30% epiboly, in this case, none of the promoters had very high scores which caused the threshold to be lower than in the other samples.

In comparison to human promoter feature composition, zebrafish promoters have significantly fewer bidirectional promoters. This observation can be explained by the fact that teleosts have undergone another round of genome duplication so there was less pressure to retain all bidirectional genes. Zebrafish genome seems to have fewer CpG island containing promoters (from about 85% in human to 52% in zebrafish), but much more INR containing promoters (from 43% in human to 70% in zebrafish). The reason behind the difference in the content of CpG island is that due to the differences in the nucleotide composition of the zebrafish genome, identifying CpG islands by using standard thresholds is challenging. Because there are fewer CpGs in promoter regions, there is a higher chance for a CpA dinucleotide to be present. CpA dinucleotide is a central part of the INR motif which could explain a higher proportion of INR containing promoters in zebrafish.

Differences in proportions of different core promoter elements between human and zebrafish could be attributed to differences in the nucleotide composition of promoter regions, by the fact that I was comparing developmental stages in zebrafish to human cell lines and tissues or that regulatory profiles of these two organisms are different.

#### 2.3.1.3 Sequence heatmaps

Sequence heatmaps visually present the signal coming from specified genomic ranges. These heatmaps are showing the strength of the PWM signal in the region, where higher the signal, a sequence motif is more similar to the original PWM. It has been shown that TATA-box motif is characteristic of promoters with a sharp initiation pattern. Figure 2.2 shows TATA-box heatmaps of promoter regions for active genes in Muscle (Figure 2.2A) and zebrafish developmental stage Prim 25 (Figures 2.2C). Promoters are centred on dominant TSS (dTSS) and the regions of 500 nt from both sides of dTSS are presented. When we sort width so that sharp promoters are at the top (the first segment of the heatmap), we can observe that the top section of heatmap contains signal coming from the region around -30 nt, exactly where TATA-box should be, while there is no signal at the bottom of the heatmap. Although splitting promoters based on their width into TATA-enriched and TATA-depleted promoters, there are still many sharp promoters that don't contain TATA-box, and there are broad promoters that have TATA-box. Heatmaps that are split based on Promoter Ontology annotation show a much stronger signal coming from -30 region and there are fewer promoters here than in sharp promoter class. Metaplots of these heatmaps, that can be found underneath the heatmap, show average signal at each nucleotide of the promoter region for two groups separately. The first heatmap was split into sharp and broad promoters, while the second one into Promoter Ontology annotations (TATA-less and TATA). Here, we can clearly observe that although sharp promoters are enriched in promoters with TATA-box, we see slight enrichment in broad promoters too. In contrast, when promoters are classified based on Promoter Ontology annotations, the peak from -30 position is significantly stronger (0.16 versus 0.04) than in sharp promoters, and there is no signal in no-TATA group.

Heatmaps for the INR motif in the human dataset (Figure 2.2B), when split by IQ width into broad and sharp, don't show any difference in signal strength. Metaplot confirms how both sharp and broad promoters show similar INR signal levels around dTSS. This suggests that the INR motif in human doesn't have a preference for broad or sharp promoters. After clustering these promoters with PO, a class with no INR shows no signal around dTSS and strength of the signal is in the range of signal coming from more distant regions of the promoter. For promoters that

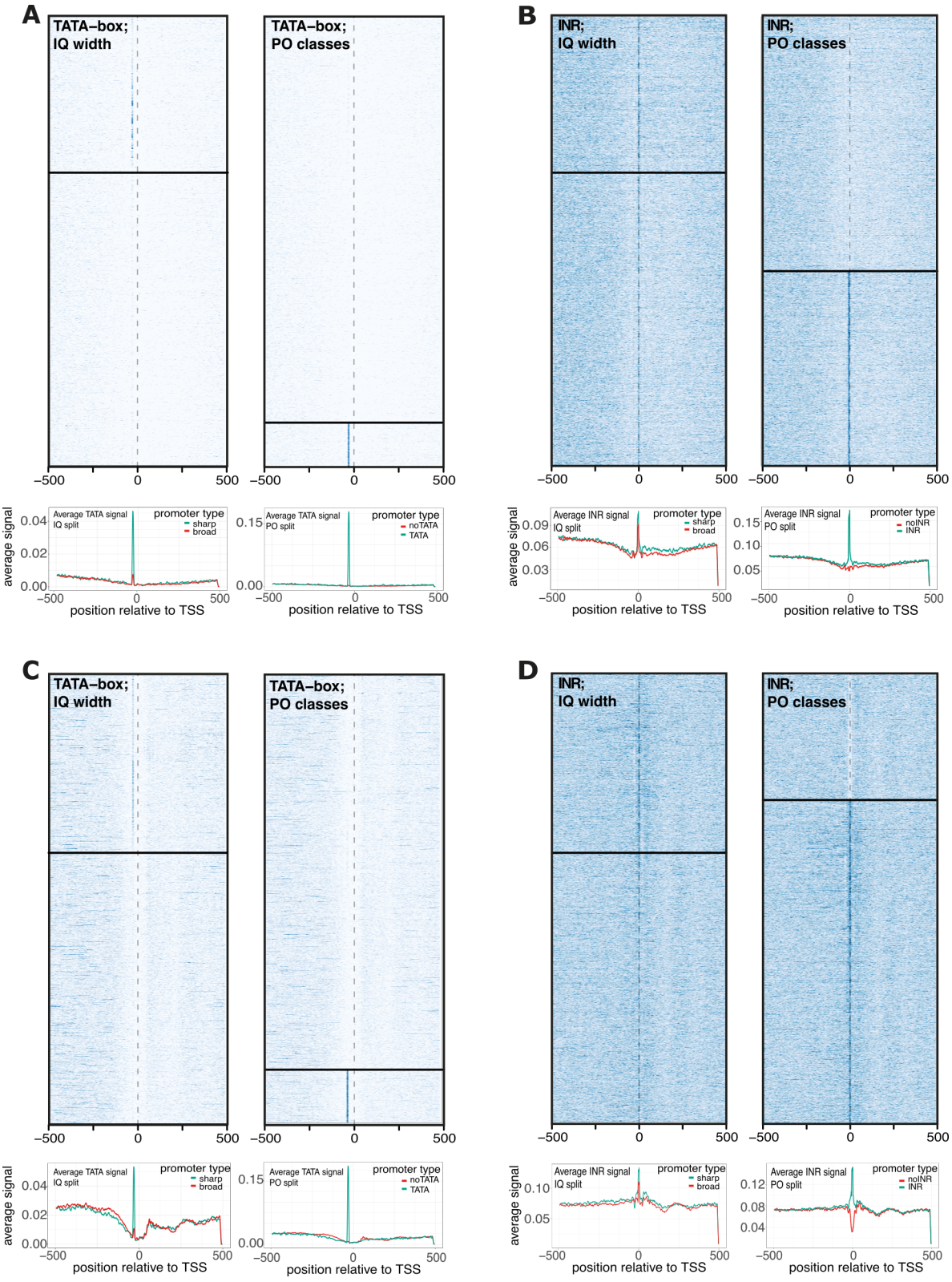


**Figure 2.2: Heatmaps of PWMs signals across promoters.**
Signal presented is a score obtained by PWM scan. In each pair of heatmaps, first one presents promoters sorted by IQ width with a split on sharp and broad promoters. The second heatmap in a pair is split based on Promoter Ontology classification.



*(continued)*

**Figure 2.2:** Heatmap signal is quantified per genomic position and presented in a metaplot that can be found under each of the heatmaps. Assignment of TATA containing and TATA-less promoters for human muscle sample is presented in panel **A**, and zebrafish Prim 25 in panel **C**. Initiator heatmaps for human muscle sample are presented in section **B**, and zebrafish Prim 25 in section **D**.

---

were annotated as containing INR binding site, this class on a metaplot clearly shows enrichment of INR around dTSS in this class. Interestingly, in zebrafish heatmaps for INR (Figure 2.2D), when I stratified promoters by IQ width I observed similar trend as in human promoters, but in case of PO classification, there is a strong depletion of INR signal in promoters not containing INR. Sharp promoters show depletion of INR signal around -30 bp position. That is a signal coming from TATA promoters that in that region has enrichment for weak (A and T) nucleotides.

### 2.3.2 Promoter Ontology can complement Gene Ontology results

The true potential of Promoter Ontology lies in its ease of use, where it could be used in all differential expression analyses alongside to Gene Ontology analysis to get an even deeper understanding of the differences between identified groups of genes. To present this idea, I reanalysed results from a published study (Tarifeño-Saldivia et al. 2017) where authors analysed gene expression of the three most common pancreas cell types isolated from adult zebrafish transgenic lines. The main aim of that study was to identify novel markers and distinct signalling pathways characteristic for each of these cell types. Namely, authors have isolated acinar, ductal, and three types of endocrine cells (alpha, beta and delta cells) and analysed their transcriptomes by doing RNA-seq.

I have obtained their RNA-seq samples and performed differential expression analysis using DESeq2. To identify genes that are characteristic for each of the cell lines, I performed pairwise Wald test. Significance levels from each of comparisons were corrected for multiple testing using Benjamini and Hochberg method. Genes having an adjusted p-value $< 0.05$ and a fold change greater than 4 were considered significantly different between two compared cell types. RNA-seq data from alpha, beta and delta cells were combined together into endocrine expression sample. This was possible since the library sizes of these samples were in a similar range and in this way the authors reduced the number of pair-wise comparisons. Genes that were not significantly different in

any comparison were considered as genes in common.

Gene Ontology analysis revealed that genes characteristic of acinar cells are predominantly involved in metabolic and biosynthetic processes (Figure 2.3A). Ductal cells are enriched for processes such as: "cell adhesion", "response to external stimulus" and "regeneration" (Figure 2.3B). Interestingly, although endocrine cells were enriched for functions like: "cell-cell signalling" or "G-protein coupled receptor signalling pathway", many neuronal terms were enriched for this set of genes, Figure 2.3C. This could be explained by the fact that this sample was created in silico by combining three cell types that are subcomponents of endocrine cells. This way, the authors did not include all endocrine cell types found in a pancreas, which could account for the lack of some endocrine-specific functions. In addition, since libraries were of similar sizes before combining them, the contribution of each cell type to the expression profile was the same, although, in reality, these cells are not present in the similar proportions. This could cause underrepresentation of more abundant cell-types and overrepresentation of less abundant cell-types in the total expression profile of endocrine cells. Nevertheless, neither of these drawbacks could explain how are neuronal function terms overrepresented in this sample. This could only be attributed to inadequate cell collection from transgenic zebrafish line.

Genes characteristic of four identified groups have distinct functions. In addition to finding the function of these genes, I analysed them using Promoter Ontology to identify whether their promoter structure differs. The common group of genes was strongly depleted for TATA-box and GC-box while being enriched for CpG islands and BREd elements (Figure 2.3E). This promoter signature is characteristic of housekeeping genes. Gene Ontology analysis for this group of genes confirmed that many overrepresented functions are housekeeping (Figure 2.3D). Acinar and ductal genes are, on the other hand, enriched for TATA-box and Initiator, while depleted for CpG islands. Although these gene groups share general tissue-specific promoter signature, they show considerable differences: ductal genes are enriched in GC-box and upstream BRE element, while acinar cells are enriched for XCPE1 motif and CCAAT-box. Endocrine-specific genes share enrichment in some promoter elements with both ductal and acinar specific genes. They are enriched for XCPE1, GC-box, MED-1, upstream BRE element and Initiator. Interestingly, endocrine specific genes don't

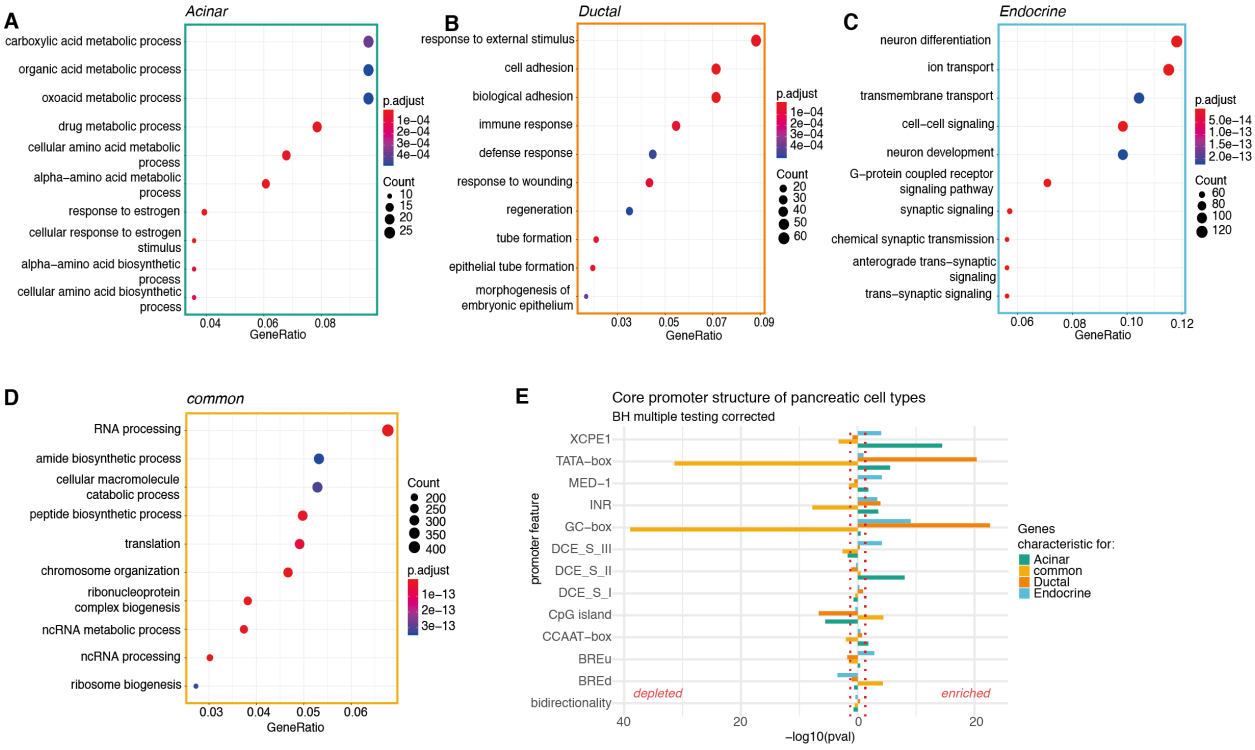


**Figure 2.3: GO and PO enrichment results for four pancreatic cell types.**
Gene ontology enrichment results for genes differentially expressed in only acinar, ductal and endocrine cells are presented in section **A-C**, while GO enrichment for all common genes in section D. (**E**) Promoter Ontology enrichment of core promoter features for these genes. The left side of the graph is representing significantly depleted promoter features, while the right side shows enriched features. A significance threshold of $p < 0.05$ is denoted with a red dashed line. In all gene group comparisons, all genes active in pancreas were used as a background for GO and PO.

seem to have a tissue-specific promoter signature that consists of depletion of CpG islands and enrichment of TATA-box. This observation suggests that this group of genes regulates its promoter activity differently and that way gene expression is primed differently. The only potential drawback to this observation is the fact that since this sample was created *in silico* could change ratios of gene levels obtained, which could disable promoter over-representation to find true promoter structure enrichment.

In this case study, I have shown that Promoter Ontology can inform about promoter structure for a group of genes. In this example, I have identified genes specific for different functions within a pancreas are also regulated in a distinct way. Acinar and endocrine-specific genes rely on XCPE1 to activate transcription since their proportion of TATA-box containing genes is smaller than in ductal genes that contain significantly more TATA-box promoters. A similar trend was also observed with the GC-box.

### 2.3.3 Promoter features of orthologous genes are conserved during evolution

Advancements in genomics are changing the way we understand evolution. Being able to compare genome sequences of organisms across different taxa, and also within genomes, enables a deeper understanding of evolutionary processes that shaped today's genomes. A key evolutionary events shaped the function and structure of genes. The best-studied ones are speciation, gene duplication, horizontal gene transfer and other gene rearrangements. Genes found in diverse species that evolved by speciation are called orthologs. Orthologs evolved from a common ancestral gene that diverged due to evolutionary pressures. Usually, orthologous genes retain the same function in two species.

Since the last common ancestor of zebrafish and humans, genomes of these species diverged largely. Most notably, teleost fish taxa, of which zebrafish is a member, underwent a whole genome duplication about 320 million years ago. This event allowed the emergence of genes with a novel or specialised functions. Nevertheless, when comparing human and zebrafish genes many carry out the same function. In addition, numerous orthologs have been identified. Due to whole genome duplication, some zebrafish genes to this day remained in two copies. Therefore, three subclasses of

human - zebrafish orthologs have been identified. *One2one* group of genes that in both genomes remained as a single copy. Genes that are found in two or more copies in the zebrafish genome, while having a single copy in human genes are classified as *one2many*, and finally, *many2many* genes are present in multiple copies in both genomes.

Although sequences of orthologous genes diverged, their function in large fraction remained the same. To test if promoter structure was also conserved in these genes, we performed Promoter Ontology analysis comparing different classes of orthologs, and comparing them between the two species. Alternatively, promoter structure could be on a more relaxed evolutionary pressure and there are no common promoter features between these species.

I obtained a list of putative human-zebrafish orthologs from Ensembl. Ensembl classifies genes based on the number of zebrafish genes assigned to a human gene (*one2one*, *one2many* and *many2many*). Due to the availability of CAGE data for zebrafish where we only have developmental samples, I have used Prim 25 (36 hpf) sample that is the most developed stage in the dataset. For comparison with the zebrafish developmental sample, I have used ESC as a human sample. Table 2.3 shows a number of identified orthologous genes active in CAGE samples.

**Table 2.3:** Number of orthologs identified for different orthology classes in human and zebrafish.

| type of orthology | human | zebrafish |
|---|---|---|
| One2one | 6595 | 7431 |
| One2many | 2216 | 4103 |
| Many2many | 258 | 445 |

I ran Gene Ontology enrichment using R package ClusterProfiler . For the background set of genes in this analysis, I have used all active genes in the sample. In addition, I also compared promoter structure between orthology classes and the background. The results of these analyses are shown in the Figure 2.4. Promoters for *one2one* orthologs contain genes that are strongly depleted of TATA-box, INR, GC-box and CCAAT-box, and enriched for XCPE1 and CpG islands (Figure 2.4A). This observation suggests that *one2one* orthologs are enriched for housekeeping alike

promoter structure. Indeed, GO enrichment analysis confirms that these genes are enriched for housekeeping functions.

For *one2many* orthologs, GO enrichment predicts that they are mostly involved in the development of the nervous system and regulation of cell movement (Figure 2.4B). Promoters of these genes are enriched for Initiator, GC-box and DCE_S_III in both organisms. Interestingly, human promoters are depleted for CpG island while zebrafish promoters are enriched for the same feature. However, this could be explained by the fact that in this comparison samples are not a perfect match between these two organisms. I tried to use best matching samples, however, zebrafish sample is the whole embryo sample while for human sample, I used ES cells. ESCs are pluripotent stem cells that have a gene expression program that allows them to self-renew and to respond to developmental signals and differentiate into the required cell type. On the other hand, gene expression of a Prim 25 embryo is a compilation of expression profiles of all cells existing in an embryo. Prim 25 embryos are at the final stages of organogenesis, a stage in development where all major organs form. In this period, some organs, like hearth, are already developed and active, while remaining organs precede to develop. This difference could be causing a discrepancy in the composition of promoter features. By annotating and analysing promoters of all active genes within a developing embryo, a great proportion of tissue-specific genes will be included. That could explain an enrichment of GC-box in one2many orthologs since the background set of genes will have many more tissue-specific genes and, in proportion, less housekeeping genes. In addition, human promoters are enriched for TATA-box while zebrafish promoters are enriched for TATA-box alternative XCPE1. These results suggest that core promoters of human *one2many* orthologs are much more tissue-specific than zebrafish promoters. *One2many* group of genes consists predominantly of genes that remained in two copies in the zebrafish genome after the whole genome duplication and the genes that were duplicated multiple times throughout the evolution. These two groups are distinct in terms of their function and conservation. Genes that arose through tandem duplication often contain very simple regulatory landscape that enabled gene to remain active even after translocation into a new environment. On the other hand, genes that remained in both copies after the whole genome duplication are known to be enriched for genes associated with signalling pathways and developmental genes and are most likely increasing the fitness of the organism. These genes have a complex regulatory repertoire and

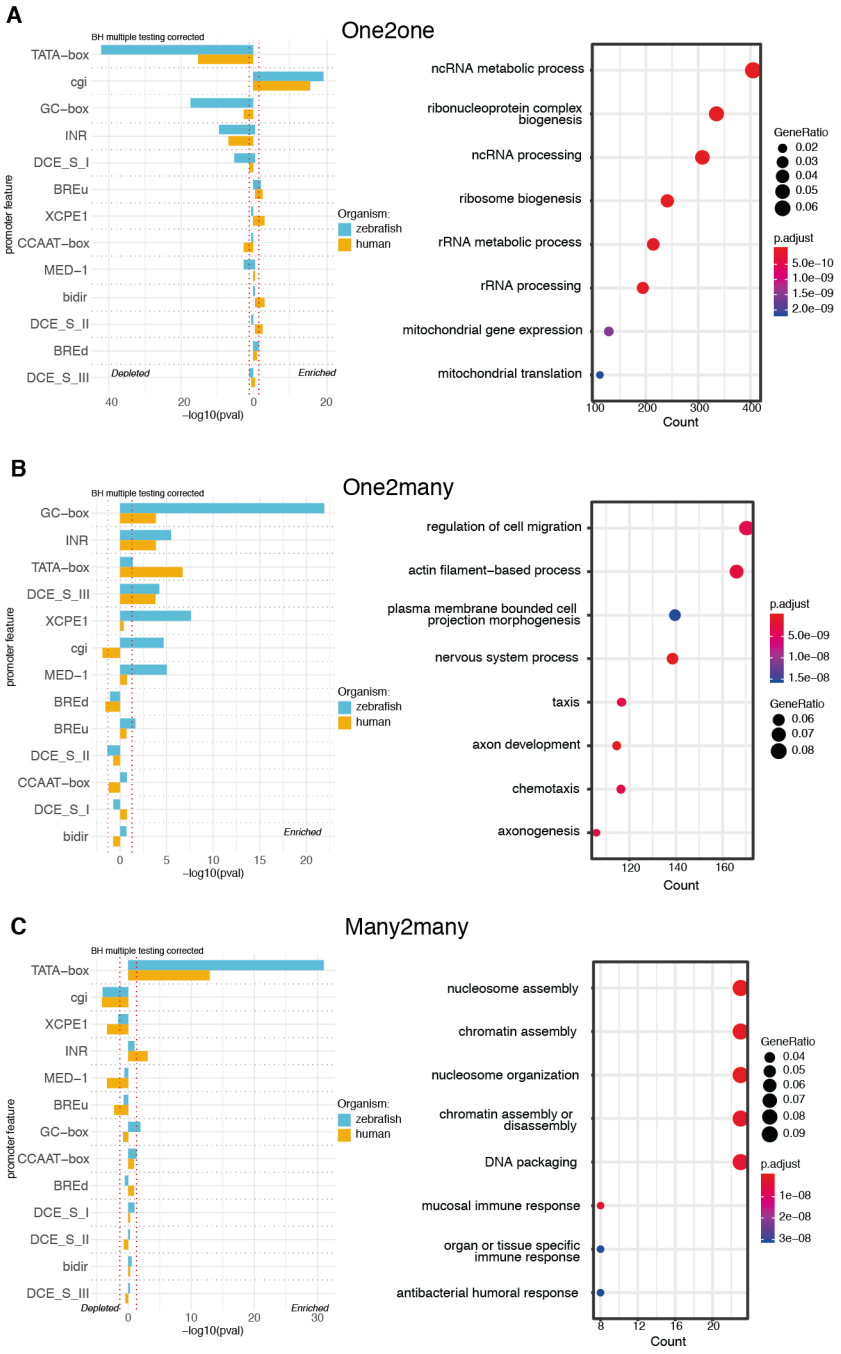


**Figure 2.4: GO and PO enrichment results for three types of human - zebrafish orthologous genes.**
Promoter Ontology results are presented for both zebrafish and human genes that were active in Prim 25 and ESC sample respectively. The left side of the graph is representing significantly depleted promoter features, while the right side shows enriched features. A significance threshold of $p < 0.05$ is denoted with a red dashed line. Gene ontology results represent significantly enriched terms for ESC sample. (**A**) Results for one2one orthologs, (**B**) results for one2many orthologs and (**C**) results for many2many orthologs.

are regulated by both distal and proximal regulatory elements. Clearly, these two groups of genes have very different promoter architectures. To address this, I have further subclustered this group of genes by counting how many associated zebrafish genes have been identified by BioMart for each human gene. *One2two* group of genes should contain only genes that evolved after the whole genome duplication and *one2many* are all other genes, present in more than 2 copies in the zebrafish genome. Unfortunately, I was not able to obtain any GO enrichments or promoter structure characteristic for these classes. This is most likely caused by the limited BioMart orthology mapping because of which I was not able to obtain clear separation between these two groups of genes.

*Many2many* orthologs contain genes that are involved in chromatin assembly and immune response (Figure 2.4C). These genes look like canonical tissue-specific genes with strong enrichment for TATA-box and depletion of CpG islands and XCPE1.

Remarkably, all three orthology classes are showing the same trend when we compare promoters of human and zebrafish. Zebrafish promoters show bigger effect sizes, but this could be due to more orthologous genes identified in CAGE samples used. This suggests that indeed, promoter structure, just like gene function is conserved during evolution.

### 2.3.4 Gene Ontology terms reveal differential core promoter usage

Recent studies have shown that there are at least three distinct promoter architectures (Lenhard, Sandelin, and Carninci 2012), with the additional ones being identified. Based on the function of a gene, promoter structure is adapted to gene's function. That is why tissue-specific genes frequently have a narrow initiation region with characteristic TATA-box motif and depletion of CpG islands. On the other hand, promoters of housekeeping genes contain wide initiation regions that are often overlapped by a CpG island. In addition, these promoters don't require TATA-box for their function.

To better explore the hypothesis that promoters have adapted based on the gene function, I analysed the promoter structure of various gene groups that were clustered based on their function. This way I could obtain a higher power to detect specific promoter structures of more specific gene groups, e.g. instead of all tissue-specific genes having common promoter structure, potentially, there

are subgroups of tissue-specific genes whose promoters adapted to that function. To do this, I have analysed Gene Ontology gene groups. GO annotates all genes with predefined terms that describe gene's function, molecular activities of gene products and localisation where gene products are active. In addition, all genes are connected with genes that share some of its features.

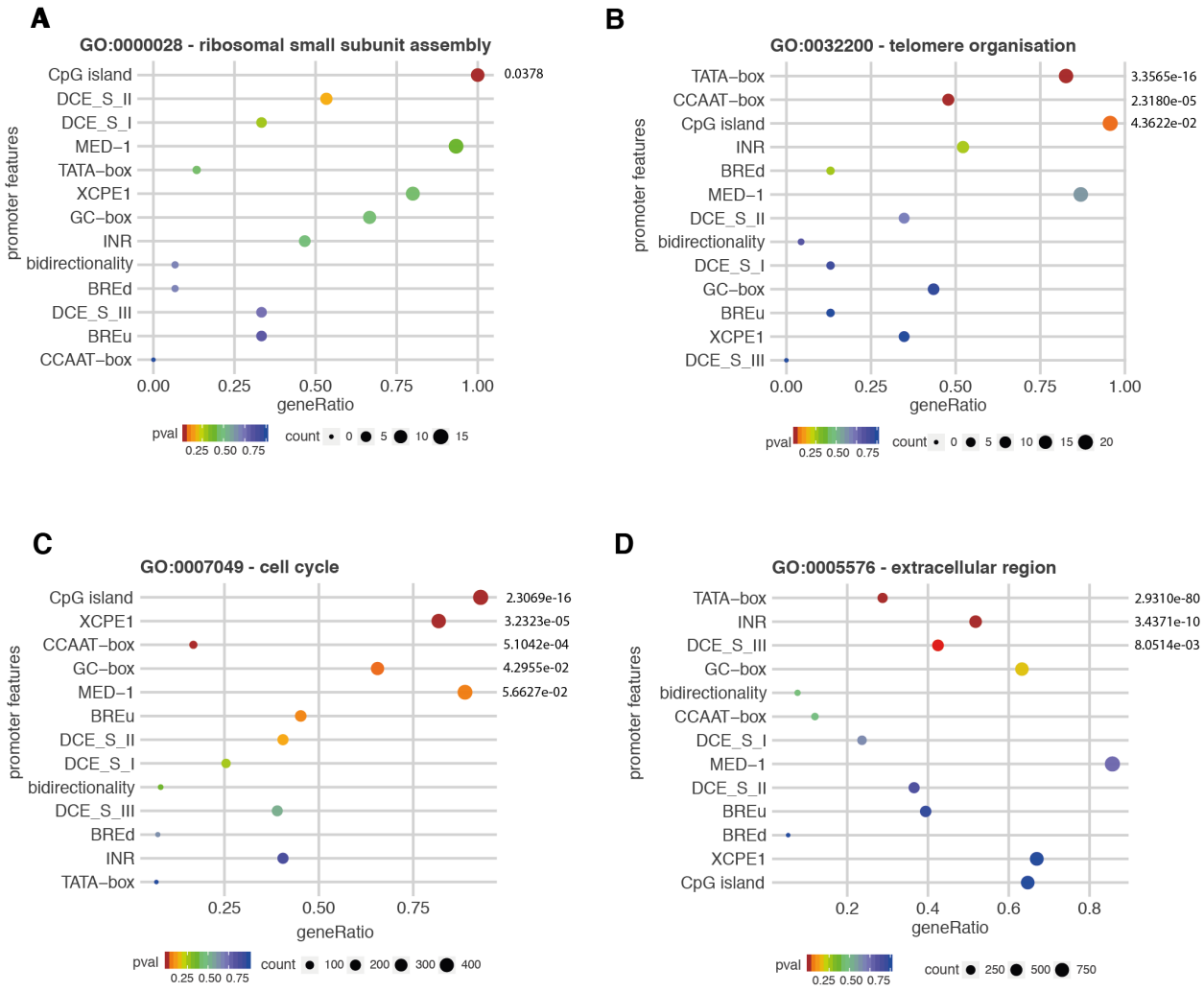


**Figure 2.5: Examples of Promoter Ontology results for four GO gene groups.** Promoter features are sorted by the level of enrichment with exact enrichment stated on the right of the figure. (**A**) Results for Ribosomal small subunit assembly gene group in which all promoters have CpG island. This group is relatively small with only 15 genes. (**B**) Telomere organization gene group. This group describes 23 genes, many which contain both TATA-box and CpG island. (**C**) Cell cycle gene group contains typical housekeeping promoter profile with strong enrichment for CpG islands and depletion of TATA-box and final example of (**D**) extracellular region, a large gene group whose promoters are significantly enriched for TATA-box and Initiator.

I have extracted all gene ontology groups from Gene Ontology Consortium. The three

ontologies (Biological process, Molecular function and Cellular compartment) contain 44 917 gene groups in total. To obtain all genes associated with a GO group, I queried BioMart by using GO.db R package. Most of these groups are small gene groups. To make promoter structure analysis more robust, I have omitted all GO groups with fewer than five genes annotated. This way, 2846 GO groups remained for further analysis. I ran Promoter Ontology on each GO group using all genes active in muscle CAGE sample as a background. In this analysis, I have identified 2414 GO groups that contained at least one promoter feature significantly different from the background. Figure 2.5 presents Promoter Ontology results for four GO groups. In the first example (Figure 2.5A), all genes in a gene set "Ribosomal small subunit assembly" contain CpG island in their promoter region. This observation suggests that for the initiation of expression, these genes, due to their high GC content, rely on destabilisation of nucleosomes around the promoter region which allow general transcription factors to bind. Ribosomal subunit genes are known to contain TCT motif in their core promoters. Promoter Ontology by default uses JASPAR PolII collection of core promoter motifs, and TCT motif is not one of those. To make comparisons across different gene groups consistent, I did not include that motif in the Figure 2.5A. However, when I annotated core promoters using TCT PWM from JASPAR, out of 15 genes in this group, 13 genes contained TCT motif around their TSS. Another example of a GO group that is enriched for CpG island is a GO group "cell cycle" (Figure 2.5C). Along with strong enrichment for CpG island, this group is also enriched for XCPE1, CCAAT-box and GC-box. These promoter elements are known to enhance expression and exert specific promoter activity that could be required by genes that are tightly regulated in the cell cycle.

An additional example is presented in Figure 2.5D. This is a group of genes belonging to the GO group "extracellular region". These genes are strongly enriched for TATA-box and Initiator, along with DCE_S_III. Genes in this group belong to signalling molecules, hormones and structural proteins, all tissue-specific functions that confirm promoter structure findings. A final GO group presented in Figure 2.5 is "telomere organisation" GO group (Figure 2.5B). This group of genes is both significant for TATA-box and CpG island feature along with CCAAT-box enrichment. This promoter structure is not common since TATA-box is often found in sharp promoters, while CpG islands are usually part of broad promoters. It could be considered that although both of these

features are significant, they don't necessarily have to be present on the same promoter. In case of "telomere organisation" genes, that is not the case since CpG island is present in 96% of promoters, while TATA-box is present in 83% of promoters, which means that there is a high overlap between these two features. However, besides genes involved in telomere organisation, there are genes that possess both functional elements. Some of those genes are MyoD1, erythropoietin and $\alpha$-globin.

#### 2.3.4.1 Similar GO gene groups are more likely to have common core promoter features

GO terms cluster genes into functional categories. Due to its hierarchical nature, child terms represent a more specific function. Conserved functional role of genes annotated into the same GO term could suggest that their promoter structure is also conserved. To explore this hypothesis, I tested if genes having similar function should have similar core promoters too. In addition, I explored the difference in core promoter structure between parent and child gene groups; is it the case that with a more specialised function, child terms concentrate some core promoter features or they obtain novel features.

GO vocabulary consists of three ontologies: biological processes (BP), cellular compartments (CC), molecular functions (MF). The composition of core promoter features differs for these groups (Figure 2.6A). GO terms belonging to biological processes ontology have the highest proportion of terms significantly enriched for TATA-box core promoters while cellular components contain most terms significant for promoters with CpG islands. Biological processes ontology is a significantly bigger ontology with more GO terms compared to other ontologies. Since GO ontologies are built in a hierarchical fashion, bigger ontologies have more branches, which in this case would represent more gene groups with a more specific function. This can explain a higher proportion of classes significant for TATA-box. In addition, CC ontology contains significantly more gene groups enriched for CpG island. Many housekeeping genes are annotated by the location of where gene product resides, some examples would be transmembrane proteins or nuclear proteins. This could provide an explanation for the enrichment of CpG containing promoters in CC annotated genes.

Next, I have analysed promoter structure of 1899 parent-child GO term pairs for which

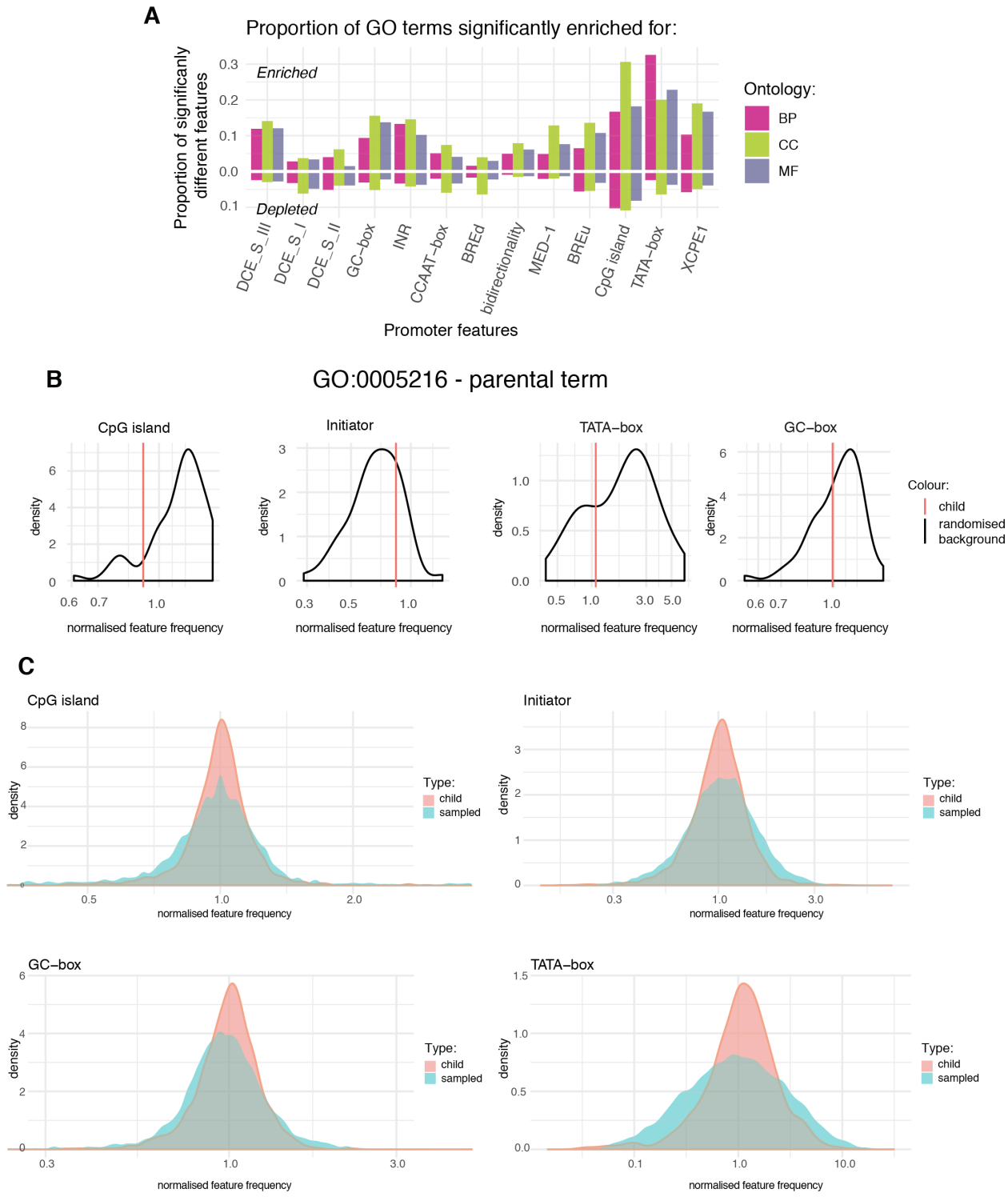


**Figure 2.6: Promoter features of GO parent-child gene groups.**
(**A**) Proportion of significantly enriched promoter features in all GO gene groups having more than 5 genes active in muscle tissue. Results are stratified for three GO ontologies and significantly enriched gene groups are presented above zero, while depleted gene groups below. (**B**) An example of GO parent-child promoter feature inheritance.

*(continued)*

**Figure 2.6:** Proportion of promoters containing promoter feature are normalised based on parental feature, so gene groups closer to 1 are more similar to parental proportions. Child proportion of promoter features is reported as a red bar, and as a contrast 100 randomly sampled GO groups of similar size to child groups are shown as a density distribution. (**C**) Density distribution of normalised promoter feature proportions for all child terms of all GO gene groups (red). In addition, for all GO groups, randomly sampled 100 GO terms are merged and presented in blue.

---

both parent and child had at least 5 genes active in muscle tissue. All parent-child relationships in this analysis are either "is a" or "part of". The first question I was interested in was if the child terms have more similar promoter structure to parental structure than random GO gene groups. To answer this, for each GO gene group, I have randomly sampled 100 unrelated gene groups. These gene groups represented the background with which I compared child promoter structure. To make all frequencies of occurrence easier to compare, for all gene groups, I have normalised the frequency of occurrence of promoter features by parental frequency. This way, gene groups having very similar promoter structure to parental structure will have occurrence values close to the value of one. For each parental GO group, I have compared background distribution of normalised frequency to normalised frequency of child GO terms. An example of a similarity comparison of some promoter features for GO:0005216 GO term (ion channel activity) can be found in the Figure 2.6B. In this example, it can be observed that child normalised proportions are close to the parental promoter feature proportion, while the distribution of randomly sampled GO terms is not necessarily centred around parental, but rather, shows a much wider spread of values. Deviation from parental frequency depends on the initial occurrence of parental promoter features. This is the reason why in the case of TATA-box, which is present in less than 10% of promoters, a much bigger spread of values can be observed than in more abundant features, like CpG island or Initiator which are present in about 80% of promoters.

To globally quantify the similarity in promoter structure between child terms and randomly sampled GO gene groups, I have combined all normalised frequencies of all randomised GO gene groups and all child normalised frequencies. This way, I expected that child distribution will be significantly more centred around parental value. Distributions for the four most characteristic promoter elements can be seen in Figure 2.6C. In all distribution pairs, child proportions are

significantly more centred around the parental proportion. Extra support has been provided in the Table 2.4. It presents statistics for sampled and child distributions of all JASPAR PolII promoter features. For all features, interquantile range (IQR) of distributions is more narrow for a child than randomly sampled GO gene groups. To show that child distribution are indeed more centred around parental value, I am reporting $5^{th}$ and $95^{th}$ quantile where for all child distributions both quantiles are closer to 1. Kurtosis is a statistical measure that describes "tailedness" of a distribution, where higher values describe more heavy-tailed distributions. When compared, in almost all promoter features, the distributions of random GO gene groups are more heavy-tailed than child distributions. The only exception to this observation are BREd and MED-1. The last column of the table provides statistical significance calculated by analysing similarity of the child and random distribution using Kolmogorov-Smirnov test.

**Table 2.4:** Statistics for child and sampled distribution of occurrance frequency for promoter features.

| | child | sampled | child | sampled | child | sampled | child | sampled | |
|---|---|---|---|---|---|---|---|---|---|
| Promoter feature | IQR | IQR | kurtosis | kurtosis | 5% quantile | 5% quantile | 95% quantile | 95% quantile | p value |
| CpG island | 0.157 | 0.245 | 33.92 | 43.867 | 0.741 | 0.6 | 1.259 | 1.495 | 2.220e-16 |
| TATA-box | 1.053 | 1.544 | 17.945 | 120.726 | 0 | 0 | 3.550 | 5.318 | 0 |
| GC-box | 0.229 | 0.280 | 2.821 | 3.408 | 0.691 | 0.638 | 1.346 | 1.426 | 5.449e-05 |
| Initiator | 0.366 | 0.474 | 10.732 | 16.965 | 0.590 | 0.530 | 1.623 | 1.867 | 2.666e-11 |
| BREd | 1.332 | 1.466 | 23.081 | 10.383 | 0 | 0 | 3.0 | 3.455 | 4.727e-07 |
| BREu | 0.360 | 0.491 | 5.913 | 11.174 | 0.523 | 0.434 | 1.615 | 1.875 | 6.531e-10 |
| CCAAT-box | 0.929 | 1.146 | 24.947 | 47.830 | 0 | 0 | 2.907 | 3.5 | 1.082e-08 |
| DCE_S_I | 0.558 | 0.706 | 9.855 | 32.26 | 0.333 | 0.3 | 2.143 | 2.455 | 5.465e-06 |
| DCE_S_II | 0.399 | 0.498 | 12.583 | 39.146 | 0.514 | 0.398 | 1.710 | 1.827 | 2.408e-09 |
| DCE_S_III | 0.355 | 0.455 | 10.732 | 16.965 | 8.233 | 14.262 | 1.6 | 1.805 | 2.11e-09 |

| | child | sampled | child | sampled | child | sampled | child | sampled | |
|---|---|---|---|---|---|---|---|---|---|
| MED-1 | 0.108 | 0.136 | 18.924 | 17.958 | 0.853 | 0.817 | 1.162 | 1.19 | 6.492e-06 |
| XCPE1 | 0.174 | 0.237 | 6.194 | 4.075 | 0.753 | 0.676 | 1.277 | 1.387 | 3.282e-10 |

After showing that in general, child GO gene groups are more similar to their parental group than randomly sampled groups, I decided to look into specific cases where child terms are significantly different from parents in some of the promoter features. Observed difference could be attributed to several reasons. It could be argued that child gene groups that adapted to their more specific function by concentrating or diluting promoter features. On the other hand, promoter features could be divided between children GO groups randomly. To test if there is a pattern in promoter structure inheritance, I first identified all child GO gene groups that possess significantly different promoter composition. For each of 1899 child-parent pairs, I ran a hypergeometric test to determine if the split of promoter feature in a smaller gene subset (child) was random. In this case, the occurrence of promoter features of parental gene group was used as a background, while child occurrences were used as query. Finally, 770 pairs had at least one promoter feature significantly different between a parent and a child.

Figure 2.7 shows an example where one child acquired significantly more TATA-box than parental GO group or any other child. Parental GO, in this case, is GO:0000785: "chromatin", and significantly different child GO group is GO:0000786: "nucleosome". GO hierarchies are built as a directed acyclic graph which allows terms to have multiple parents which could result in some child gene groups being more different due to other parental terms. For this reason, I tested if the structure of GO hierarchies of child groups are significantly different from each other. To do this, I have performed a pair-wise semantic similarity analysis of all child and parent terms. Semantic similarity informs how similar two GO gene groups are based on the common ancestor terms and their annotation statistics. Results of semantic

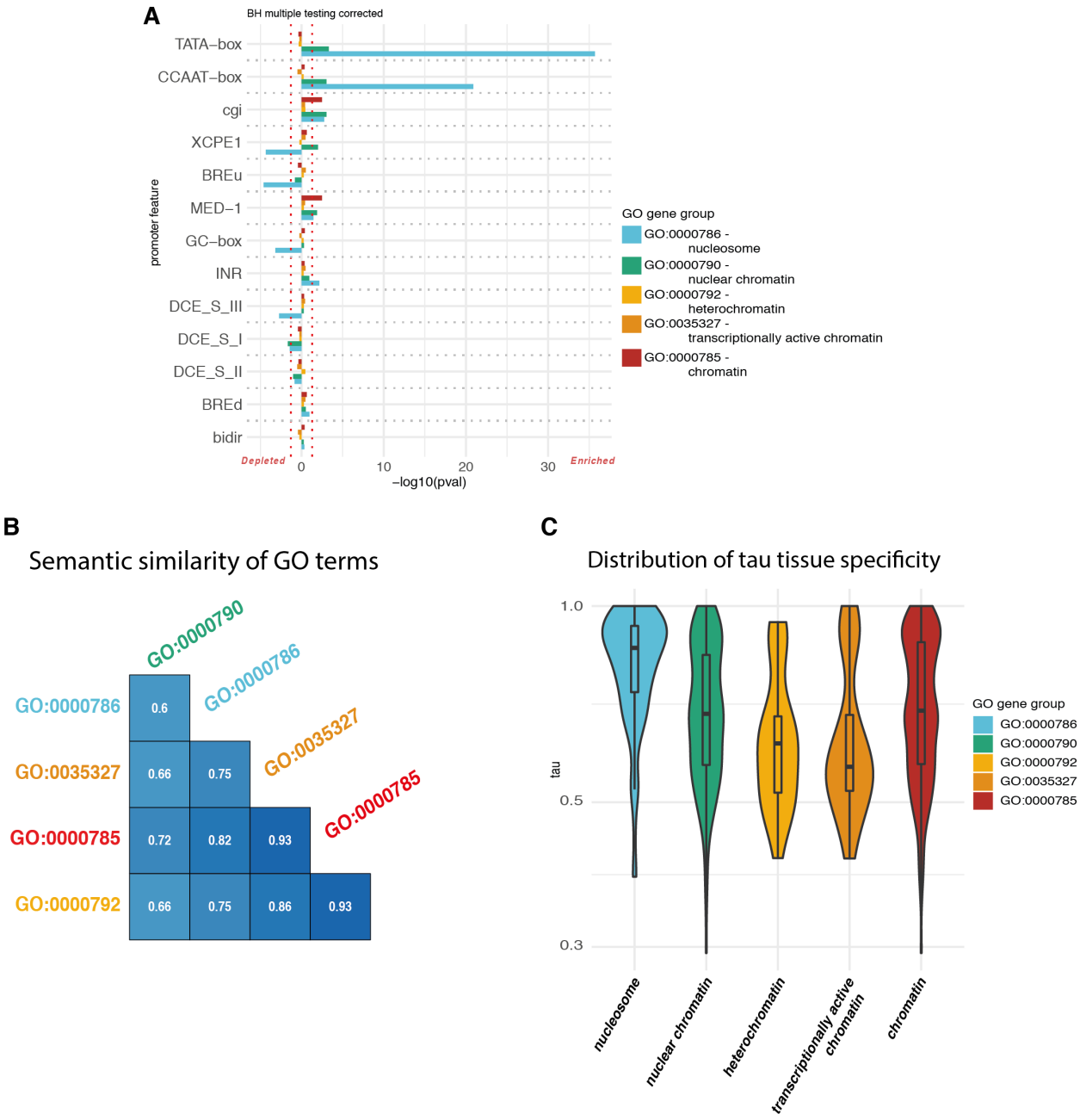


**Figure 2.7: An example GO parent-child pair in which child possesses significantly different promoter structure.**

(**A**) Promoter Ontology enrichment of core promoter features for GO gene groups. The left side of the graph is representing significantly depleted promoter features, while the right side shows enriched features. A significance threshold of $p < 0.05$ is denoted with a red dashed line. In all gene group comparisons, all genes active in human muscle tissue sample were used as a background for PO. (**B**) Pair-wise semantic similarity of analysed gene groups. Wang method for semantic similarity based on the graph structure of GO was used. In this case GO groups with similarity close to 1, have a similar GO ontology structure, while GO groups with values close to 0 don't have GO ancestors in common. (**C**) Distribution of tissue specificity values for each of analysed GO groups. $\tau$ coefficient was calculated using RNA expression from 16 adult human tissues from Illumina's BodyMap 2.0 dataset.

similarity analysis can be found in Figure 2.7B. GO gene groups that are most similar to the parental group are heterochromatin and transcriptionally active chromatin. Surprisingly, the most dissimilar group to all other gene groups is nuclear chromatin. This observation could be explained by the fact that this GO term has additional parental terms: nucleus and nuclear chromosome part that have more genes in common with nuclear chromatin term than chromatin. Despite semantic differences, promoter structure of these gene groups are alike, except for nuclear chromatin being enriched for TATA-box. In case of this parent-child GO gene groups, the semantic similarity was not able to explain the difference in promoter structure of one of the child terms. Next, I decided to analyse if some of these gene groups specifically became more tissue-specific and in that way became more specialised for their role. To do this, I have calculated $\tau$ index using expression information of 16 human tissues extracted from Illumina's Human BodyMap 2.0. Figure 2.7C represents the distribution of $\tau$ values for each of these gene groups. Chromatin and nuclear chromatin gene groups have a relatively uniform distribution of $\tau$ values, with similar amounts of genes uniformly expressed across tissues and tissue specific genes. On the other hand, the nucleosome gene group is significantly more tissue specific than any other group. Due to the fact that these genes, in comparison to their paternal group, specialised for tissue specific function, their promoter in turn, also show significant enrichment in TATA-box and CCAAT-box promoters along with being depleted for TATA-box alternative, XCPE1.

## 2.4 Discussion

In this chapter, I developed a novel method for the analysis of overrepresented core promoter features in the group of genes. The presented method allows the user to define promoter regions from a CAGE-seq experiment. It then annotates promoters with defined features of interest. Currently, the method supports PWM-based and genomic range-based features to be annotated. Once the set of promoter annotations is available, this method performs over-enrichment analysis, for a group of genes, to identify core promoter features overrepresented in the sample. Finally, in the case of multiple gene groups, it is possible to compare their core promoter structure.

I presented the annotations of core promoters with PolII motifs for four human and four zebrafish samples. Human samples originated from two cell lines and two tissue samples. Zebrafish samples were embryonic samples from four different time points, two before MZT, and two after it. This way, it was possible to compare the differences in promoter composition of genes active in different stages of development. I used core promoter annotations to analyse core promoter structure of orthologous gene groups between human and zebrafish. Different orthology gene groups between themselves show distinct core promoter composition. However, core promoters over-representation of the same orthologous gene group between two species remained the same. Finally, I showed gene groups with similar biological function are more likely to also have similar core promoter features than a random gene group.

Importantly, many core promoter features require precise positioning relative to the TSS to be functional. The Promoter Ontology method relies on the promoter location defined by CAGE-seq. CAGE-seq can map transcription initiation events at single nucleotide resolution. This way, Promoter Ontology can remove many false-positive annotations. However, this could be a potential drawback of this method since CAGE-seq data is not readily available and for many tissue types CAGA-seq data is not available. Promoter Ontology could be applied to the CAGE-seq data of similar tissue. In the extreme case,

RNA-seq data could be used to calculate overrepresentation of promoter features, but in that case, annotations of motif-based features are not going to be as accurate since dominant promoter used would be UCSC annotated promoter.

Further, Promoter Ontology may be used alongside Gene Ontology analysis. In this way, users are able at the same time, for their group of genes of interest, discover the potential biological function of these genes and their over-represented core promoter features. By expanding annotations with the additional sequence motifs and epigenetic features such as histone marks and DNA methylation, Promoter Ontology could provide even more detail into the way gene groups are regulated.

In this chapter, when looking at the enrichment of promoter features of GO gene groups, I used GO gene groups obtained from BioMart which contain only genes involved in a given biological process. However, in the comparison of promoter structure between parent and all child terms. For both parent and child, I have extracted genes that are involved in all offspring GO terms. Some other platforms that allow extracting of genes from GO terms, like QuickGo (Binns et al. 2009). By expanding these two groups of genes to include all offspring terms, I include more genes that are related by their biological process or molecular function. Having larger groups of genes gives more power to detect differences between child and parent promoter structure.

Promoter Ontology reports significantly enriched promoter structures for a given set of genes. However, sometimes it would be beneficial to find a subset of the original list of genes that would be even more enriched for a set of promoter features. A similar idea is used in Gene Set Enrichment Analysis to define a leading-edge subset of genes (Subramanian et al. 2005). The leading-edge subset in that analysis can be interpreted as the core set of genes that account for the majority of the enrichment signal in the whole set of genes. Leading-edge alike definition of an optimal subset of genes could be implemented in two ways, supervised or unsupervised. In the supervised implementation, the user could provide a set of promoter

features for which enrichment would be optimised. On the other hand, in the unsupervised implementation of the method, significantly enriched promoter features would be defined by running Promoter Ontology with the entire set of genes. A subset of genes would be defined based on the highest enrichment of the promoter features that were significantly enriched in the first run of Promoter Ontology. The obtained subset of genes could have more power to detect some additional promoter features enriched, to see to which extent enrichment of all promoter features changed, I would rerun Promoter Ontology again.

This advancement would make Promoter Ontology a useful tool for the identification of additional promoter classes and their equivalent architectures in other organisms since in that case Promoter Ontology would have more power to detect subsets of genes with characteristic promoter composition.

# 3 Spatio-temporal complexity of gene expression during embryonic development in zebrafish

## 3.1 Introduction

Embryonic development is characterised by a gradual differentiation of embryonic cells, from phenotypically identical cells in early stages of development to fully differentiated cells with distinct biological functions a few days later. If organisms did not have this ability, all cells would behave the same and development of tissues and organs would not be possible. This process has been extensively researched, leading to the identification of genes important in early development. Genes essential for embryonic differentiation were first identified in *Drosophila* by their loss-of-function phenotype (Nüsslein-Volhard 1994). The majority of these genes in amniotes are transcription factors and signalling molecules. For instance, in zebrafish axis formation, 90% of genes essential for this process are signalling molecules and transcription factors (Schier and Talbot 2005). An interesting feature of these genes is that they have an elaborate spatio-temporal pattern of expression across the embryo. Spatiotemporal regulation of gene expression allows some cells to express a gene and make use of its gene product, while in other cells to have this gene inactive. Precise gene expression is a key requirement for proper embryonic development.

A well-established method for analysing spatio-temporal pattern of expression is ISH. ISH is a technique that identifies the location of the gene expression at a specific time point during development. In addition to localisation, ISH can also be used to measure gene expression quantities within the identified tissue. However, quantification of expression from ISH can be very difficult to optimise and compare across experiments, since it depends on

the length of hybridisation reaction and the sensitivity of signal detection. For accurate quantification of gene expression, NGS methods have been used. RNA-seq provides a comprehensive quantification of all genes active in the sample with significantly better reproducibility than ISH. NGS methods were instrumental for understanding different gene structures, co-activity of genes their regulation etc. However, these methods are unable to devise a spatial component of gene's expression. In early developmental experiments, sequencing is often performed on the whole embryos. Although in those experiments we measure the activity of all genes in the sample, a signal we receive is averaged across all cells in the sample. Due to the drawbacks of both methods, a complete understanding of spatiotemporal developmental expression is lacking.

In this chapter, I have analysed annotations from all ISH studies performed on zebrafish documented at ZFIN and combined this information with RNA-seq expression measurements from the whole embryo. Having these two datasets combined, I was able to identify genes with the highest spatio-temporal complexity. Spatio-temporal gene expression complexity manifests in a selective expression of a gene in defined regions of an organism at a specified level of activity. I hypothesized that developmentally essential genes will be the ones with the highest expression complexity. To be able to compare gene expression of genes whose localisation is not overlapping I have created a coefficient of anatomical specificity that describes how precisely an anatomical structure is defined. This coefficient enabled me to learn general patterns of gene expression across development. For example, I could identify if a gene's expression is becoming more specific for a set of anatomical structures or is the expression of a gene present throughout all anatomical systems in similar quantities. With the extracted information of spatiotemporal expression, I built a resource that enables browsing all available ISH, patterns of expression and dynamics of anatomical specificity. Besides, this resource supports finding genes expressed in specified anatomical structures and comparison of gene expression patterns.

## 3.2 Methods

### 3.2.1 Data sources and initial filtering

Spatial distribution of gene expression across the organism is an important factor determining gene expression complexity. Gene expression complexity can be described as the variability of expression. The more complex expression is, the more variability gene shows across different cell types and across time. To obtain the information on how spatially divergent gene expression is, I have used mRNA *in situ* hybridisation data from zebrafish development. All zebrafish mRNA *in situ* gene expression data used in this study come from the ZFIN database ("The Zebrafish Information Network - Gene Expression Data" 2019). This dataset contains curated gene expression information from 5618 publications. Each publication is represented as a set of descriptors that specify the experimental method used, genes analysed, where and when gene expression activity was observed. Some genes were analysed only by a single or a small number of publications. To get more robust results on the spatial distribution of expression, I filtered out all genes that were supported by fewer than five publications. Although this dataset is curated, I have encountered several mistakes in the annotations that reoccur throughout the dataset. For example, some anatomical structures, used to annotate gene expression, did not match the developmental stage of the analysed embryo. To ensure that data can be used for data mining, I have introduced additional filtering steps. For this, I have extracted all anatomical structures present at each developmental stage from anatomical descriptions obtained from ZFIN's AO system (Van Slyke et al. 2014). With this controlled vocabulary, I ensured that all data points report expression in anatomical structures appearing in the corresponding stage. Finally, I have omitted data from all publications that had at least 50% of non-matching anatomical structure and developmental stage in which these structures were reported.

### 3.2.2 A measure of anatomical specificity derived from anatomical descriptions

After filtering, the entire dataset described the expression in 1078 anatomical structures. Of these anatomical terms, many described expression in a specific cell type, while some other terms described much larger structures encompassing multiple cell types or even whole anatomical systems. Clearly these anatomical structures will have different contribution towards spatial dynamics and expression complexity. In order to define how much each anatomical structure contributes to the spatial divergence of expression, I have created a coefficient of anatomical specificity that informs how precisely defined an anatomical term is. The higher the value of the coefficient is, more specific anatomical structure is.

To establish a measure of specificity I extracted anatomical structure descriptions from zebrafish AO system. Not all anatomical structures were included in the AO system, so I manually curated 258 structures that were mentioned in publications and linked them to corresponding structures from which they develop.

The relationships "part of" and "is a" from AO system were used to create hierarchies of anatomical systems. While creating hierarchies, the only anatomical terms that did not have its parental terms were anatomical systems, so they were used as roots of 20 hierarchies representing anatomical systems throughout development. Starting from whole systems, first I extracted terms that they can be subdivided into: this process was repeated until all structures were linked in hierarchies. Anatomical terms that were described by "develops from" relationship, were kept at the same position in the hierarchy. I assigned a coefficient of anatomical specificity to all structures; it represents the relative position of an anatomical term in the hierarchy. I tested how very general terms separate from specific ones by using three different coefficients:

- Linear coefficient assigns, to all structures at a specific level of the tree, a value that is higher by one than the previous, lower level.
- Exponential coefficient assigns an exponential value of the corresponding level to all

structures in that level.

- Finally, I tested a coefficient that accounted for the shape of the tree. For each level of the hierarchy this coefficient assigns a value that represented the total number of terms that are more general than terms on that level (above that term).

For each coefficient, I calculated variation in the distribution of specificity scores for each developmental stage. Exponential coefficient shows the best separation of anatomical terms in the period of organogenesis, while linear coefficient separates early stages the best. Since most of the data points come from organogenesis, I continued using the exponential coefficient. All coefficient values were normalised so that values are in the range of 0 to 1.

### 3.2.3 Imputation of missing data points

To define the optimal number of stages to impute, I tested what would be the information gain to the expression table if we would impute the gaps in the expression of various widths. Along with the information on how many new data points would be introduced to the expression table, we considered the extent of morphological change that an embryo undergoes in the imputed period. The maximum gap of 3 developmental stages was considered as the optimal amount of information gain and noise added to the system.

There were several cases in which I was able to impute data points:

- In the case of a gene which has reported expression in the same anatomical structure in the stages before and after the missing stages, the same structure was assigned to the missing stages along with its descriptors.

- In a case where structures before and after the gap were different, but belong to the same anatomical system, I would check whether succeeding structure develops from the earlier structure. In that case, structures would be assigned according to the timeline of structure development.

- If the two structures did not develop one from the other, I assigned only the information

about expression in the corresponding anatomical system and not the structures.

All other cases could not be confidently imputed, so I did not assign any additional information about expression in those stages.

### 3.2.4 Similarity in gene expression of anatomical systems and WGCNA

From the expression table, I calculated Pearson correlation coefficient of expression specificity for each pair of anatomical systems. To make the correlation more accurate, I calculated correlation separately for each of the developmental stages. If for a specific pair of anatomical systems there were fewer than ten coexpressed genes, I would not calculate correlation score for that pair.

Weighted gene coexpression network analysis (WGCNA) was performed to identify genes that have similar gene expression profiles across development. For that analysis, I used WGCNA R package (Langfelder and Horvath 2008). Expression table was transformed to a stack of matrices where each matrix represented the expression profile of a single gene with rows representing the anatomical structures and columns developmental stages. Pearson correlation was then calculated for every pair of matrices. To reduce noise and emphasize strong correlations between coexpressed genes, I weighted the Pearson correlations by taking their absolute value and raising them to the power ß. ß was defined after calculating mean connectivity and goodness of fit for a range of coefficients by using pickSoftThreshold function from the WGCNA package. Weighted correlations were then transformed into topological overlap matrices (TOMs) by using TOMsimilarityFromExpr function. Modules of genes with similar expression reports are finally detected by performing soft clustering on the TOM dissimilarity values. The TOM dissimilarity values, in turn, represent the connection strengths between genes in the network and are used to build a dendrogram.

The modules of coexpressed genes from WGCNA were functionally annotated with GO enrichment test. GO enrichment analysis was performed in R using the GOstats package

(Falcon and Gentleman 2007). Genes from each module were tested against all genes from expression table for overrepresentation of particular terms in biological process and molecular function ontology by using hypergeometric test. The p-values obtained from each test were corrected for multiple testing by Benjamini-Hochberg procedure (Reiner, Yekutieli, and Benjamini 2003).

### 3.2.5 Exploratory factor analysis

For the exploratory factor analysis (EFA), I prepared a wide matrix in which the expression of each gene was described by all combinations of anatomical structure and stage in which that structure was observed. The values used in that matrix were anatomical specificity values of each structure. Next, I calculated the pair-wise correlation of all stage-structure pairs. Using the correlation matrix, I calculated eigenvalues of all covariates by performing principal component analysis. Based on the distribution of eigenvalues I have decided to perform exploratory factor method for a set of factors (4, 12, 18, 32, 54, 98, 131, 256, 378, 425 and 585) using "oblimin" as factor rotation method. These factors were chosen as the values that sample the "elbow" and minimal and maximal number (eigenvalue $> 1$) of factors from the eigenvalue distribution curve (Figure 3.5A). These factors determine the final size of the output matrix. Exploratory factor analysis was performed using psych package in R (Revelle and Revelle 2007). From the obtained factor loadings, I extracted the weights used to describe each of the genes by using "tenBerge" method. It finds weights such that the correlation between factors for an oblique solution is preserved. Visualisation of the obtained factor weights was done using pheatmap R package (Kolde 2012).

### 3.2.6 Adult tissue specificity and developmental temporal specificity calculation

To calculate adult tissue specificity, I downloaded RNA-seq data from (Hu et al. 2015). This is a study that profiled gene expression of 8 adult zebrafish tissues at three different temperatures. I obtained FASTQ sequencing files of these samples under GEO

accession number GSE62221. All files were mapped to danRer10 genome by using STAR (Dobin et al. 2013). Mapping was performed with slight modification of default parameters: "outFilterScoreMinOverLread 0.3; alignSJoverhangMin 15; outFilterMismatchNmax 33". Reads were then quantified per gene they were overlapping, and final gene count table was created. Normalisation and variance stabilizing transformation of gene counts was performed using the DESeq2 package (Love, Anders, and Huber 2014). Variance stabilizing transformation was run with blind = TRUE parameter. After extracting normalised TPMs for all eight tissue samples, I used formula for $\tau$ index (Equation 3) to obtain tissue specificity information for all genes active in these samples.

To calculate temporal specificity, I obtained RNA-seq data from (White et al. 2017). This study profiled the expression of mRNA during the 18 stages of zebrafish development, from 1 cell to 5 days post-fertilisation. The data were mapped just like (Hu et al. 2015) study. This dataset contained five biological replicates for each timepoint of development, so when normalising the data, I used variance stabilizing transformation with blind = FALSE parameter. Temporal specificity is calculated as an inverted normalised measure of information content. An equation of temporal specificity can be found in Equation 4:

$$\text{temporal specificity} = 1 - \frac{H(x)}{\log_2 N_t} \tag{4}$$

where $H(x)$ represents Shannon entropy of a normalised expression during the time course, and $N_t$ represents the length of the timecourse. Shannon entropy can be calculated as follows:

$$\text{Shannon entropy: } H(x) = -\sum_{i=1}^{n} p_i \log_2 p_i \tag{5}$$

where $p_i$ represent normalised expression level in a given time point while $n$ is the length of the timecourse.

### 3.2.7 Random forest prediction of tissue and temporal specificity

For each of the factor matrices computed with EFA, I performed two random forest predictions: One predicting adult tissue specificity, the other predicting developmental temporal specificity. Before running predictions, I performed random forest tuning of *mty* factor using tuneRF function from randomForest R package with parameters ntreeTry = 500, stepFactor = 1.5 and improve = 0.01. *mty* factor defines how many variables are going to be tested at each split of the decision tree. After defining optimal *mty* for each of the models, I ran random forest prediction using 500 trees with importance parameter set to TRUE. Later, I assessed the percentage of variance explained by each of the generated models. A factor matrix that yielded the highest cumulative percentage of variance from two random forest models was chosen for further analyses. Random forest models built from that factor matrix were combined together using combine function from randomForest package.

### 3.2.8 An index of gene expression complexity

From the final prediction model, I extracted the importance of each variable and used them to weight the initial EFA matrix. The sum of all weighted factors represents a factor of complexity of gene expression. Gene Ontology enrichment of biological processes analysis was performed on the group of 20 genes with the highest complexity factors. GO analysis was performed using the clusterProfiler package To complement results of the most complex genes, I did the same analysis on the set of 20 genes with the lowest complexity score. Annotations of GRB target and bystander genes were obtained from the previous group member, Ge Tan.

## 3.3 Results

### 3.3.1 The majority of ZFIN data describes organogenesis

To get comprehensive information about spatiotemporal gene expression during vertebrate development, I summarised all of the mRNA *in situ* hybridization data from the ZFIN database. ZFIN maintains a curated database of annotated mRNA *in situ* hybridisation figures from all research articles analysing zebrafish development. Their database contains annotations of expression localisation for 12,188 genes from assays reported in 5,618 research articles. After compiling gene expression data from all these articles, I have generated an expression table, that, for each gene, summarises temporal and spatial developmental expression in the embryo. A data point from an expression table is defined as information about a single gene in a single stage of development, expressed in a single anatomical structure. Basic statistics about the expression table can be found in Table 3.1 and Figure 3.2.

**Table 3.1:** Basic statistics from the resulting expression table.

| **Records in expression table** | 167,873 |
|---|---|
| Data points derived from all records | 291,296 |
| Genes whose expression was assayed | 12,188 |
| Publications incorporated | 5,618 |
| Anatomical terms | 1,078 |
| | |
| **The most frequently used anatomical terms** | |
| *whole organism* | 115,700 |
| *myotome* | 7,582 |
| *hindbrain* | 6,155 |
| Median of anatomical term appearance | 10 |

The expression table contains 291,296 data points. On average, each publication contributes 15 data points. The scopes of articles were not the same, so some studies just reported gene as being expressed in the whole organism while other studies reported all observed structures in detail. The most frequent anatomical descriptor used was "whole organism", and 115,700 data points are using it to define the expression. This term was mostly used by high-throughput genetic screens that interrogated the majority of genes in the database, but were not detailed enough to provide more precise spatial information of expression. The median number of occurrences for 1,078 anatomical terms in the database is 10, which means that the majority of anatomical terms are only used by a handful of publications.

The distribution of the expression table data points over the embryonic time course is shown in Figure 3.1A. Data points are distributed unevenly and most of the information comes from organogenesis, a developmental period when the majority of organs are formed. Very early stages are not frequently investigated due to technical difficulties of manipulating early embryos and because there is a limited number of anatomical structures to investigate. Research on zebrafish past the 33rd stage (Long-pec, 48 hpf) undergoes strict ethical regulations leading to scant publications dealing with late development. A detailed explanation of developmental stages can be found in Appendix Table A.1. Number of genes analysed in each of the developmental stages is still following the same trend like the total number of datapoints, but the difference between early stages and organogenesis is not as pronounced. In early developmental stages, researchers were consistently analysing 2,000-2,200 genes. Again, the developmental stage with the highest number of genes is Prim 5, a stage after which embryo is developing melanin in the skin and is not translucent anymore. Surprisingly, there are only 717 genes assayed in adult fish, which is three times less than early developmental stages. When looking at the distribution of anatomical structures used to describe expression (Figure 3.1C), there are on average 12 structures used in early developmental stages. This could be explained by the fact that very early embryos, before gastrulation, do not have many

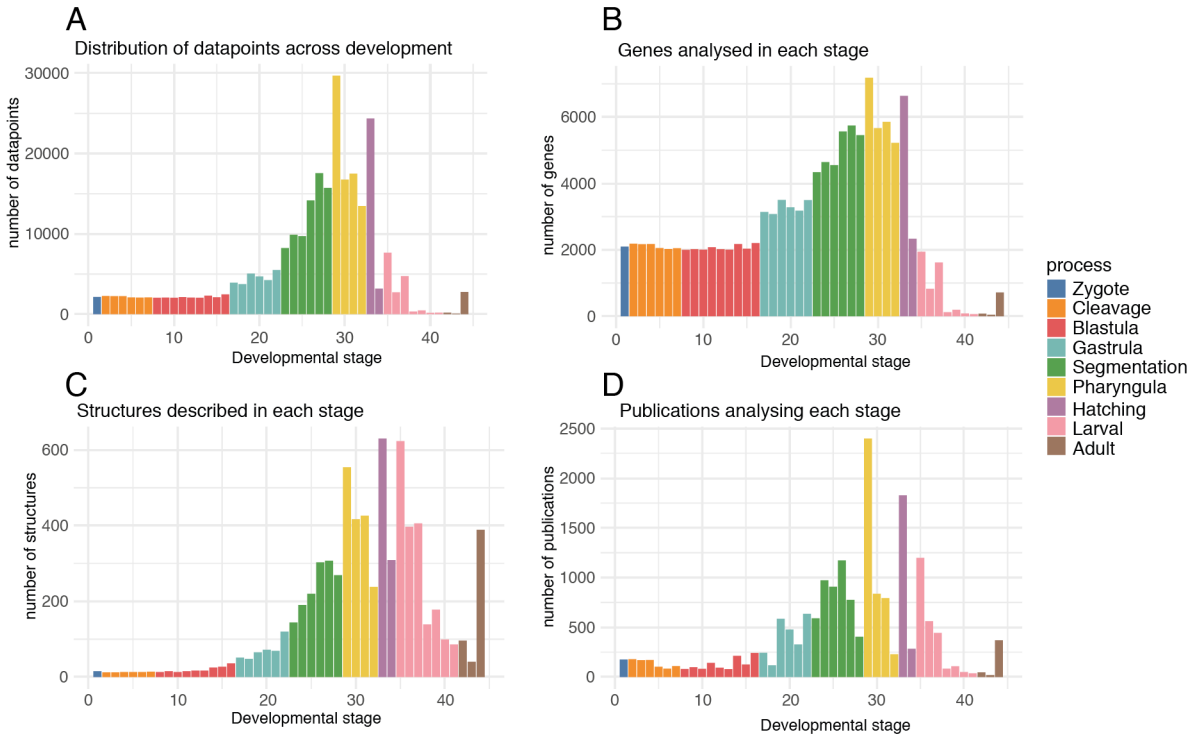


**Figure 3.1: Summary of the expression table created from ZFIN wildtype fish expression database after quality filtering.**
Every bar represents a standard ZFIN developmental stage for zebrafish, from zygote to adult fish. Bars are coloured according to the developmental process. (**A**) Distribution of total expression records in expression table. Zebrafish embryo is not transparent after 48 hpf, which here corresponds to the 33rd stage. There is a dramatic decrease in the number of datapoints just after 33rd stage. (**B**) Distribution of genes analysed in each of developmental stages. (**C**) Number of structures used to describe the location of gene expression for each stage. Before gastrulation, only a few anatomical structures are present. This causes very low values in the initial 15 stages of development. (**D**) Number of publications describing gene expression in embryos at a particular stage. There is a striking preference for the assays at stage 29 (Prim 5), 24 hpf.

formed structures. On the other hand, 389 structures were used to describe expression in adult fish. When we take into consideration that the expression table contains annotations for only 717 genes, that means that on average a new anatomical structure was used to annotate expression of every other gene. Despite the fact that there are not many data points describing expression from larval stages, these data points used disproportionately more anatomical structures than other developmental phases. Again, this could be explained by the fact that by that time embryo already developed all the major organs, so expression could be detected in many more structures. Finally, when I looked at the distribution of publications describing embryos across development, again, the majority of articles analysed embryos 24 and 48 hours post fertilisation (stages 29 and 33). Another interesting trend could be observed from this distribution. Researchers are more likely to analyse the first stage of a developmental process in comparison to the later stages after the segmentation phase. This can be seen as a disproportional quantity of publications analysing stages 29, 33 and 35. This trend could be identified in other distributions previously mentioned in Figure 3.1, but here is particularly evident. These results suggest that articles analysing later stage of a developmental phase are likely to analyse more genes and describe more structures than the ones analysing earlier periods of the same developmental phase.

After doing initial data mining on the expression table, some inconsistencies were found. One clear example was publication ZDB-PUB-041116-1 reporting expression of *tp73* gene in brain in *zygote*. It is clear that *zygote* does not have a brain, but to investigate systematically how often genes are annotated with structures that are not even developed in that developmental period, I created a controlled anatomical vocabulary. A controlled vocabulary of anatomical terms for each developmental stage tells which anatomical structures are present. I created it by extracting descriptions of anatomical structures from ZFIN's Anatomical Ontology system. Then, by using controlled vocabulary, I filtered out all data points that did not match anatomical terms with their developmental stage. For the publications reporting more than half of their annotations in the incorrect structures, I omitted

them from further analysis since I cannot be sure that the annotations of appropriate structures are indeed correct. Eight percent of all initial data points were reporting expression in anatomical structures that are not present in the corresponding anatomical stage. Additionally, due to the sparsity of data for the genes supported by fewer than five publications, those genes were discarded from future analysis. After filtering, I obtained detailed gene expression information for 919 genes making 70615 expression data points. By doing these filtering steps, we have reduced our dataset from 12188 to 919 genes This is a dramatic decrease, but since this dataset will be used to conclude about spatio-temporal dynamics of gene expression, only the genes with a substantial coverage of gene expression information are useful for devising confident conclusions.

Since the expression table was built by merging all relevant publications (5618 form Table 3.1), the cumulative information from all data points for gene expression for many genes was not continuous. Often, developmental time span of two publications, after merging their information, lacked a few developmental stages. If a gap in expression information is short enough so that gene expression could not be significantly different from the flanking stages, I could confidently impute expression information (see Materials and Methods, Imputation of missing data points). From the filtered set of genes, 138 genes had expression information covering complete period of their expression. Two thirds of genes contained gap in the expression information that was shorter than three developmental stages, so their expression was imputed in the missing stages. I imputed a total of four percent of expression data points, which enriched expression information in stages that were poorly covered. After quality filtering and imputation of datapoints in the expression table, it consisted of 73914 datapoints describing expression of 919 genes across zebrafish development.

### 3.3.2 A measure of gene expression complexity derived from anatomical expression data

The newly enhanced expression table uses 826 anatomical structures to annotate localisation of expression. To be able to compare spatio-temporal dynamics of gene expression, I need to be able to compare all anatomical structures with the aim to conclude which gene has more complex expression distribution and by how much. These comparisons in many cases will not be possible, since structures are not related and direct comparisons are impossible, or sufficient annotations in AO system about these structures do not exist. To reduce the number of anatomical structures used, I have established a coefficient that, using previously created anatomical hierarchies, informs how precise the localisation of each structure is, compared to all other anatomical structures belonging to that anatomical system. In addition, I have kept the information about the closest relatives of all structures, so that if I need to further reduce anatomical structures, I can assign specificity of the most similar structures. The most general term used in publications is "whole organism". Since this term does not contribute to spatial expression information, its coefficient is set to 0. Anatomical systems are the next most general terms, and their coefficient of specificity was set to 1. From anatomical descriptions provided by AO system, I created 20 hierarchies of anatomical structures, each for different anatomical system. Each hierarchy was describing the substructures of each anatomical structure belonging to that anatomical system. In these hierarchies, anatomical structures that were close to the root of hierarchy (e.g. central nervous system) are much more general than structures close to the leaves of hierarchies. From the position on hierarchies, I assigned a specificity coefficient to each structure, so that structures close to the root have very low values, whereas more derived structures have higher coefficient values. The largest hierarchy belongs to the neuronal system. It contains 42% of all anatomical terms and all structures are distributed on 18 different levels. Due to the uneven size of anatomical hierarchies, I normalised all specificity scores so that coefficients of each hierarchy are distributed in the range from 0 to 1.

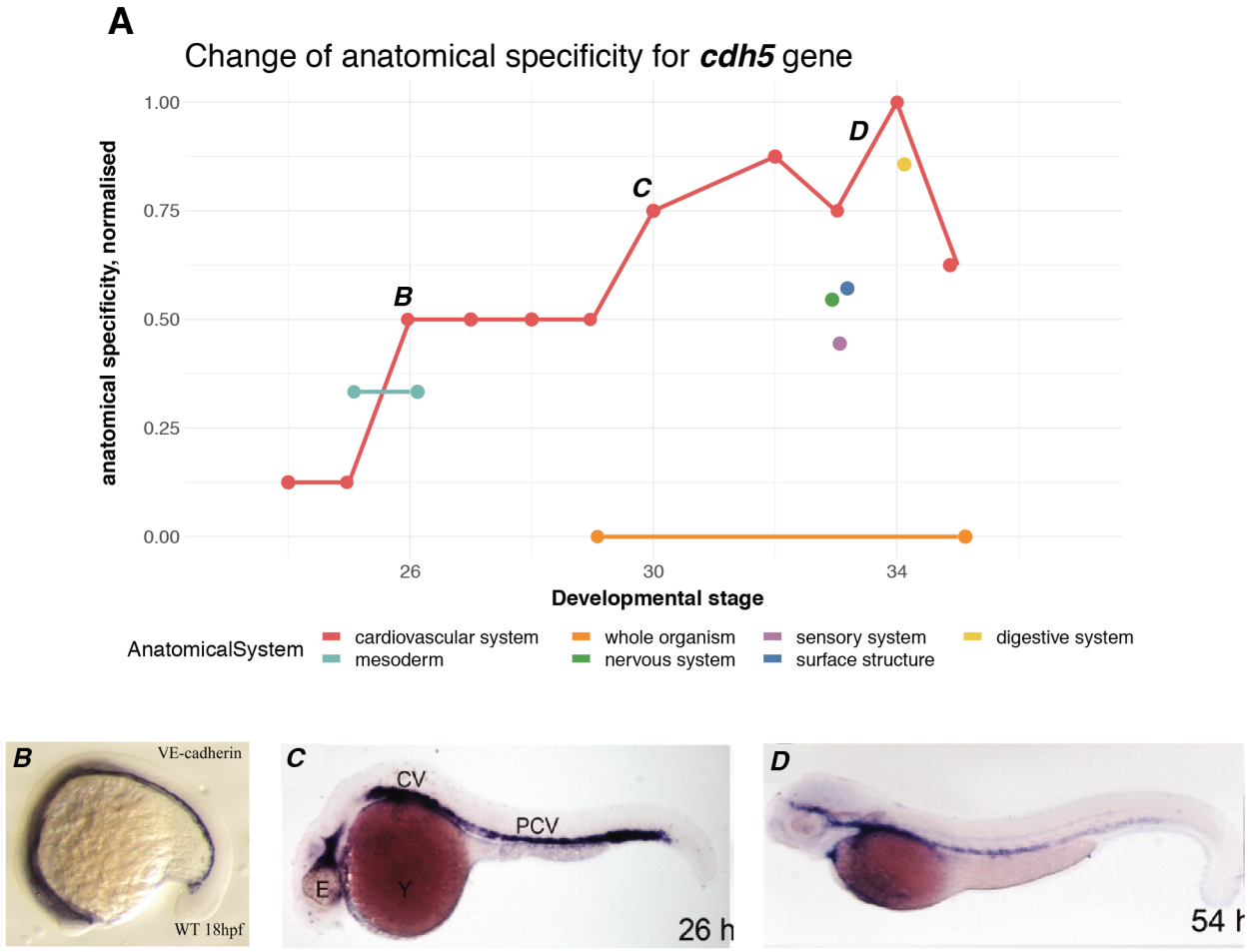


**Figure 3.2: Spatio-temporal expression of cadherin 5 (*cdh5*) gene.**
(**A**) Dynamics of anatomical specificity through development. cdh5 is active from stage 24 (Segmentation - 5 somites) until stage 33 (Hatching - Long pec). This gene is predominantly expressed in the cardiovascular system (red line) although some publications reported expression in other anatomical systems (noise). It shows a gradual increase in anatomical specificity through time, see also panels (**B-D**). (**B**) ISH of *cdh5* for an 18 hpf embryo from Jin et al., (**C**) and (**D**) ISH of *cdh5* from Flores et al. for embryos from 26 and 54 hpf respectively. The corresponding embryo stages are represented on the graph by the letter of a panel. ISH embryos show spatial reduction of signal across embryo that confirms increase in anatomical specificity.

The coefficient of anatomical specificity is used to interrogate gene expression changes regarding anatomical localisation throughout embryonic development. Figure 3.2A shows the change in anatomical specificity for cadherin 5 (*cdh5*). The expression table suggests that cadherin 5 expression begins at ZFIN stage 24 (Segmentation - 5 somites) and is observable until stage 33 (Hatching - Long pec). It has been reported that this gene mediates cellular contact in vascular endothelium, and that it is the main structural protein in junctions between endothelial cells (Lampugnani and Dejana 1997). It is not surprising to see that this gene is mostly expressed in the cardiovascular system, together with mesoderm in the early stages of its expression. Median anatomical specificity in cardiovascular system during the first stages of the expression of this gene is 0.125, which signifies that the gene is expressed in the majority of structures developed in segmentation. As the development progresses, median anatomical specificity of this gene increases, suggesting that expression of this gene becomes more specialised to better defined anatomical structures. To confirm changes in anatomical specificity from the Figure 3.2A, Figures 3.2B-D are presenting mRNA *in situ* hybridisation figures from ZFIN database. Figure 2B represents embryo in Segmentation:14-19 somites stage taken from (Jin et al. 2007). Gene expression is observable from head to tail, across the majority of the embryo. Figure 3.2C represents expression patterns in Pharyngula:Prim-15 embryo, taken from (Flores et al. 2010). At this stage, it looks like the expression is much more restricted to specific structures of the embryo. Anatomical specificity at this stage is also higher than it was in the previous figure. Finally, Figure 3.2C shows an embryo 54 hours post fertilisation. This embryo has even more restricted localisation of expression. Accordingly, the median value of anatomical specificity is also slightly higher than in a previous figure. In addition, it is worth noticing an example of noise on this graph from the publications that reported expression of this gene in whole organism in two stages (orange line at the bottom). There are also additional annotations of *cdh5* expression for a group of anatomical systems that can be found in stages 33 and 34. These reports are most likely noise since surrounding stages do not support this information.

Similarly to the example of *cdh5*, I have generated graphs of anatomical specificity for all 919 genes present in the expression table. These graphs can be used to distinguish housekeeping genes (i.e. genes expressed in whole organism at near constant level) from tissue-specific or developmentally regulated genes. Figure 3.3A shows an example of a housekeeping gene (*actb1*) that has consistently low values of anatomical specificity during its period of expression. Again, it can be observed that there are noisy reports of expression in the late stages of development. Namely, there are publications investigating expression only in the late stages of development which annotated expression in a few, specific, anatomical structures, although the majority of other publication state expression in the whole organism.

Another group of genes that can be identified using anatomical specificity dynamics are tissue-specific genes. Those are genes that are specifically expressed in one anatomical system or cell type, but can show dynamics in terms of anatomical specificity during the period of their expression. Figure 3.3B represents a tissue-specific gene *phox2a* that is specifically expressed in the nervous system. It's activity was identified at the end of gastrulation and beginning of organogenesis and is observable until the end of organogenesis. During this period, changes in localisation of expression were observed, but its expression was confined within the cells constituting nervous system.

In contrast to the previous two panels of the Figure 3.3, panel 3.3C represents anatomical specificity dynamics for *sox2*, a developmental gene. Gene *sox2* encodes for a transcription factor that has an essential role in the maintenance of pluripotency of non-differentiated stem cells. Due to its instrumental role in development, its expression is expected to be detected in many embryonic structures in which this TF regulates developmental processes. Its anatomical specificity confirms this hypothesis, showing how this gene is expressed in many anatomical systems at different specificities with a highly dynamic pattern. Particularly interesting observation is, while looking at a single stage, gene in different anatomical systems at the same time can be expressed at different levels of specificity. In

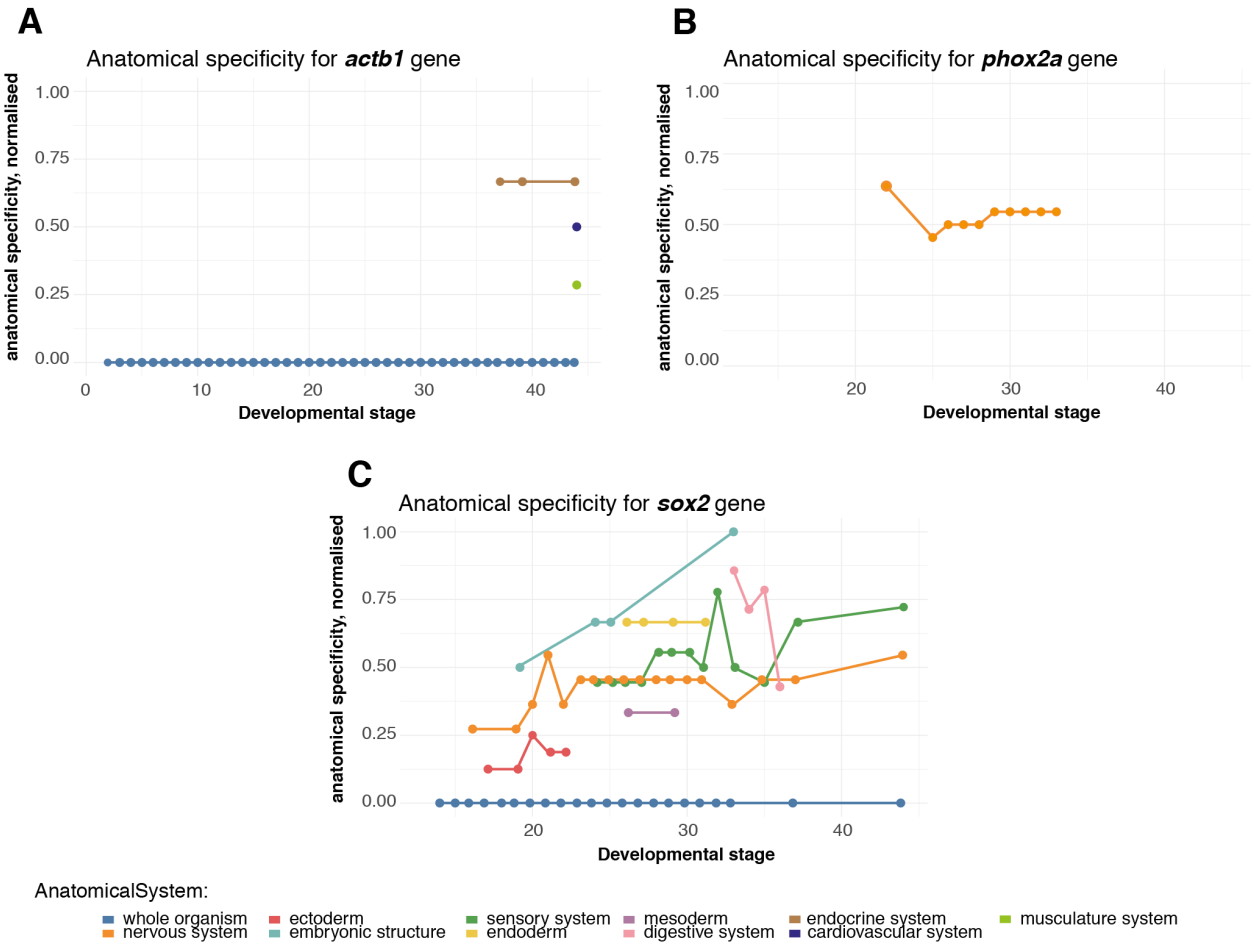


**Figure 3.3: Anatomical specificity dynamics for different gene groups.**
(**A**) Housekeeping gene actin b (*actb*) shows a steady profile of anatomical specificity throughout development, with some noise in the adult stages due to publications annotating only a few very specific structures. (**B**) Tissue specific gene *phox2a* is expressed in neuronal system at various levels of specificity. (**C**) Developmentally important gene *sox2* is expressed in numerous anatomical systems with a very dynamic profile of anatomical specificity. All graphs represent median value of specificity for anatomical system at each stage of development.

some, it is more generally expressed, while in others its expression is more positionally restricted. This observation suggests that *sox2* has regulatory mechanisms able to drive the expression of different levels of anatomical specificity in different anatomical systems.

### 3.3.3 Gene clustering based on anatomical specificity

Although anatomical specificity is able to discern between major gene classes, I wanted to test if it is able to further sub-classify genes based on their spatiotemporal gene expression pattern and see if these genes share other common features. In general, clustering is done by assigning genes to one of the clusters, based on the similarity measure. These methods are also called hard clustering methods, since a gene can only belong only to a single cluster. On the other hand, fuzzy clustering methods allow genes to be clustered in multiple clusters and their membership is reported for each of the clusters. Membership coefficient indicates to which extent does a gene belong to a cluster. Reporting membership coefficient is particularly important in cases when a gene is as similar to multiple clusters. By using fuzzy clustering, I should be able to assign genes to all clusters in which they are above a defined membership threshold. More importantly, genes that are not similar to any of the clusters will not be force assigned to one of the clusters, but rather omitted from clustering. Doing that gives more information into what are the specific features of gene clusters. I performed fuzzy clustering on expression table by using WGCNA (Langfelder and Horvath 2008; Langfelder, Zhang, and Horvath 2008). To reduce the size of the whole expression table and to make it less sparse, I have grouped all anatomical terms used to describe the expression into their corresponding anatomical systems and the value used, as a representation of the whole system, was the sum of all anatomical specificity coefficients reported. This way I created 919 x 20 matrix, where each row corresponded to a gene, and columns sum of anatomical specificity for each of anatomical systems. The first step of WGCNA algorithm is to calculate Pearson correlations for all pairs of genes in the network. Since gene annotations are noisy due to incomplete annotations, I weighted all correlations by

taking the absolute value of the correlation and raising it to the power of $\beta$. This proces will reduce the contribution of lowly correlated genes in network generation, while at the same time emphasizing strongly correlated genes. To define the coefficient $\beta$, WGCNA utilises the fact that gene expression networks exhibit an approximate scale-free topology (Barabasi and Albert 1999). I have tested a range of $\beta$ powers, from 1 to 20, and for each of them I ran network topology analysis. $\beta$ was chosen as an inflection point of calculated scale-free topology fit indexes as a function of the tested $\beta$ values. Optimal $\beta$ value was determined to be 12. Statistics of scale-free topology and the mean connectivity of the network can be found in the Figure A.1. Finally, obtained weighted correlations are corresponding to the connection strengths between genes in the constructed network.

After creating a network with WGCNA, I obtained 6 modules of coexpressed genes and a group of unclustered genes. Dendrogram of gene dissimilarities and their assignment to modules can be seen in the Figure 3.4A. Modules were defined by pruning dendrogram branches. WGCNA uses dynamic branch pruning methods for module identification from a dendrogram depending on their shape (Langfelder, Zhang, and Horvath 2008). This is the reason why modules are discontinuous in the module assignment ribbon. The largest module is the red module with 438 genes, whereas the yellow module is the smallest with only 31 genes. There were 86 unclustered genes. To test if our network modules that were identified by combining all anatomical specificity scores have unique gene functions, I ran GO analysis for each of the modules. All genes in current expression table were used as the background gene set, and genes belonging to each of the modules as the target gene set. The results of GO enrichment analysis for WGCNA modules are presented in the Figure 3.4 C-H. Figure 3.4C presents GO enrichment results for the red module, which is strongly enriched for nervous system development functions. This results is expected since, historically, zebrafish is a model organism for development and neuronal system function and most publications will be looking at genes involved in these process. GO enrichment for the second largest module (cyan) suggests that this group of genes is involved in immune processes and responses to

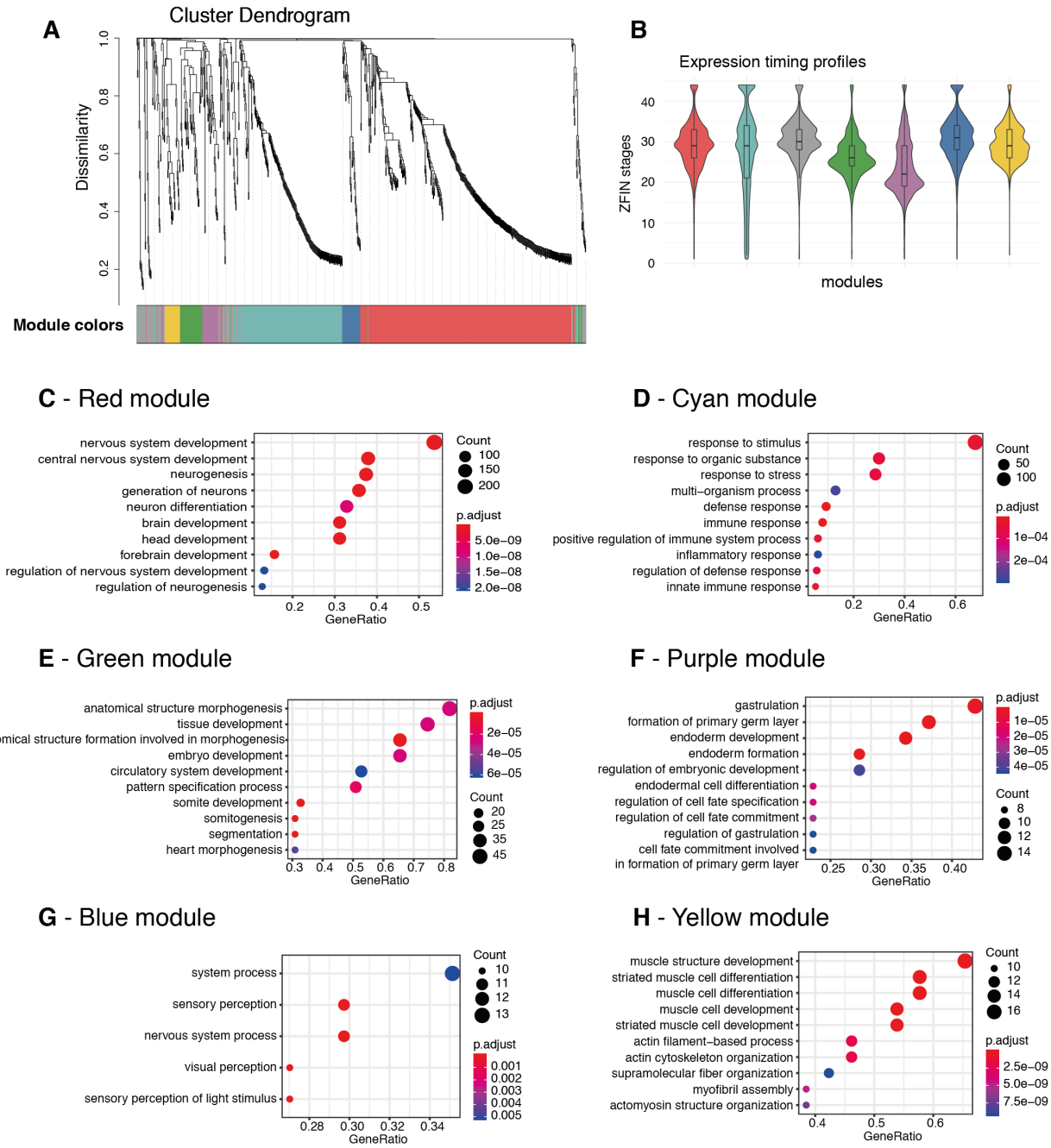


**Figure 3.4: WGCNA analysis on the expression table.**
(**A**) A dendrogram of TOM dissimilarity in which each branch represents a gene. At the bottom of the dendrogram there is a ribbon with module assignments. (**B**) Expression timing profiles of identified WGCNA clusters. (**C**) - (**H**) GO enrichments of biological processes for identified gene modules.

stimuli (Figure 3.4D). This is an intriguing result since our annotation doesn't include any anatomical structures belonging to the immune system. One of the reasons for AO system to not include immune system structures is the fact that immune system develops after the first five days of development, a period in development where number of studies analysing development drastically reduces. Green and purple module are enriched for tissue development and gastrulation respectively. Two smallest clusters, blue and yellow, are specifically enriched for processes in developing sensory and musculature system.

To further analyse the enrichment of Cyan module, I tested if genes belonging to Cyan module are active later in development in comparison to other modules. If this was the correct observation, GO enrichments could be explained by the fact that these genes specialise later in development than most expression records from expression table. Figure 3.4B presents expression timing profiles for WGCNA modules. It is evident that purple module that is enriched in gastrulation processes peaks first. Cyan module, on the other hand, seems like is expressed throughout with minimal enrichments during the organogenesis. This observation could suggest that these genes are involved in housekeeping functions. The increase in the expression in the adult fish might be contributing to the most of the immune system GO terms.

WGCNA analysis clearly showed that genes with similar patterns of anatomical specificity take part in similar biological processes. In addition, this analysis showed how, although this dataset is incomplete, with the current amount of annotations I am able to obtain distinct functional gene groups which suggests that these annotations contain enough information to functionally discriminate genes with different expression patterns.

In this analysis, I have combined all developmental stages into one representative coefficient, and even with this, I obtained modules of genes expressed in distinct developmental periods. Combining all stages into one coefficient reduces detailed information of dynamics on a stage basis. To overcome this challenge and use entire expression table information that

is a very sparse matrix, I next did an Exploratory Factor Analysis to reduce the number of descriptors used to describe gene expression.

### 3.3.4 Exploratory Factor Analysis

The expression table provides a detailed view of the developmental localisation of gene expression. In order to conclude about spatio-temporal dynamics of gene expression, it is necessary to reduce the current use of 826 anatomical terms across 44 developmental stages into a smaller set of factors that will capture the maximal amount of information. When combined, the information from each developmental stage and anatomical term contains 3724 variables that explain expression. Since a matrix with 3714 columns would be sparse, I have reduced this matrix. To do that, I performed EFA. EFA identifies latent variables by identifying similar variables and grouping them together into a single factor. This way, I was able to reduce the number of variables used to describe expression with a minimal loss of information provided from the original variables. EFA is able to eliminate multicollienar variables by applying different rotation methods that ensure that the new factors are orthogonal. EFA was used in this analysis because when variables don't have common features, EFA will not report underlying factors with high support. Some other dimensionality reduction methods, like PCA, in that case, will report principal components that explain the maximal amount of variance in the data. This could be misleading since instead of reporting an error-free factor, they report a well-defined component that models all variance in the underlying data.

Factor analysis initiates the analysis with the assumption that there are as many factors as the provided variables. In the first step, a matrix with the same number of variables is created, but these new variables represent linear combinations of the original variables with some weight coefficients. This means that information from the original matrix in this step is transformed and not reduced or lost. A common way to do a transformation is to calculate eigenvalues and eigenvectors. Next, data is transformed in the direction of calculated eigenvectors and all new factors are represented by using eigenvalues. A factor

explains more variance than the original variable if its eigenvalue is higher than 1.

To create a reduced matrix of *in situ* annotations, I have removed all structures that had fewer than 5 reports of expression in the whole dataset. From there, the correlation between each pair of variables is computed using all complete pairs of observations on those variables. Eigenvalues of factors from the expression table are sorted from the highest to the lowest value on the Figure 3.5A. I created an EFA model including all factors whose eigenvalue was higher than 1. This model included 585 factors. In addition, I have created ten additional, smaller models to test if the reduction in the number of informative factors reduces prediction power of our dataset. These models were built by using 4, 12, 18, 32, 54, 98, 131, 256, 378, 425 factors respectively. Factor loadings for the model with 131 factors can be seen on the Figure 3.5C. In order to get information about how much does each factor contribute to gene's expression information, I have extracted factor values by using tenBerge method.

In addition to dimensionality reduction, factor analysis can be used to construct novel coefficients. The most common way to construct a coefficient is to sum all factors. A problem of this method is that not all covariates that make up the coefficient have the same explanatory power, and therefore adding them all up results in bias. To discover the explanatory power of calculated factor matrices, I ran random forest prediction where factor matrices were used to predict expression pattern extracted from RNA-seq datasets. With the importance measure extracted from random forests and factor values used to describe gene expression, I created gene expression complexity measure.

### 3.3.5 Random Forest prediction of RNA-seq extracted features

To get novel insights about spatio-temporal dynamics of gene expression during development, I used annotations of spatiotemporal expression extracted from ZFIN to predict expression patterns extracted from RNA-seq data. This way, by doing predictions using ZFIN annotations, I can investigate what are features that are important to explain RNA-seq

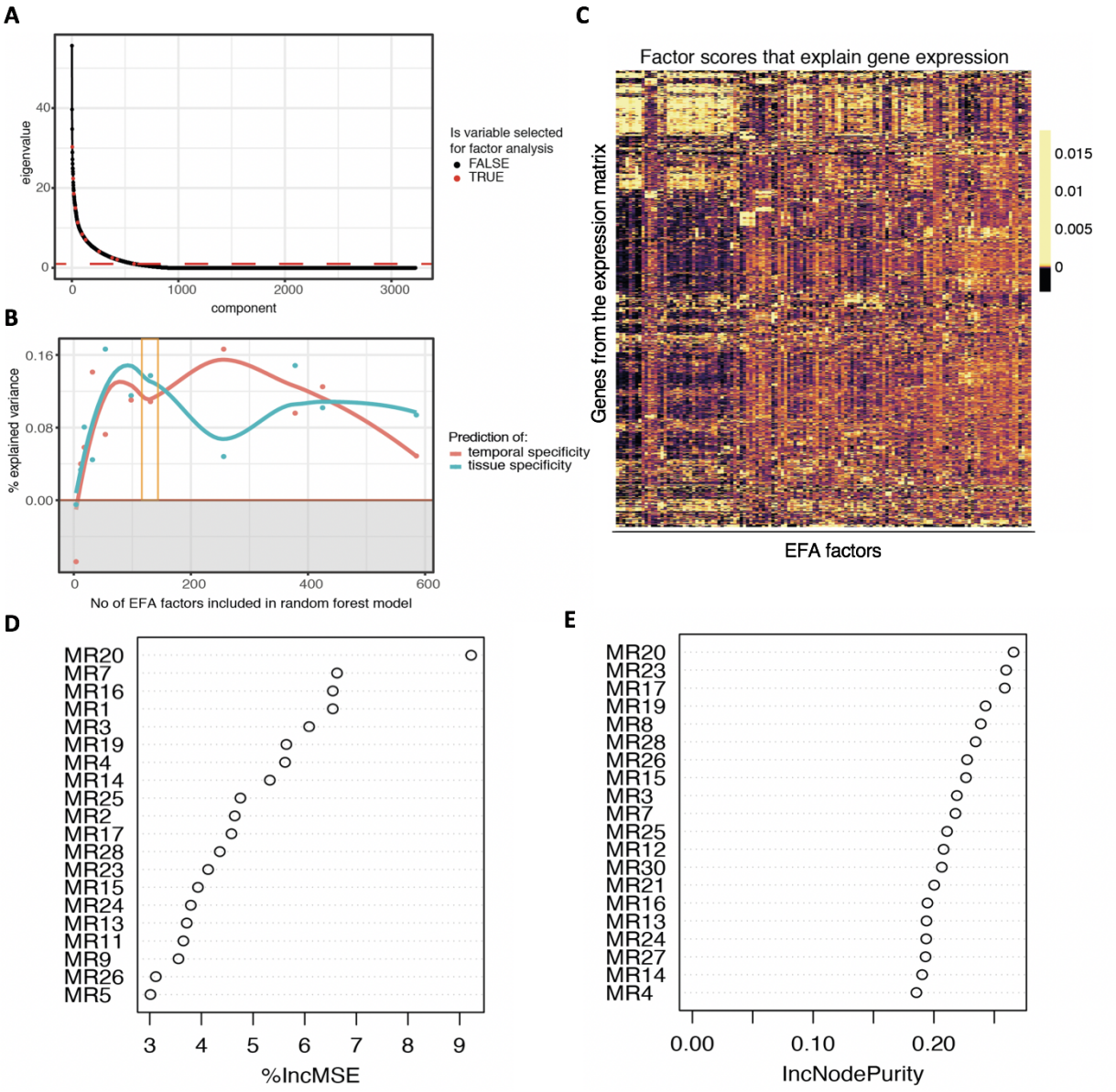


**Figure 3.5: Results of exploratory factor analysis (EFA) and random forest (RF) predictions.**
(**A**) Eigenvalues for principal components calculated from pair-wise correlation matrix. Eigen values are sorted in descending order, from the most informative to the least informative ones. Red line denotes eigenvalue of 1. (**B**) Percentage of explained variance for RF predictions using matrices with different number of factors in EFAs. For each number of factors included into prediction model, two predictions were performed, one predicting temporal specificity, another predicting tissue specificity. Orange highlighted area shows a factor (131) at which the cumulative explained variance is the highest. (**C**) Factor loadings explaining gene expression annotations from ZFIN. Rows of this matrix represent genes and columns 131 factors extracted from EFA. Colour scale is distributed in even proportions of factor loading values from black (the lowest values) to yellow (the highest values). (**D**) And (**E**) variable importance from combined RF model. **D** represents percentage increase in mean squared error that a factor contributes to, and **E** shows increase in node purity.

pattern of expression. In addition, based on the predictors used to explain the expression of a gene, I am able to conclude on the complexity of spatio-temporal expression. Unfortunately, ZFIN dataset is incomplete since it is created by collating currently available research articles. For many genes, only limited expression information is available. In addition, annotations of expression provide a binary readout of gene expression.

Extracted factor values were then used to explain tissue and temporal specificity of annotated genes. Tissue specificity was calculated from (Hu et al. 2015), while temporal specificity was calculated from (White et al. 2017). White et al. study measured gene expression during the development using RNA-seq, but since developmental samples were created from the whole embryo, they don't have the power to define a spatial component of observed expression. Although we do not have RNA-seq samples from embryo tissues, to approximate tissue-specificity, I used adult tissue-specificity in this analysis, since genes that are expressed in multiple tissues in the adult fish will most likely have been expressed in multiple tissues in development.

With the created reduced factor matrices, I have run random forest predictions of tissue specificity and temporal specificity. Random forests are able to run regression to predict only a single variable. For this reason, I have trained separate random forest models that predicted tissue specificity and temporal specificity. Finally, these two sets of random forest models were combined to get a common, larger, random forest model. Before running random forest analysis, I tuned the *mtry* parameter needed to run the algorithm. *mtry* defines how many variables will be randomly sampled and used as a condition at each tree split. For all analysis, I used random forest with 500 trees created. In all the models built, the accuracy of the model stabilised sooner than 500 trees, so there was no need to increase this parameter. While training a random forest model, the mean prediction error is calculated. It is calculated by taking a mean of prediction error for each sample (in my case gene), using only the trees that did not use this gene in their bootstrap sample. Figure 3.5B

shows the accuracy of trained random forests with a variable set of initial factors extracted from ZFIN database. In the case when only four most informative factors were used, the percentage of explained variance was negative for both prediction models. Negative explained variance means that the created model has no predictive power. Unexplained variance is attributed to the true random behaviour or lack of model fit. With the increase of a number of factors included in the RF model, the proportion of explained variance increases for both types of prediction. In the models predicting tissue specificity, variance explained by these models starts decreasing after the model with 54 factors and stabilises by 425 factor model. In the case of models predicting temporal specificity, the proportion of explained variance fluctuates among different models. Since these two types of prediction are ultimately going to be combined together into one model, I choose to use the model in which cumulative accuracy was the highest. In this case, the model with 131 factors had the highest combined accuracy. Random forests can be used to rank the importance of variables in a regression. Figures 3.5D and 3.5E show the variable importance parameters for the combined model of a factor matrix with 131 factors.

### 3.3.6 Coefficient of gene expression complexity

Gene expression is regulated in a number of ways, from epigenetic modifications on the DNA and histones to RNA transcription and binding of TFs to specific regulatory sequences. Transcription itself requires the precise function of many proteins and RNAs. All these elements allow a gene to be specifically expressed through time and space. In this chapter, I define gene expression complexity as a concept that is related to the number of elements which are required to drive gene expression precisely. A consequence of numerous regulatory elements is an ability for a gene to finely regulate its expression across space and time. In this analysis, I have used the information of spatiotemporal gene expression as an estimate of gene expression complexity. By using ZFIN's *in situ* hybridisation expression information to predict tissue and temporal specificity of expression, I have incorporated both

aspects of expression diversity, its quantity and spatio-temporal distribution.

The importance parameters extracted from random forest models were used to weight factor matrix in order to create a novel coefficient of gene expression complexity. Figure 3.6A presents 20 genes having the highest complexity coefficient out or all genes described by the feature matrix. On the other hand, Figure 3.6B presents genes with the lowest score of gene expression complexity. Many of the genes with the highest complexity score are developmentally important genes. This observation is not surprising considering that developmental genes are reported to have the highest diversity in their expression repertoire. The majority of genes with the lowest scores of expression complexity are housekeeping and tissue specific genes. Figure 3.6C represents three examples of genes with high expression complexities. It is noticeable that these genes are expressed in multiple anatomical systems and at different levels of anatomical specificity. In addition, when looking at anatomical specificity of expression within a single anatomical system, dynamic behaviour can be seen. A gene can be expressed very generally in early development and, as the time progresses, be expressed in more specific and better defined anatomical structures. When looking at a single developmental stage, these genes are expressed more specifically in some anatomical systems, while in other systems their expression was widespread. In contrast to genes in the Figure 3.6C, genes in Figure 3.6D are genes with some of the lowest complexity coefficient. These genes belong to the group of genes who are expressed in the whole organism throughout development, or along with the "whole organism" annotations, have a low number of additional annotations. In this case, like in the case of gene *tnfb*, additional more specific annotations in the adult stage (stage 44) are due to a publication that was looking very specifically into the expression of this gene in the digestive system. Since they were analysing only specific structures, their annotations yielded a higher score of anatomical specificity. However, despite these additional annotations, random forest model scored these genes as low complexity.

To confirm that genes with the highest gene complexity are involved in development,

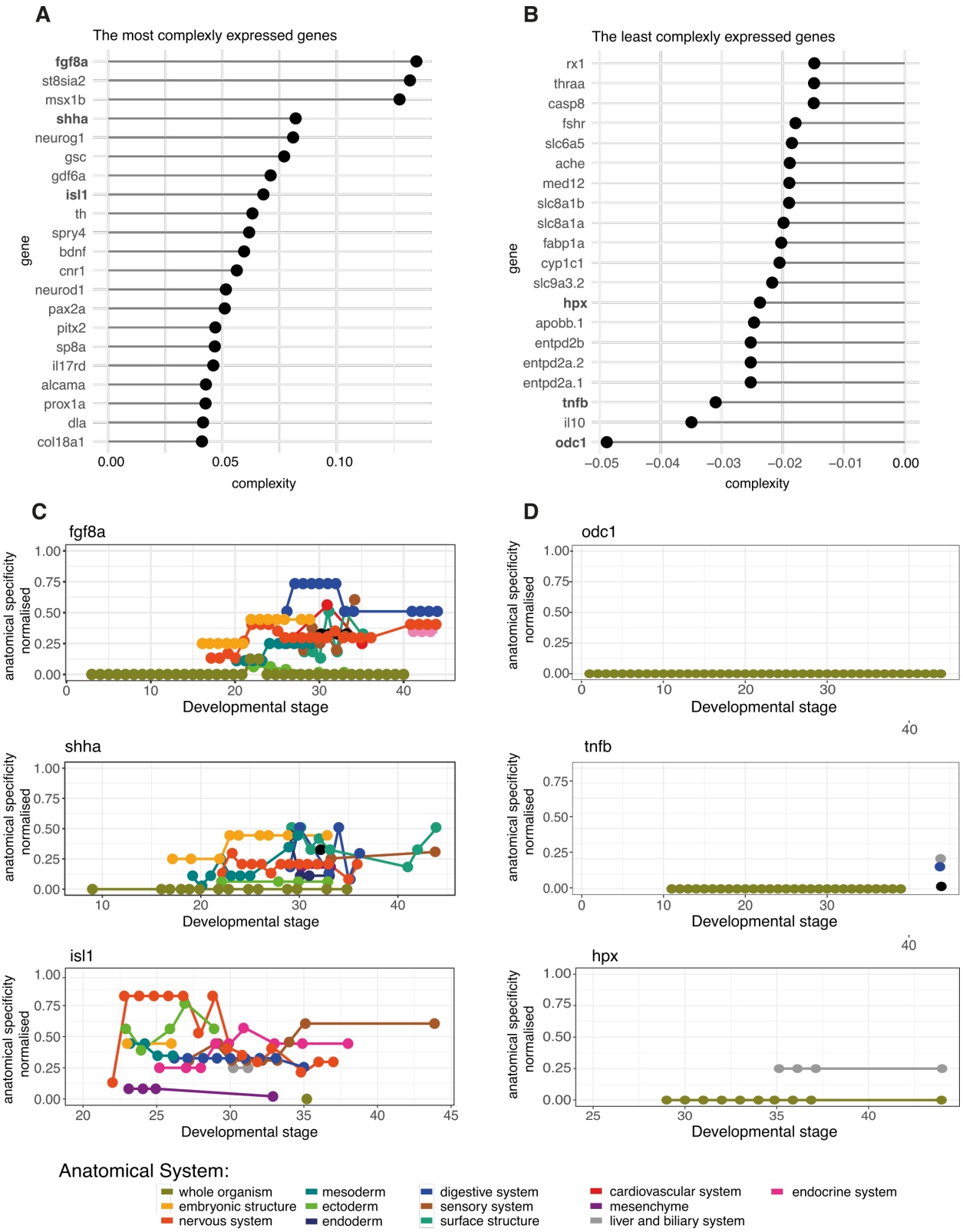


**Figure 3.6: Characteristics of genes having the highest and lowest complexity from the RF model.**

(**A**) Twenty genes with the highest complexity coefficient. Bolded genes are the ones for which their anatomical specificity is displayed in panel C.



*(continued)*

**Figure 3.6:** (**B**) Twenty genes with the lowest complexity coefficient. Bolded genes are the ones for which their anatomical specificity is displayed in panel D. (**C**) Anatomical specificity dynamics for genes *fgf8a*, *shha* and *isl1*. (**D**) Anatomical specificity dynamics for genes *odc1*, *tnfb* and *hpx*.

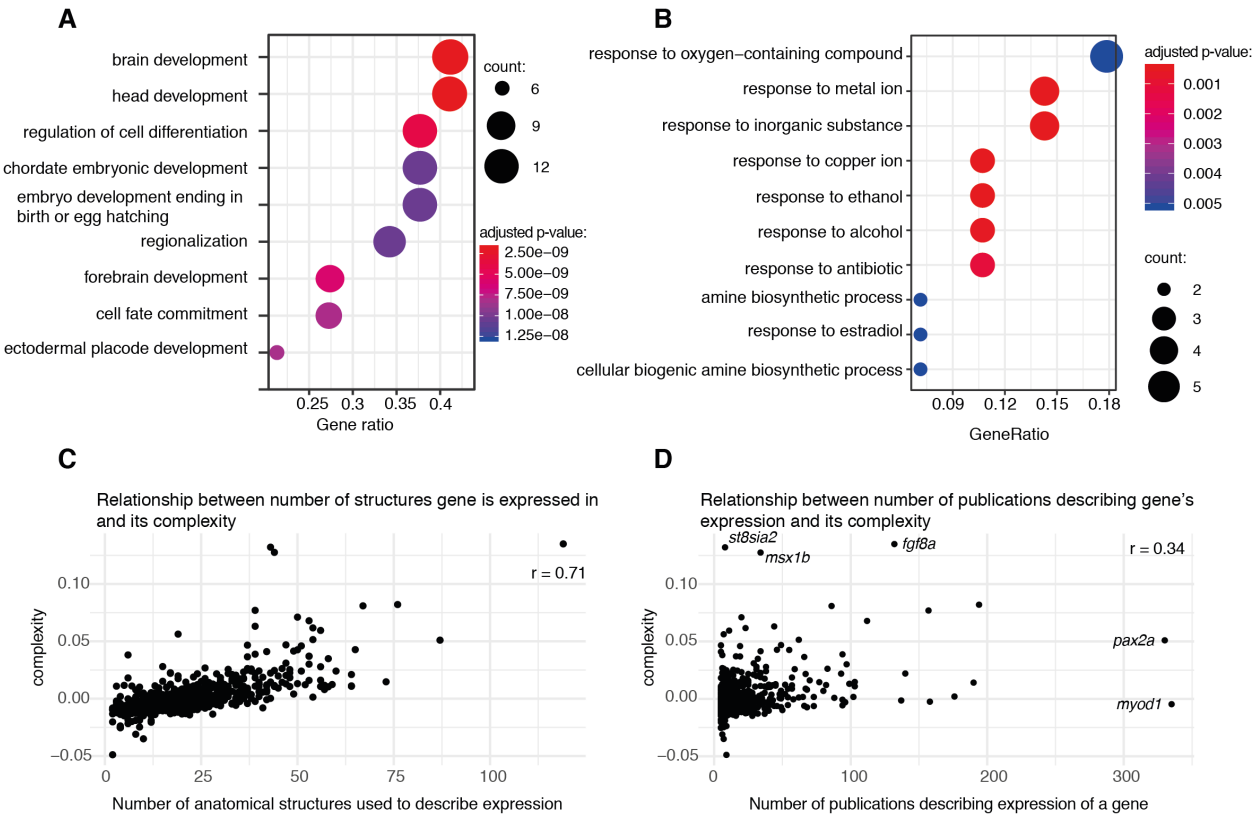


**Figure 3.7: Additional gene annotations for genes from RF predictions.**
(**A**) GO enrichment of biological processes for 20 genes with the highest complexity score (**B**) GO enrichment of biological processes for 20 genes with the lowest complexity score (**C**) Correlation of gene expression complexity score and number of anatomical structures used to describe the expression. Pearson correlation coefficient is shown in the top right corner (**D**) Correlation of gene expression complexity score and the number of publications analysing that gene. Pearson correlation coefficient is shown in the top right corner

I ran GO analysis of 20 most complex genes in this dataset, results shown in Figure 3.7A. In this analysis, as a background, I used all genes for which we have calculated complexity. Indeed, these genes are involved in embryonic development, with many early developmental processes like regionalisation and regulation of cell differentiation among the most significant ones. In addition, I have investigated the function of the genes with the lowest complexity

score. Again, I ran GO analysis of the 20 least complex genes while using all genes in the dataset as a background. Figure 3.7B shows results of GO analysis for this sample. The majority of significantly enriched terms are involved in the response to a chemical. It is surprising to see that these genes are not enriched for true housekeeping functions like replication, translation, and many metabolic processes. However, this can be explained by the fact that most genes involved in those functions are not supported by more than 5 *in-situ* hybridisation studies and are discarded from downstream complexity analyses. In the current list of genes with the lowest complexity coefficient value, there are groups of transmembrane protein genes (*slc* and *entpd* family) that have receptors for various chemicals. This caused GO enrichment to have "response to chemical" terms as enriched.

A potential problem of my coefficient of gene expression complexity is that, due to the differences in the number of data points annotating expression per gene, this model could be overfitting genes with many annotations and falsely making them more complex than in reality. Genes assayed in many publications will have a higher diversity of structures annotated which could, in turn, cause these genes to be more complexly scored than in reality. Random forests as ensemble methods that fit the prediction model many times and chose a combination of models that overfit to a lesser extent than other prediction models. To investigate if this is indeed an issue with my models, Figure 3.7D shows the relationship between the number of publications describing the expression of a gene and its calculated complexity of expression. There is a weak Pearson correlation between these two variables r=0.34. Gene *myod* is annotated by more than 300 publications, and in turn, its expression complexity is around zero. This is expected since this gene regulates muscle cell differentiation and regeneration and it is primarily expressed in the musculature system. On the other hand, gene *max1b* is annotated by fewer than 50 publications and is among genes with the highest complexity score. Although this gene is not analysed extensively, it has been shown that it is involved in processes like: embryonic organ morphogenesis, fin regeneration, and regulation of neuron apoptotic processes which suggest that this gene, indeed, has a complex expression

pattern.

Another interesting relationship is between the number of unique anatomical structures used to annotate gene expression and gene complexity shown on the Figure 3.7C. In this case, Pearson correlation is much stronger (r = 0.71) and very few genes don't follow a general trend of increase of expression complexity with the increase in the number of unique structures used. This relationship is expected since the more structures gene is expressed in, more likely it is to have different ways of regulating expression in different body parts, and have more complex gene expression.

Finally, after obtaining the list of genes with the highest gene expression complexity, and showing how these genes are indeed involved in developmental processes (Figure 3.7A), I explored their promoter structure. I hypothesised that genes with complex expression profile should have more diverse promoter profiles and that these genes will relay on long-range regulation of gene expression. In order to identify a significantly enriched promoter feature in this group of genes, promoter feature should have a significantly different frequency of given feature in comparison to the background set of genes. However, a problem of this dataset as a whole is that, because it is based on the set of genes studied by at least five research articles, it is strongly enriched for genes that have many promoter features. This observation can be seen on a figure Figure 3.8A that shows how 919 genes in this dataset, when compared to all genes active in the development, are significantly enriched for TATA-box, GC-box, Initiator, MED-1 CCAAT-box motifs and for bidirectional promoters. Because of this strong bias towards genes with all of the mentioned promoter features, It will be challenging to identify promoter structure enrichment. Figure 3.8C presents Promoter Ontology results for the 20 genes with the highest gene expression complexity coefficient. None of the promoter structures analysed were not significantly enriched after adjusting p-values for multiple testing. The same result is observed with the set of the 20 genes with the lowest expression complexity coefficient (Figure 3.8D).

To test if the genes with the high gene expression complexity use long-range regulation to drive precise expression, I tested if these genes belong to genomic regulatory block (GRB) target genes. GRBs are genomic regions spanned by HCNEs that are used as regulatory inputs for a target gene in the GRB region (Akalin et al. 2009). The target genes are usually transcription factors involved in the regulation of embryonic development and neuronal processes. Along with target genes, GRBs can harbour other genes that are not sensitive to the regulatory inputs from HCNEs. These genes are called bystander genes. Figure 3.8B shows how with the increase in the gene expression complexity, the proportion of GRB target genes increases, while the proportion of bystander genes remains stable. This observation suggests how elevated gene expression complexity coefficient can indicate that gene possesses elaborate gene regulation, or that it is intricately expressed throughout an organism.

## 3.4 Discussion

In this chapter, I have established a measure of spatio-temporal gene expression complexity in zebrafish embryonic development. I define spatio-temporal gene expression complexity as the ability for a gene to be specifically expressed in a limited proportion of cells at the specific level of activity. To derive a measure of complexity, at least three pieces of information are required: location, the developmental period and levels of the observed expression. Until recently, methods that could extract all three information did not exist. To overcome this problem, I have combined gene expression data from multiple assays.

First, I used mRNA *in situ* hybridisation assays to learn about the spatial distribution of gene expression in development. Here, I summarised expression data for the wild-type fish from ZFIN. This database contains annotated mRNA *in situ* hybridisation images from 5618 publications assaying expression of 12188 genes. Although this is a fantastic resource to study zebrafish gene expression localisation, the fact that it has been collated from many published studies makes this resource vulnerable to batch effects. Different groups can perform hybridisation protocol differently and cause expression reports to disagree. To overcome

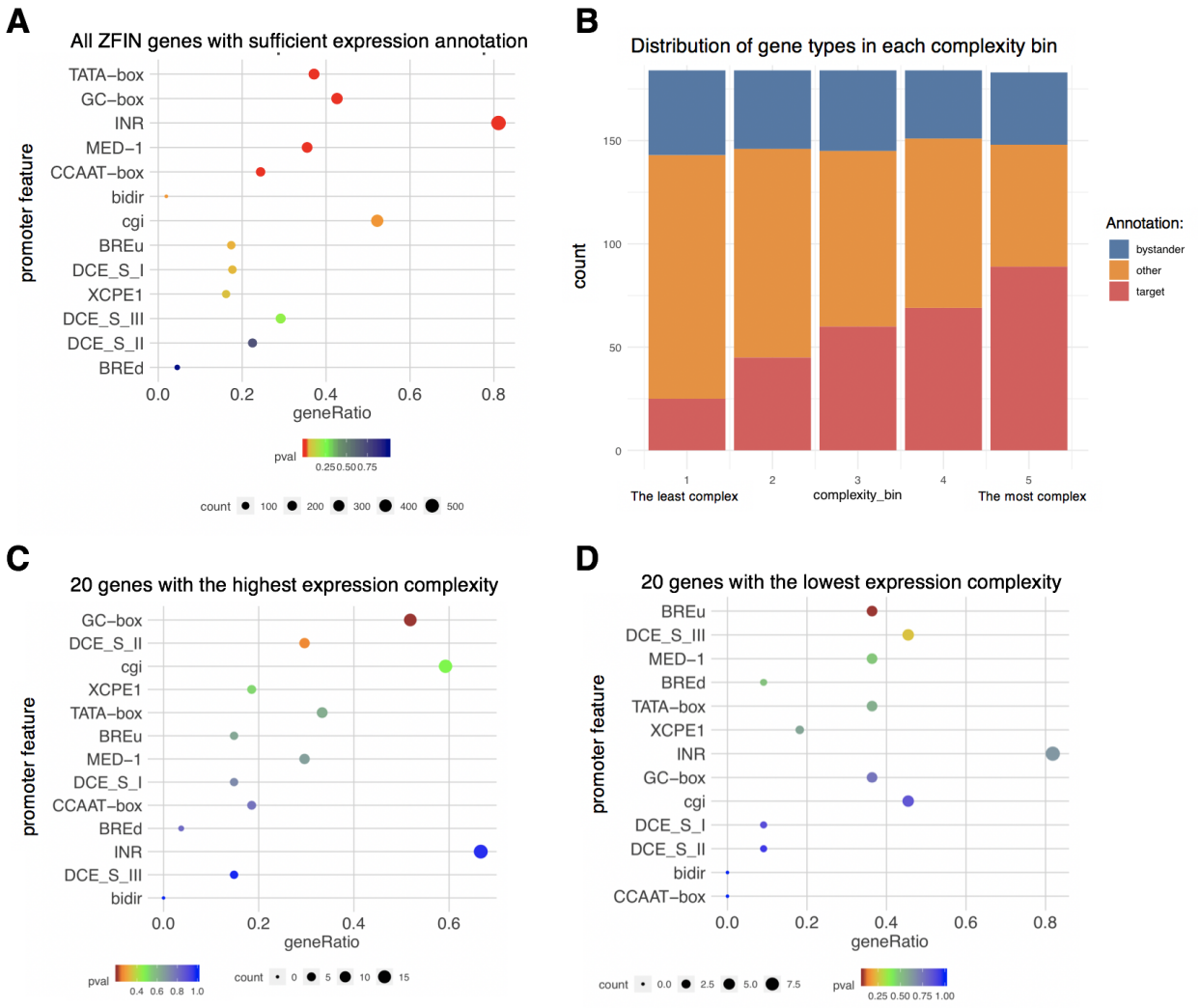


**Figure 3.8: Additional gene annotations for genes from RF predictions.**
(**A**) Promoter feature enrichment for all ZFIN genes analysed by at least 5 publications. The background in this analysis were all genes active in Prim 5 developmental stage. (**B**) Distribution of genes, split into five bins of gene expression complexity, belonging to GRB target and bystander annotation. The first bin consists of genes with the lowest gene expression, while fifth bin consists of gene with the highest gene expression complexity. (**C**) Promoter feature enrichment for the 20 genes with the highest gene expression complexity. The background in this analysis were all genes from ZFIN with the annotations from at least 5 publications. (**D**) Promoter feature enrichment for the 20 genes with the lowest gene expression complexity. The background in this analysis were all genes from ZFIN with the annotations from at least 5 publications.

this problem, I used the Zebrafish Anatomical Ontology system to discard reports that used anatomical structures not present in the corresponding developmental stage. In addition, I analysed gene expression information for genes that were assayed by at least 5 publications. This way I ensured that each gene has sufficient localisation information.

I propose a novel measure of anatomical specificity that, for each anatomical structure, defines how precisely its location is defined. Structures encompassing entire anatomical systems will have very low anatomical specificity since expression can be observed in a great proportion of the organism. On the other hand, specific cell types or single cells will have the highest score of anatomical specificity. By using anatomical specificity measure, one can learn about changes in gene expression localisation across time, and even compare changes in expression localisation within an anatomical system. Anatomical specificity showed how genes with the most dynamic profile of anatomical specificity in development are developmentally important genes. This observation is in the agreement with the results from Tomancak et al. have shown how transcription factors and signalling molecules are gene groups with restricted localisation of expression and that these genes are often found active in different organ systems.

After summarising gene expression localisation, I explored publicly available datasets analysing levels of gene expression in development. Finally, I used anatomical specificity to predict features extracted from RNA-seq developmental timecourse. Although this prediction was underpowered to properly explain RNA-seq features which resulted in a limited ability to explain variability from RNA-seq, however, by extracting feature importance of all covariates, I was able to determine the contribution of each covariate to describe gene expression profile. After combining the contribution and loadings of each covariate I obtained a measure of gene expression complexity. This measure was then used to explore the expression complexity of genes in ZFIN dataset. Indeed, developmentally important genes are the genes with the highest complexity information.

Unfortunately, since ZFIN dataset reflects the interests of the scientific community towards gene expression analysis, many of the genes with uniform expression pattern are not included in this dataset. Zebrafish was historically a model organism for development and neuronal system, the majority of genes analysed in this chapter are implicated in functions in these two domains. In addition, because I imposed that each gene analysed should be assayed by at least 5 publications, many of the true housekeeping genes were in this way discarded. Tomancak et al. did a similar investigation in *Drosophila* (Tomancak et al. 2007). They analysed images from embryonic *in situ*s and correlated them with microarray expression levels. However, in their analysis, they designed the experiments and produced *in situ* stainings which made their results much more standardised.

Anatomical specificity informs about localisation of gene expression, and can be thought as the alternative to tissue specificity coefficient. However, since anatomical specificity is developed based on the annotations of gene expression and not on the ratio of expression values across different tissues, it is a more stable measure of tissue-specificity than $\tau$ index. In addition, it can be used in the situations where there is not enough data to calculate $\tau$, like in embryonic development.

Zebrafish AO system is a part of Uberon, a cross-species anatomical ontology (Mungall et al. 2012). Inclusion of Zebrafish AO system to Uberon required Zebrafish AO system to be precisely annotated by using the traditional anatomical classification system. The advantage of this is that the annotations of anatomical structures and their relationships are reliable and comparable to AOs of other organisms. Additionally, since Anatomical specificity was constructed from Zebrafish AO, it would be possible to apply Anatomical specificity to define gene expression complexity in other organisms. Defining gene expression complexity in other organisms is going to be crucial to determine if the same type of genes are indeed the most complex. Also, gene expression complexity information from different organisms would help alleviate the problem of this dataset, the lack of annotations for some

genes.

During the analysis for this chapter, I encountered inconsistencies in the annotations that I removed from my analysis. At the same time, I have reported these inconsistencies to the ZFIN curators that have then acted upon it.

In this chapter, I have shown that genes with the highest gene expression complexity coefficient often belong to the group of GRB target genes. Those genes have a complex regulatory landscape with many enhancers in the region regulating their expression. The reason why these genes don't have particular motif-based feature significantly enriched could be due to the background that is already highly enriched for many motif-based features, but also it could be because these promoters might not be relying on the sequence motifs in the close proximity to drive their expression but rather they are regulated by distant regulatory elements. The fact that these genes belong to GRB target genes is indirect proof that they indeed relay on distant regulatory elements to drive their expression. Additional insights would be gained by analysing Promoter capture Hi-C data that would enable identification of distal promoter-interacting genomic regions (Mifsud et al. 2015).

# 4 Patterns of gene co-expression across developmental cell groups

## 4.1 Introduction

The profile of a cell's gene expression governs its phenotype. Subtle differences in gene expression during embryonic development and differentiation cause cells to have distinct functions and give rise to cell heterogeneity. During this process, histologically identical progenitor cells undergo distinct differentiation processes to become specific cell types. Historically, gene expression during differentiation was researched by analysing populations of cells. Since pooled cell populations do not provide an accurate measure of expression for individual cells, but rather average information from all pooled cells, the role of cell heterogeneity had not been explored genome-wide until now.

One approach for a better understanding of the relationships between genes across the genome during development is to construct gene co-expression networks. These networks describe gene relationships based on their correlated expression across samples. A subset of genes with highly correlated expression levels are biologically interesting because they imply a common regulatory mechanism, which often implies participation in similar biological processes. Gene co-expression networks have been used to compare expression patterns across tissues (Mack et al. 2019; Saha et al. 2017) and species (Eidsaa, Stubbs, and Almaas 2017; Nowick et al. 2009).

A general feature of co-expression networks is the existence of a limited number of highly connected genes, whereas the majority of genes have very few interactions (Barabási and Oltvai 2004). Highly connected genes are more pleiotropic than genes with few connections. They also show reduced variation in gene expression within one cell type due to stronger selective constraint in comparison to genes on the network periphery (Mähler et al. 2017).

Correspondence of co-expression networks across tissues was not extensively studied.

Tissues consist of combinations of different cell types, which in turn have a combination of housekeeping and tissue-specific genes. Because of this diversity, genes that are highly connected in one tissue co-expression network are often less connected in other tissues. Genes that are highly connected in multiple tissues should be especially highly pleiotropic.

In this chapter, I explore patterns of gene co-expression calculated from a scRNA-seq dataset examining expression shortly before and during zebrafish gastrulation, a developmental period in which significant changes in gene expression occur. scRNA-seq dataset analyses individual cells instead of whole tissues and embryos which provides higher sensitivity for detecting variability in gene expression. Genes that are frequently co-expressed in distinct tissues might share a great part of their regulatory profile. This is particularly important when analysing developmental genes that show elaborate regulation repertoire. I identified genes specific for a single cell group and compared their core promoter structure to those from genes active throughout all groups of cells in this dataset. Afterwards, I identified modules of co-expressed genes for each cell group. Finally, I integrated co-expression information from all clusters to obtain a measure of global gene co-expression across different cell groups of this dataset. Using global co-expression, I created networks of co-expressed genes that enabled me to identify co-expressed genes from different developmental pathways and to identify their common core promoter features.

## 4.2 Methods

### 4.2.1 Initial data and quality filtering

Single cell RNA-seq data analysed in this chapter was obtained from (Farrell et al. 2018). This study analysed the transcriptomes of 38,731 cells from 694 embryos by using Drop-seq technology (Macosko et al. 2015). The embryos spanned 9 hours of zebrafish development, from high blastula (3.33 hpf) stage which is just after maternal to zygotic transition to a 6-somite (12 hpf) stage which represents stage just after completion of gastrulation. The

raw data were obtained from NCBI GEO under accession GSE106587. This dataset was pre-filtered for low-quality molecular identifiers so I was able to proceed with aligning reads. Reads were aligned against the Ensembl Zv10 reference genome using Bowtie2 algorithm with the following parameters: – phred33 and –reorder (Langmead and Salzberg 2012). To obtain information on gene expression per cell analysed, DigitalExpression program from Drop-seq tools v1.01 suite was used (Macosko et al. 2015). Digital counting of transcripts is performed so that for each gene, a list of unique molecular identifiers (UMIs) per cell was generated. All UMIs that had an edit distance of one were merged. Finally, the total number of UMIs per gene per cell were counted and this number is reported as the number of transcripts of a gene in a cell. Cells for which more than 45% of their expression was coming from mitochondrial RNA were excluded from further analysis. In addition, cells whose transcriptomes were of low complexity were omitted. The complexity of the library was determined based on Farrell et al. who required a cell to have a specific number of genes and UMIs identified (Farrell et al. 2018). Separate intervals were determined for each developmental stage to accommodate a decrease in the transcriptome complexity during development. It can be observed that the number of UMI counts and genes active reduces as the development progresses. This could be explained by the fact that late stage embryos have reduced nuclear mRNA content in comparison to early stage embryos. Used intervals are provided in the Table 4.1:

**Table 4.1:** Quality control intervals which a cell needed to satisfy to remain in the analysis.

| Stage | Gene count | UMI count |
|---|---|---|
| High | 1000 - 7500 | 1500 - 40000 |
| Oblong | 625 - 7500 | 1500 - 30000 |
| Dome | 800 - 3800 | 2000 - 20000 |
| 30% Epiboly | 625 - 3000 | 1000 - 17500 |
| 50% Epiboly | 600 - 4000 | 1500 - 25000 |
| Shield | 600 - 2500 | 1000 - 15000 |

| Stage | Gene count | UMI count |
|---|---|---|
| 60% Epiboly | 600 - 3500 | 1500 - 22500 |
| 75% Epiboly | 600 - 3200 | 1400 - 20000 |
| 90% Epiboly | 500 - 3500 | 1000 - 20000 |
| Bud | 500 - 3200 | 1000 - 17500 |
| 3 - Somite | 500 - 3000 | 1000 - 12500 |
| 6 - Somite | 500 - 3000 | 1000 - 12500 |

### 4.2.2 Detection of variable genes

Expression of many genes in single cell RNA-seq datasets reflects technical variability or stochastic noise due to the single-cell protocols. To focus on the genes whose expression varies across different cell types and periods of development and not technical noise, I have focused this analysis only on the genes that show higher variability in expression across cells than what would be expected for a gene of similar expression but whose variability is only caused by technical noise. Variable genes were defined as in (Pandey et al. 2018). where the authors fit a null model that explained technical noise based on gene expression. Biologically variable genes were those whose variation of expression was at least 1.35 times greater than the null model variation. This analysis was done for each developmental stage separately. Finally, for later analyses, I used the union of all detected variable genes.

### 4.2.3 Creation of expression matrix and normalisation of the counts

Gene expression table was generated by adding all identified transcript counts to the corresponding gene. From quantified expression information, I created a matrix in which columns represented cells and rows represented genes. This expression matrix was then normalised to account for library size of each cell. Expression quantities of each cell were divided by the total number of detected transcripts in a given cell and multiplied by the

median number of UMIs detected in cells. All values were then log2 transformed.

### 4.2.4 Spectral clustering of single cells

Single cells were clustered based on their expression profile by using spectral, graph-based clustering using URD R package. Spectral clustering computes clusters based on principal components of the data. The first step of URD's clustering is to generate a graph of k-nearest neighbours (knn). To be able to generate clusters of varied sensitivity, I calculated knn graphs for seven k values (10, 30, 50, 60, 100, 150 and 200) from the normalised expression matrix. Based on the generated graphs, Louvain and Infomap clusterings were performed to generate clusters of cells with similar expression profile. All of the generated clusters were investigated to define an optimal model based on cluster dispersion and network modularity.

### 4.2.5 Identificattion of co-expressed genes

Modules of co-expressed genes were identified using WGCNA (Langfelder and Horvath 2008). A recent study showed how, when calculating co-expression networks, using counts that have been normalised on a sample level introduces additional bias. Using normalised values introduces unexpected co-variation, especially for lowly expressed genes (Crow et al. 2016). For this reason, for each identified Louvain cluster, I regenerated an expression matrix with raw UMI counts per gene. From the raw expression matrix, for each Louvain cluster, I created a signed network using all genes that had more than 5 tpm expression across all cells in a cluster. The first step of network creation is the determination of a soft power parameter. For each Louvain cluster, I tested a range of powers (1 to 10 and 12, 14, 16, 18 and 20) to investigate the optimal tradeoff between scale-free topology and mean connectivity of a network. For each of the analysed powers, I calculated the fit to a scale-free topology model and mean connectivity (including parameters such as the median number of connections per node, the mean number of connections each node in the network has, and the maximum number of connections for each of a range of specified powers) by using pickSoftThreshold

function from WGCNA package. Based on the values of these two functions, I picked a soft threshold for each Louvain cluster. More precisely, I chose a power value after which scale-free topology fit $R^2$ is above 0.8 and mean connectivity is in the asymptotic part of the curve. For clusters whose $R^2$ never reached 0.8, but the values were above 0.5, I used the inflexion point of scale independence curve as chosen power. Clusters whose maximum scale-free topology fit $R^2$ was below 0.5 were excluded from further analysis since I considered their clustering to be too noisy for downstream analyses. Using the determined soft-thresholding powers, I calculated a weighted adjecency matrix using adjecency function from WGCNA. This function calculates the pair-wise correlation of gene expression. Correlation values were weighted by raising them to the provided soft-thresholding power. The adjacency matrices were then transformed into TOMs by using TOMsimilarityFromExpr function. Co-expression modules were finally detected by performing soft clustering on the TOM dissimilarity values. Clusters identified by soft clustering were identified by using Mfuzz R package (Futschik, 2007).

### 4.2.6 Gene co-expression across Louvain clusters

All genes expressed in at least two Louvain clusters were included in the calculation of co-expression across the clusters. For each gene, I have annotated in which Louvain clusters they were found active, and in which co-expression module within clusters they were clustered. All genes that were assigned to a "grey" cluster in more than 60% of Louvain clusters were omitted from the further analysis. To calculate how often genes are coexpressed across Louvain clusters, I calculated pair-wise Jaccard similarity for all genes. For each pair of genes, I found the union of all cluster-module combinations in which they were identified. Jaccard similarity was calculated as a ratio of the intersection and union of cluster-module assignments for those two genes. The calculated similarities were then used to generate Gene Co-expression networks. From Jaccard similarities for each gene, I calculated Shannon entropy of their co-expression. Shannon entropy in this context represents the uncertainty of

co-expression across different cell clusters. More specifically, genes that show high uncertainty of their co-expression means that the fact that they are expressed in one cluster does not inform about their co-expression in other clusters.

### 4.2.7 Network of coexpressed genes

Vectors of Jaccard similarities for each gene were clustered using k-means clustering. Clustering was repeated for a number of k values, namely, I run k-means obtaining 8, 10, 12, 14, 18, 22, 25, 50, 82, 152 and 222 clusters. Based on the total within-cluster sum of square distance for each of clustering runs, I found the optimal number of clusters. The identified clusters were then inspected for enrichment in gene function and promoter structural elements. I run Gene Ontology analysis using clusterProfiler R package with all genes in this matrix as a background. For the enrichment in promoter features, I ran Promoter Ontology pipeline using zebrafish 30% epiboly promoter annotations as a reference. Again, all genes present in the co-expression matrix were used as background.

From the gene co-expression frequencies across Louvain clusters, I created network graphs by using ggraph R package. Here, nodes of a network represent genes, whereas edges indicate the magnitude of their co-expression. Since the interaction between genes in the network is determined based on the co-expression of gene alone, many of these interactions could be false positive. To identify putative direct targets, on an example of two transcription factors, I used ChIP-seq data analysing binding of transcription factor *cdx4* from the whole embryo. ChIP-seq data for *cdx4* was obtained from (Paik et al. 2013) (downloaded from Sequence Read Archive under accession code GSE48254). ChIP-seq reads were quality surveyed with FastQC and aligned to zebrafish danRer10 genome build using Bowtie2 version 2.3.0 (Langmead and Salzberg 2012), allowing a maximum of one mismatch in a seed region of 36 nt. TF binding peaks were then called using MACS2 (Feng, Liu, and Zhang 2011) with default settings. To identify which genes are direct targets of these transcription factors, I looked for TF binding peaks in promoter regions. Promoter regions were defined as regions

1000 nt upstream and downstream from dominant TSS, calculated using CAGE data from 30% epiboly embryos obtained from (Nepal et al. 2013).

## 4.3 Results

Single cell RNA-seq data used in this chapter are obtained from Farrell et al. The authors in that study investigated the transcriptomes of 39,488 embryonic cells and constructed developmental transcriptional trajectories. Using computed trajectories, they were able to observe strategies by which different tissue types differentiated their transcriptome profile from the common profile at the early stages of development. They have identified 25 distinct cell types that followed specific differentiation patterns. In the contrast to authors that were focused on defining developmental trajectories, I have explored if key developmental genes that are active across multiple identified cell types express differently based on the other genes active in those cells and the differentiation programme. Genes that are co-expressed share common regulatory programme, therefore, I aimed to identify co-expressing partners of key developmental genes within different cell types.

After obtaining Single cell RNA-seq from (Farrell et al. 2018) and performing quality filtering, I obtained a dataset describing the expression of 1883 variable genes with the information coming from 39,488 cells spanning 12 stages of embryonic development, from high blastula to 6-somite stage. To visualise all cells in two dimensions, I produced T-distributed stochastic neighbour embedding (tSNE) plot (Maaten and Hinton 2008). This method enables visualisation of high dimensional datasets where close neighbours are located close to each other. Based on user-supplied perplexity parameter, which controls the width of the Gaussian kernel used to calculate similarities between cells, different architectures of cells in 2D space are presented. I tested a range of perplexity coefficients, 10, 15, 20, 30, 50 and 80, and the architecture obtained with 30 coefficient produced most isolated clusters without overclustering cells. Figure 4.1A presents tSNE visualisation of 39 488 cells clustered on transcriptional similarity coloured based on the developmental stage. Clustering according

to the developmental stage can be observed, especially in the early stages of development. Clusters in lower right corner contain all cells originating from High blastula and oblong stage. Above that cluster, there is a cluster of cells from the dome stage above which there is a cluster of cells from 30% epiboly. The top right spread of cells represents cells from 50% epiboly and Sphere stage. After the Sphere stage, cell clustering based on cell stage is not as pronounced. From the middle of the graph towards the left, there is no clear clustering based on the developmental stage. However, there is a developmental trend within clusters: cells from the earliest stages on the right side of the cluster and the cells from the later stages on the left side. Cells in later stages of development show greater variability of gene expression which results in cells from those stages not being clustered together, but rather there is a subgrouping of cells with different transcriptional profiles.

To better understand transcriptional profiles of cells in this study and find groups of cells with common transcriptional profile, I have performed spectral clustering using Louvain modularity method. Louvain clustering using 30 nearest neighbours to derive cell clusters resulted in optimal cluster output. Figure 4.1B presents tSNE visualisation of the cells which are coloured based on Louvain-30 cluster assignment. There were 32 clusters identified, with the biggest cluster containing 3390 cells, while the smallest cluster contained 45 cells. Cells from High blastula and Oblong stage were clustered together in cluster 26, showing how similar transcription profiles of these cells are.

Additionally, I analysed how often the genes are expressed exclusively in one of the Louvain clusters which would suggest that those genes are specific for the function of that small subset of cells. As a contrast to genes exclusively active in one of the clusters, I analysed if there are genes that are expressed in all of the Louvain clusters. To do this analysis, I looked at the normalised gene expression of the whole Louvain cluster, instead of the expression of cells individually since individual cells are prone to technical expression noise. I calculated the total gene expression for a gene across the whole cluster and divided it by the total

cluster expression. Based on the expression proportion for all genes identified in a cluster, I devised a background expression level. All genes that had expression proportion higher than background expression level by 0.5% were considered active in that cluster. After this calculation, I have explored in how many clusters the genes are expressed. Figure 4.1C shows a distribution of gene expression frequency across all Louvain clusters. Interestingly, there are two large groups of genes, those expressed only in a single cluster, and those expressed in all clusters. A smaller proportion of genes are identified in more than one, but not all clusters.

To investigate if genes that are active only in a single Louvain cluster are stage-specific genes that are required for a short period of development, I analysed in how stage-specific gene expression is. I assigned genes as active in a specific stage with a similar method as with Louvain clusters. Namely, I considered a gene to be active in a specific stage if a proportion of reads mapped to that gene in a stage was higher than background expression level by 0.5%. This time, a great majority of genes are active in all 12 stages of development (Figure 4.1D). There are about two thousand genes expressed in up to two developmental stages.

### 4.3.1 Promoter features of genes expressed in a single Louvain cluster

Genes that are expressed only in a single Louvain cluster are genes that have a function specific to that cluster. Those genes are either specific for a developmental stage in which they are uniquely expressed, or are tissue-specific. In contrast to stage-specific genes, genes active in all Louvain clusters are expected to be housekeeping genes. They are active despite a different environment and the function of cells from different Louvain clusters. In addition, since this dataset assays developmental expression during gastrulation, it is possible that some developmental genes are active in all Louvain clusters.

To verify these claims, I have analysed promoter structure of genes active in a single Louvain cluster and of those active in all identified clusters. Figure 4.2A presents the results of a promoter structure overrepresentation analysis for genes active in a single cluster. In this analysis, all genes active in this dataset were used as a background. Clearly, these

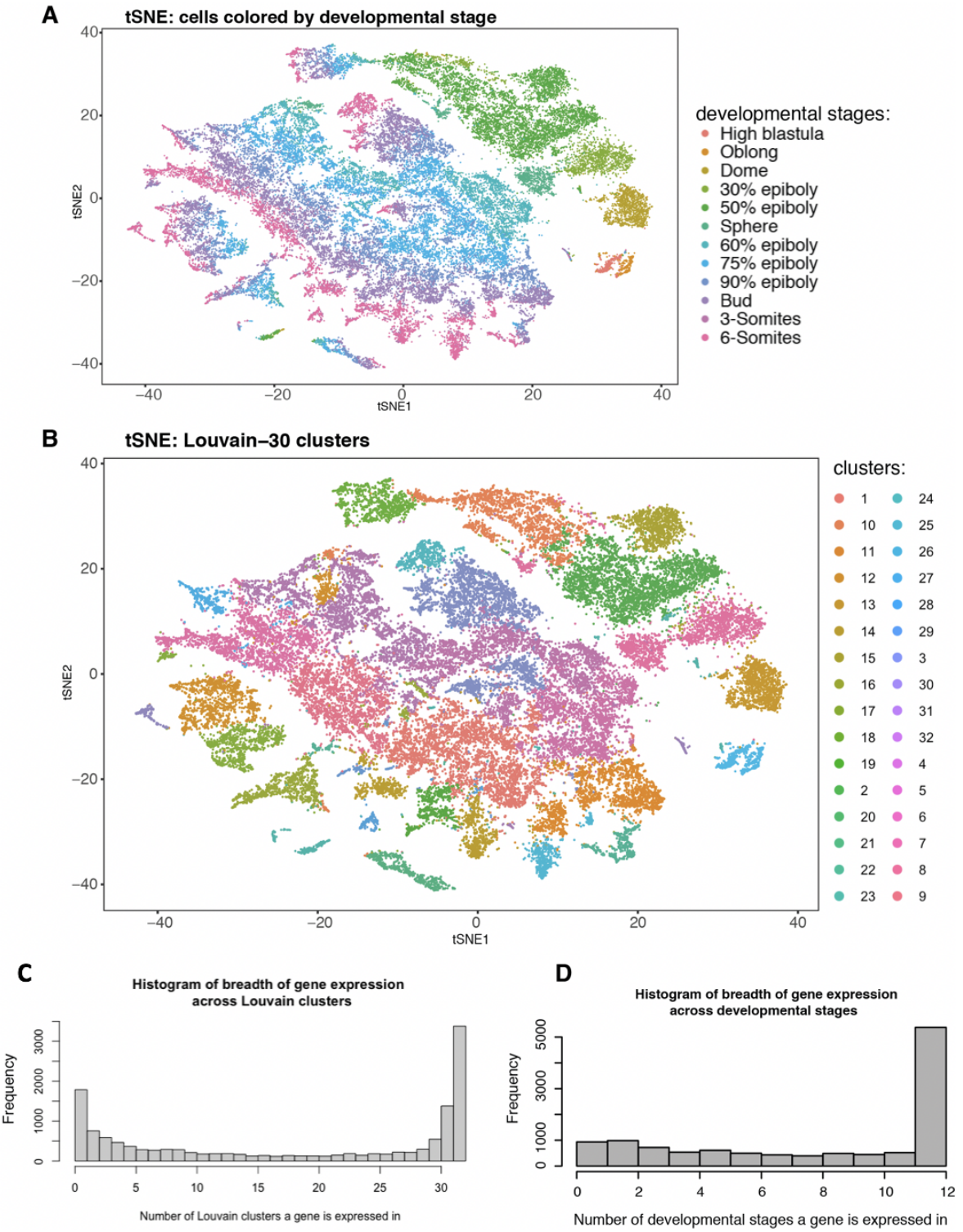


**Figure 4.1: tSNE representation of scRNA-seq dataset from 12 stages of zebrafish development.**

(**A**) tSNE representation of scRNA-seq dataset used where cells are coloured according to developmental stage. Twelve stages analysed in this study span 12 hours of development, from just after MZT to the beginning of somitogenesis. (**B**) tSNE plot of the scRNA-seq dataset where cells are coloured based on the Louvain cluster they are assigned to. (**C**) Histogram summarising in how many Louvain clusters genes are expressed. (**D**) Histogram representing the frequency of gene expression across developmental stages.

genes are indeed tissue-specific alike genes with significant position-specific enrichment for TATA-box, Initiator and GC-box. On the other hand, genes active in all Louvain clusters are genes that show enrichment for the presence of CpG island, upstream BRE element and bidirectional promoters (Figure 4.2B). Bidirectional genes were found to be enriched in DNA-repair, chaperones and DEAD-box helicases (Trinklein et al. 2004). Observing the enrichment of bidirectional genes supports the idea that genes active in all clusters are involved in housekeeping processes during development.

Figure 4.2A confirmed that genes active in a single Louvain cluster indeed have promoters with tissue specific characteristics, but for this analysis, I have used uniquely genes from all clusters. Louvain clusters contain cells that perform distinct functions, and their uniquely expressed genes could have very different promoter structures. To investigate if there are differences in promoter structure between genes uniquely expressed in different Louvain clusters, I have chosen to investigate the enrichment of promoter structures for three Louvain clusters. I have analysed only three clusters since they were the three clusters with the highest number of genes uniquely expressed in them. The remaining clusters had fewer than 50 uniquely expressed genes, which would not be sufficient to conclude about enrichment.

Figure 4.2C presents enrichment of promoter features for uniquely expressed genes from clusters 23, 26 and 31. Although, when they were analysed as one set, these promoters were enriched for TATA-box, Initiator and GC-box, when I analysed them separately, key differences emerged. Promoters from cluster 23 are most like the common promoter profile of these three clusters. These promoters are enriched for tissue-specific promoter features with enrichment of Inr, TATA-box and GC-box, and strongly depleted of CpG islands. Cluster 31 is also enriched for TATA-box. In contrast to that, cluster 26 is depleted of TATA-box. These results show that stage-specific genes active in different Louvain clusters contain distinct promoter profiles despite them all being stage specific.

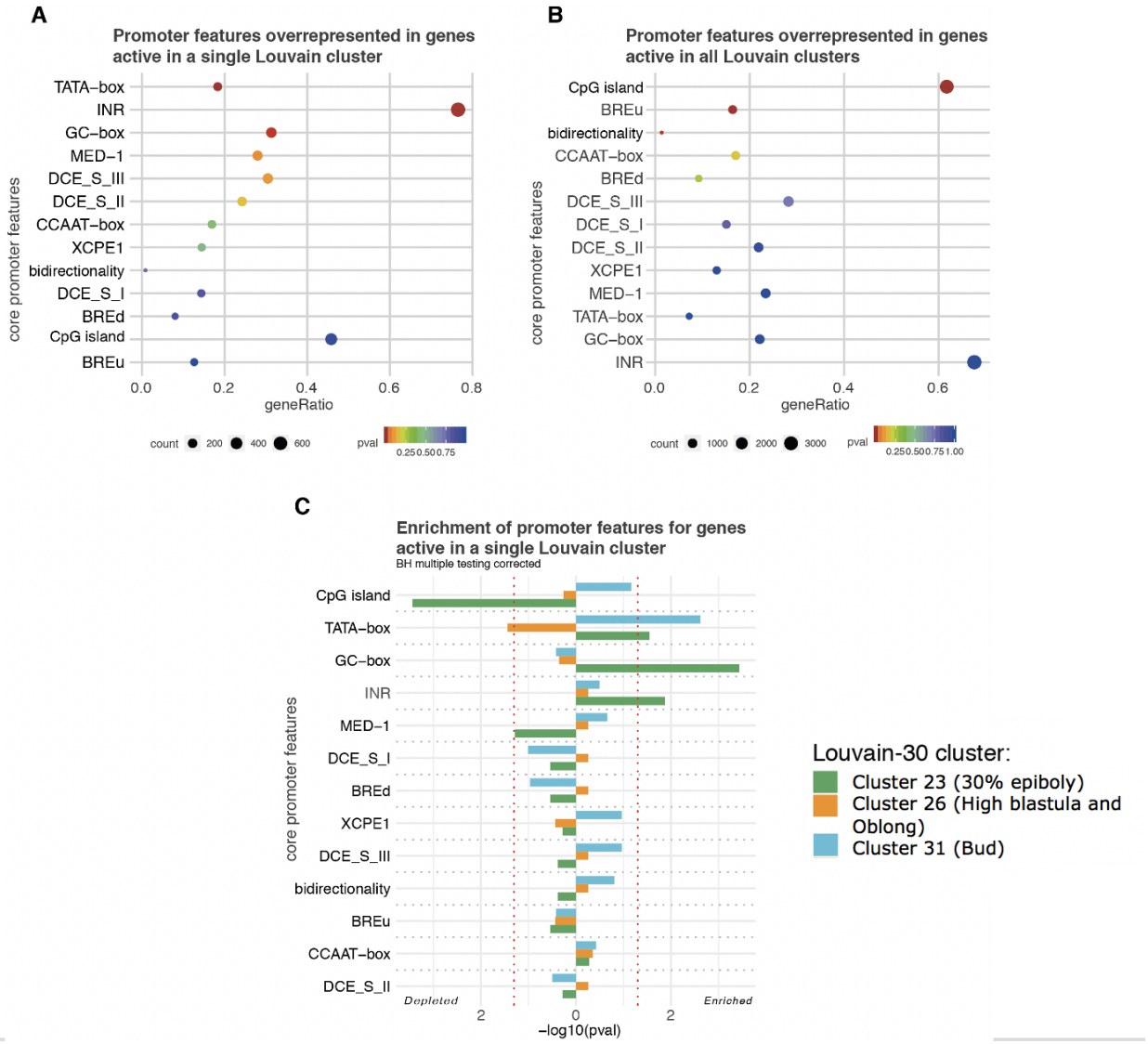


**Figure 4.2: Core promoter structure of genes active in a single or all Louvain clusters.** (**A**) Overrepresented core promoter features for genes active in a single Louvain cluster. All core promoter features represented by a red dot are significantly enriched in comparison to all genes in this dataset. For this analysis, I used Promoter Ontology algorithm using promoter annotations from 30% epiboly. (**B**) Overrepresented core promoter features for genes active in all Louvain clusters. Again, red coloured dots represent significantly enriched promoter features. (**C**) Comparison of core promoter structure for genes expressed exclusively in Louvain cluster 23 (green), 26 (orange) or 31 (blue). All bars extending left from the left dotted line are significantly depleted of that promoter feature. All bars extending right from the right dotted line are significantly enriched for that promoter feature.

### 4.3.2 Co-expression of genes across different tissues

The timing of gene expression is crucial for the proper function of a gene. Since developmental genes are expressed in many different tissues, I was curious to investigate the timing of expression across different embryo structures in which a gene is expressed. The more varied the expression timings are, the more should the gene require elaborate regulatory repertoire to result in the complex spatio-temporal pattern of expression. In addition, I analysed how often genes are co-expressed across different embryo structures. Particularly, I investigated if genes that are active in the same set of tissues co-express in all of the common tissues, or they have specific expression partners for each of the tissues they are active in. Genes that are co-expressed together in multiple tissues could require each other for the expression, while genes that are co-expressed only in a subset of tissues might have an additional mechanism that defines when these genes interact.

For this analysis, I have run WGCNA on each of the Louvain clusters. Louvain clustering separated cells based on their expression profiles and can, therefore, serve as a good estimate of different anatomical systems. WGCNA was used to identify co-expressed genes within the clusters. Before creating a weighted gene co-expression network, I calculated optimal soft-thresholding power for each cluster. Soft-thresholding power is applied to pair-wise correlations, and in that way, strong correlations are enhanced, while lowly correlated genes are penalised. For each Louvain cluster, I calculated following parameters: the $R^2$ for the scale independence, the mean number of connections each node in the network has, the median number of connections per node, and the maximum number of connections. All these parameters were calculated for a range of specified powers (1 to 10 and 12, 14, 16, 18 and 20). Soft-thresholding power that was finally defined as the value after which $R^2$ of the scale independence is higher than 0.8. Maximum $R^2$ for some Louvain clusters never reached 0.8; in that case, I used the inflexion point of scale independence curve as the chosen power.

Figure 4.3A and 4.3B show an example of soft thresholding calculation for Louvain

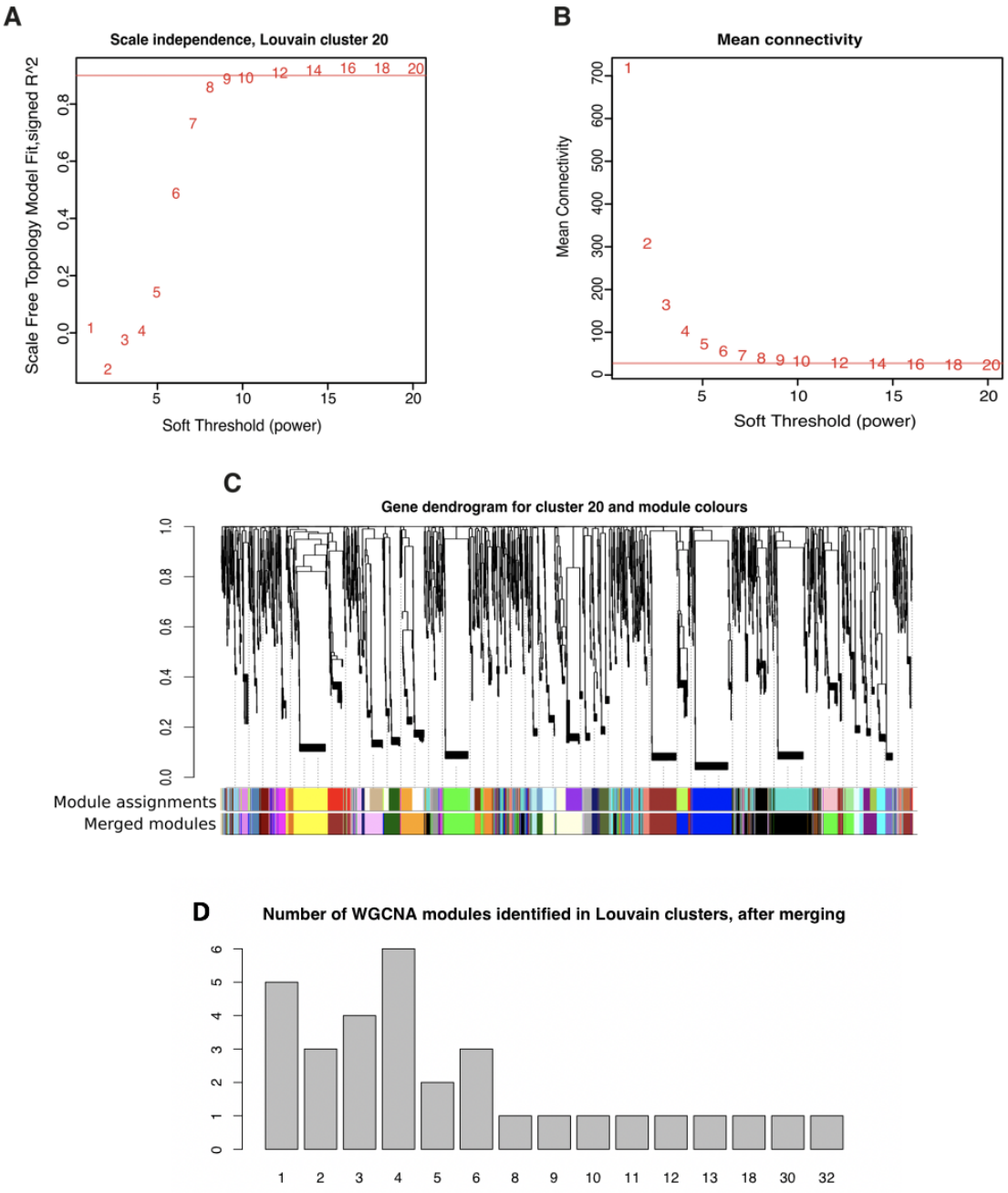


**Figure 4.3: WGCNA module identification for Louvain cluster 20.**
(**A**) Scale-free topology fit ($R^2$ on the y-axis) for a range of proposed soft thresholding values from This plot reaches a saturation level around 10 or 12, which could be optimal threshold values. (**B**) Mean connectivity of the network with respect to the soft thresholding power. With the increase of power values, mean connectivity decreases. (**C**) Gene dendrogram created by average linkage hierarchical clustering of expression values from cells in Louvain cluster 20. Y-axes represents TOM dissimilarity of expression profiles. Module assignment is represented in the ribbon below and it was obtained by soft clustering of the dendrogram. The original modules whose expression was very similar were merged and new cluster assignment can be seen in the lower ribbon. (**D**) Barplot showing how many co-expression modules were identified in each of the Louvain clusters.

cluster 20. In this example, I used 12 as an appropriate value for the soft threshold. The reason for this decision was the fact that the Scale Free Topology Model fit at 12 reaches the plateau and is above 0.8 (from Figure 4.3A). At the same time, according to Figure 4.3B mean connectivity of modules derived with a soft threshold of 12 is close to zero. The same decisions were used for the remaining Louvain clusters. Using soft thresholds, I calculated the adjacencies of clusters. By transforming the adjacency matrix, I generated the TOM and calculated dissimilarity of genes. WGCNA then uses hierarchical clustering to generate dendrogram of gene expression dissimilarities from which co-expression modules are identified.

Figure 4.3C presents the dendrogram of gene dissimilarities and the assignment of genes to co-expression modules. In this example, WGCNA algorithm initially identified 49 different modules of co-expression with sizes ranging from 443 to 24 genes. In the case of this cluster, there were no unclustered, "grey", genes. However, some of these modules have very similar expression profile and for downstream analyses, it is prudent to merge them. To quantify the similarity of expression profile for the whole module, I calculated eigenvalue of expression for each module and calculated the pair-wise correlation. For all pairs of modules whose eigenvalue correlation was above 0.75, I merged them into one module. Doing this analysis for all modules, reduced 19 modules for this cluster, so finally, I identified 30 clusters. Merged cluster calls can be seen in Figure 4.3C under a ribbon "merged modules" where it can be seen how some smaller clusters now became a part of larger clusters: for example, genes clustered under light green cluster disappeared and became a part of the blue cluster.

After merging similar modules across all Louvain clusters, Figure 4.3D represents the final number of modules identified in each of the clusters. Notably, there are five Louvain clusters in which no genes were clustered into co-expression modules, i.e. all genes remained unclassified. Those are clusters: 5, 11, 13, 26 and 30 from Figure 4.1B. Interestingly, cluster 26 is a cluster in which the earliest cells are clustered, and cluster 13 contains cells from Dome stage that is next stage in the developmental trajectory. Cluster 30 is a small cluster that

comprises 120 cells from 6-Somite stage and 26 from 3-Somites. Cluster 5 is, on the other hand, primarily made of cells from 60 and 70% Epiboly. These results suggest that in the case of cells from early developmental stages, WGCNA was not able to identify different modules of gene expression. The reason for this could be the fact that zygotic genome activates shortly before these stages and that cell transcriptome mainly consists of maternal transcripts that are being degraded. Degradation of maternal transcripts could cause remaining transcripts not to be correlated. For these clusters, I have re-run WGCNA with soft thresholding powers that were adjacent to the original value. Even in that case, no gene expression modules were identified.

A cluster that had the largest number of identified modules was cluster 31 with 32 modules, following by cluster 20 with 30 modules. These clusters were the smallest in the whole dataset and the only clusters with fewer than 100 cells. Additionally, these were the only clusters without "grey", unclustered, genes. These results suggest that in the case of clusters with low cell number WGCNA over-fitted expression profiles. For further analyses, I have discarded clusters with a single module, as well as clusters with less than 100 cells since these clusters showed low power of detecting expression modules.

### 4.3.3 Co-expression similarity across Louvain clusters

After identifying the modules of co-expression for remaining Louvain clusters, I proceeded to analyse gene co-expression across different Louvain clusters. For that, I created a matrix that annotates, for each gene, in which Louvain cluster they were active and to which module within a cluster it belongs to. For this analysis, I have discarded the unclustered (grey) genes across Louvain clusters. At this point, I was left with co-expression information for 11567 genes. Some of these genes were expressed in only a single cluster. Since I ought to calculate co-expression across different tissues, for further analyses, I have discarded genes expressed in a single cluster. This time, 8799 genes remained. For the remaining genes, I have calculated the pair-wise co-expression similarity. The similarity was defined as

Jaccard similarity of module assignment across all Louvain clusters in which these genes were expressed. Jaccard similarity represents the size of the intersection divided by the size of the union of the sample sets. Therefore Jaccard similarity of 1 would mean that genes are co-expressing in all clusters in which they were active, while the distance of 0 represents a situation in which genes are never found co-expressing in any cell cluster.

From pair-wise Jaccard similarity of co-expression, I calculated Shannon entropy of these values to test which genes have the most uniform profile of co-expression. Figure 4.4A shows a distribution of entropy values for co-expression with all other genes in the matrix. The gene with the lowest entropy is *ten1* gene whose entropy value is 6.234, while *cdx4* gene has the highest entropy with entropy value of 11.992. ten1 gene is a part of the telomere-associated complex in which it acts as a replication factor promoting DNA replication under stress conditions. On the other hand, *cdx4* gene is a homeobox-containing transcription factor that is involved in anteroposterior patterning and hematopoiesis during embryogenesis (Nes et al. 2006). Because of its multiple roles during development, in which it regulates different downstream genes, it is not surprising that its entropy of co-expression would be high.

Next, I analysed the function of 20 gens from both ends of entropy distribution. Gene Ontology enrichment results for the genes with the lowest entropy score can be seen in Figure 4.4B. These results confirm that genes with the lowest entropy are housekeeping genes involved in the cell cycle. These genes are expressed across an organism and are expressing in the complex of genes that regulates and maintains proper functioning of cell-cycle which would explain low entropy values. Gene Ontology enrichment for genes with the highest entropy values is presented in Figure 4.4C. These genes are involved in the regulation of developmental processes such as regionalisation, growth and brain development. This result suggests that developmental genes from this group are indeed pleiotropic and that they have distinct co-expression partners across different tissues.

Genes having the highest cumulative co-expression similarity across all genes are housekeeping genes with enriched GO functions highly overlapping those from Figure 4.4B. This observation is expected since housekeeping genes are defined as genes whose expression is very stable across different tissues. In addition, these genes will be expressed across all embryo cells. Therefore housekeeping genes were clustered as co-expressed in the majority of Louvain clusters and with many co-expressing partners, which resulted in high cumulative co-expression similarity.

### 4.3.4 Clustering co-expression similarity reveals clusters of similarly co-expressing genes

To find genes with similar patterns of gene co-expression across different clusters, I clustered the entire co-expression matrix using k-means clustering. K-means clustering is an unsupervised method that aims to stratify given data into k clusters in such a way that elements within clusters are as similar as possible, whereas elements of different clusters are as distinct as possible. I have clustered this matrix with a range of cluster quantities. For each k-means clustering run, I have calculated total intra-cluster variation which measures compactness of clusters. Values of total intra-cluster variation for each of the clustering runs can be seen in 4.5A. Initially, with the low number of clusters defined, the total sum of intra-cluster variation was the highest. This is because k-means clustering will always cluster all genes even if some are not similar to any of the clusters, they will be force-clustered into the closest one. With a low number of clusters defined, a higher proportion of genes will be dissimilar to defined clusters and their Euclidean distance to the centre of the cluster will contribute to the high sum of within-cluster variation. As the defined number of cluster increases, the total sum will decrease as there are fewer genes that are force-clustered. I choose 25 clusters as an optimal tradeoff between compactness of the clusters and their number since this is the point around the “elbow” of the curve, a point after which additional partitioning of the data brings scarce additional information.

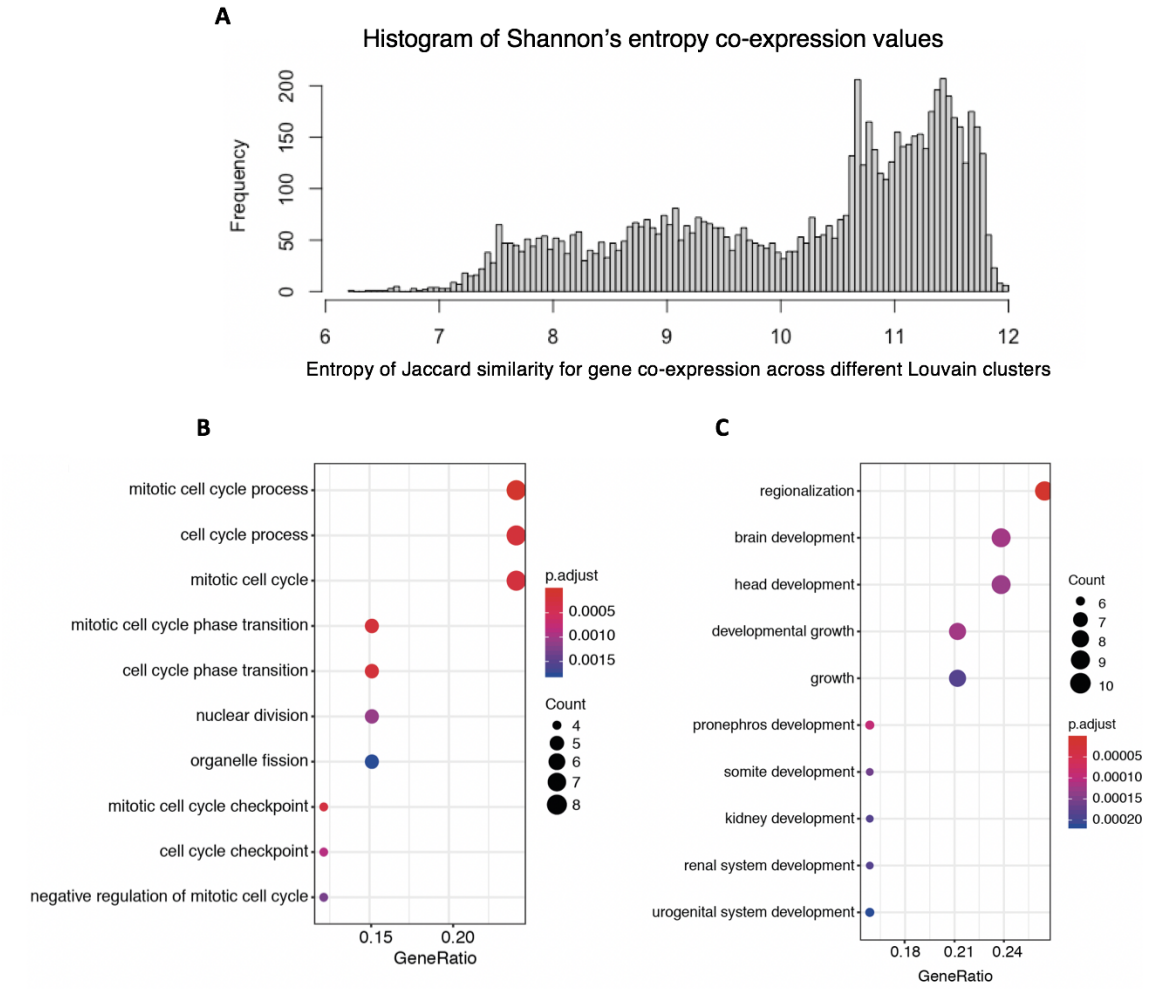


**Figure 4.4: The entropy of pair-wise gene co-expression Jaccard similarity values.** (**A**) Histogram representing the distribution of Jaccard similarity values for pair-wise gene co-expression across Louvain clusters. Genes with the high entropy values have many co-expressing partners; with some of them, they are often co-expressing across Louvain clusters, and with some only in a single cluster. Genes with low entropy values have constant co-expressing partners across all Louvain clusters in which they are active. (**B**) Gene Ontology enrichment of biological processes for genes with the lowest entropy values. (**C**) Gene Ontology enrichment of biological processes for genes with the highest entropy values.

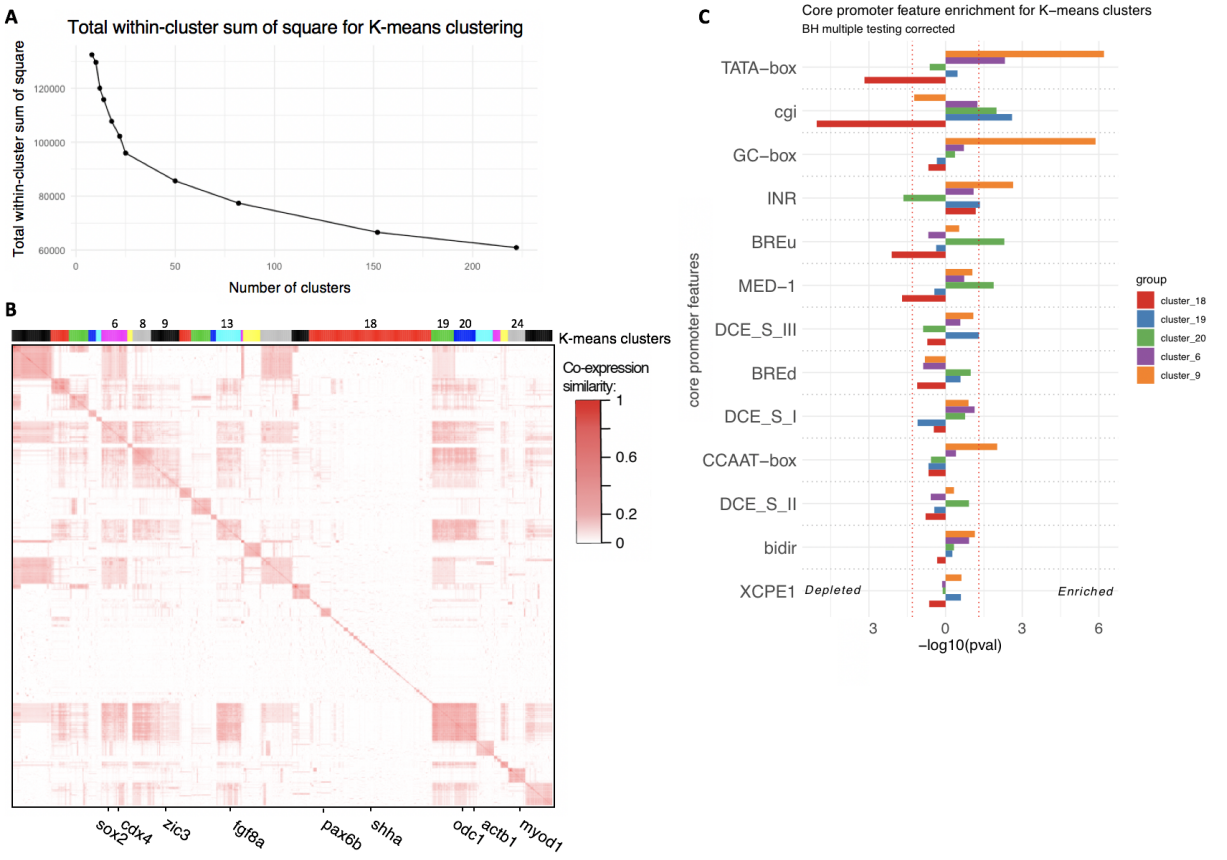


**Figure 4.5: K-means clustering of Jaccard similarity of co-expression values.** (**A**) The total within-cluster sum of the squared distance from the centre of the cluster for a different number of clusters. With the increase in the number of clusters, clusters contain more similar genes which cause the total sum to decrease. (**B**) Heatmap representing Jaccard similarity of all genes and their K-means clustering into 25 clusters. Each row and the column or this matrix represents a gene. There are a total of 8799 genes represented on this heatmap. At the bottom of the heatmap, some of the genes mentioned in the chapter are presented. All k-means assignments are represented in a ribbon at the top of the heatmap. (**C**) Comparison of core promoter structure for genes clustered into different k-means clusters. Purple bars represent cluster 6, orange cluster 9, red bars represent cluster 18, blue bars cluster 19, and finally cluster 20 is represented with green bars. All bars extending left from the left dotted line are significantly depleted of that promoter feature. All bars extending right from the right dotted line are significantly enriched for that promoter feature.

After clustering Jaccard similarity of co-expression for genes across Louvan clusters in 25 clusters, I obtained clusters shown in Figure 4.5B. These clusters vary in size and the intensity of their co-expression with other clusters. The rectangles on the diagonal represent co-expression within a cluster. In addition to co-expression within cluster, majority of clusters interact with other clusters, as seen by the signal away from the diagonal. One exception is cluster 7 which consists of 38 genes that, in this dataset, are exclusively co-expressed with genes within the cluster and are not co-expressing with other genes. These genes are involved in lysosome localization processes, as can be seen in the Table 4.2. Genes belonging to cluster 11 are, in addition to co-expressing with genes within the cluster, are also co-expressed with only a fraction of genes from cluster 3. Cluster 18 is the largest cluster that is made of distinct small groups of genes that are co-expressing with themselves and seldom with genes from the other clusters.

**Table 4.2:** Gene Ontology enrichment of Biological Processes in k-means clusters.

| K-means cluster | size | GO terms enriched | significance | in cluster; background |
|---|---|---|---|---|
| 2 | 160 | mitotic cell cycle | 2.83e-24 | 67; 404 |
| 2 | 160 | cell division | 2.28e-29 | 58; 238 |
| 4 | 61 | glycosylation | 0.0447 | 6; 61 |
| 7 | 38 | lysosome localization | 0.0059 | 4; 14 |
| 9 | 144 | tissue development | 4.85e-07 | 42; 421 |
| 9 | 144 | pattern specification process | 6.70e-06 | 19; 112 |
| 10 | 78 | organelle assembly | 1.31e-12 | 22; 109 |
| 10 | 78 | cell projection organization | 7.15e-10 | 29; 331 |
| 13 | 78 | ribosome biogenesis | 9.44e-11 | 31; 141 |
| 13 | 182 | RNA processing | 2.85e-06 | 51; 499 |
| 19 | 182 | RNA processing | 2.28e-15 | 68; 499 |
| 19 | 182 | RNA splicing | 2.11e-08 | 41; 261 |

| K-means cluster | size | GO terms enriched | significance | in cluster; background |
|---|---|---|---|---|
| 20 | 161 | RNA splicing | 5.76e-32 | 63; 261 |
| 20 | 161 | mRNA processing | 2.88e-30 | 64; 295 |
| 21 | 119 | chromosome organization | 0.0022 | 18; 170 |
| 22 | 62 | mitochondrial gene expression | 0.0094 | 9; 100 |
| 23 | 79 | protein targeting to ER | 1.71e-133 | 73; 87 |
| 23 | 79 | translational initiation | 6.61e-119 | 74; 119 |

Table 4.2 presents most significantly enriched GO terms from the biological processes ontology for each of the k-means clusters. Eleven clusters have at least one biological process enriched, with most having many more terms; for brevity, I am presenting the most significant ones. The remaining 14 clusters did not show enrichment for any of the biological processes. Majority of enriched processes in this table are housekeeping functions like RNA processing, cell division or chromosome organization, while only cluster 9 is enriched for developmental processes.

At the bottom of the heatmap from Figure 4.5B, I have shown the location of some important housekeeping and developmental genes. The gene with the highest entropy, *cdx4*, is located in cluster 6, together with *sox2* gene. *fgf8a*, the gene that was in the previous chapter defined as a gene with the most complex pattern of expression, is located in cluster 13. Gene Ontology results for this cluster define this cluster to be involved in RNA processing and ribosome biogenesis. This suggests that k-means clustering did not have sufficient power to separate all developmental genes from housekeeping genes. This task is challenging because both groups of genes are expressed in many different tissues.

*zic3* gene is located in cluster 9, the only cluster enriched for tissue development functions. Developmental genes such as *pax6b*, *pax6a* and *shh* are placed in cluster 18. *pax6* genes were expressed in 2 and 3 Louvain clusters respectively so the number of potential

co-expressing partners is greatly reduced. Finally, *myod1* gene is in the cluster 24 whereas genes *actb1* and *odc1* are in cluster 20.

Clusters 19 and 20 are very frequently co-expressed together, as can be seen from the big red square at the lower right of the heatmap. In addition, they are co-expressing with clusters 13 and 6. Clusters 13, 19 and 20 are enriched for the same biological functions, RNA processing, which explains their high frequency of co-expression. The only difference between clusters 19 and 20 is that cluster 19 is, in addition to above-mentioned clusters, also co-expressing with genes from clusters 1 and 16.

To further explore differences between clusters 19 and 20, I have analysed their core promoter features. For this analysis, I ran Promoter Ontology using promoter annotations form 30% epiboly as a reference and promoters from all genes in the heatmap 4.5B as a background. Both of these clusters are enriched for CpG islands, as would be expected for housekeeping genes (Figure 4.5C). Surprisingly, when looking at Inr, cluster 19 is enriched for Inr motif, while cluster 20 is depleted of the same motif. Despite the fact that this motif is not well defined in vertebrates, it could potentially explain why two groups of genes of the same function have slightly different co-expression patterns.

Cluster 9, despite being functionally enriched in developmental processes is enriched for TATA-box, GC-box, CAATT-box and Inr motif. Also, promoters of genes in cluster 6 are enriched for TATA-box motif. Frequent enrichment of TATA-box could be caused by the fact that by doing this analysis we have excluded true tissue-specific gene by imposing that all genes have to be mapped to WGCNA modules in at least two Louvain clusters. Also, this dataset contains only genes active after MZT and during gastrulation, a period in which many tissue-specific genes will not be active. Both of these facts reduce the number of TATA-box genes in the dataset and therefore enable a small proportion of TATA-box containing genes to be called significantly enriched. Finally, cluster 18 is depleted for both CpG islands and TATA-box.

### 4.3.5 Networks of genes co-expressed across development

After exploring the co-expression of individual genes and clusters of similar genes, I decided to create networks of genes co-expressed across Louvain clusters. Network modelling is a powerful way to interpret biological systems consisting of multiple elements that share common features. In biological systems, networks consist of nodes and edges. Depending on the context of the network, nodes can represent genes, proteins or metabolites, whereas edges represent the relationships between nodes such as physical binding, enzymatic reaction, regulation, or statistical correlation.

An example of a co-expression network is for the gene with the highest entropy in the dataset. I have extracted 25 genes that have the highest tissue co-expression similarity with *cdx4* Gene *hes6* has the highest co-expression with *cdx4* across Louvain clusters, with Jaccard similarity of 0.5. To include the edges between all genes in this subset, I calculated all pair-wise Jaccard similarity of co-expression across different Louvain clusters. Figure 4.6A shows a network around *cdx4* gene. To simplify the network, I discarded all nodes where genes were lowly co-expressed, and I represent only 40 edges with the highest co-expression similarity. As expected, *cdx4* has the most edges in this network, but despite that, other genes also show hub alike behaviour. *wnt5b*, *snai1a* and *sox11b* are examples of genes that are co-expressing with many genes in this network.

Many genes in this network are crucial for proper development. *snai1a* is a transcription factor that promotes the repression of the E-cadherin and in that way downregulates the expression of ectodermal genes within the mesoderm. *sox11b* is expressed in the developing neuronal system. It is an SRY-box transcription factor that regulates expression levels of genes involved in Hedgehog signalling, as well as the regulation of sprouting angiogenesis. *irx7* is a homeodomain transcription factor that takes part in retinal development in zebrafish. It regulates retinal differentiation by activating the expression of TFs that specify different retinal cell types downstream (Zhang et al. 2013). *ta* or *tbxta* is a T-box transcription factor

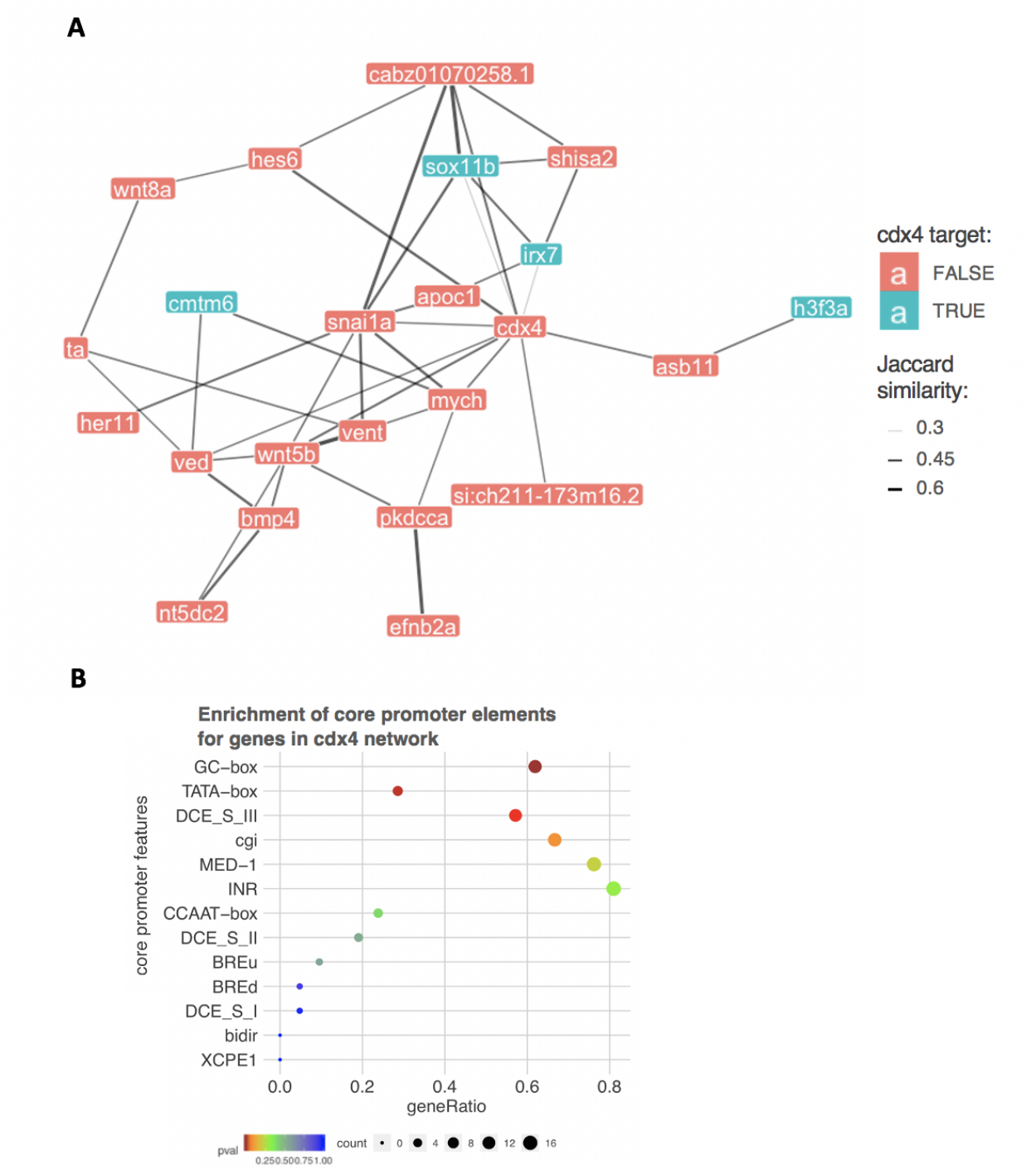


**Figure 4.6: Co-expression network for *cdx4* gene created from Jaccard similarity of co-expression values.**
(**A**) Gene network of co-expression across Louvain clusters. Nodes in this network represent genes with the highest frequency of co-expression with *cdx4* across all Louvain clusters. Edges of the network represent Jaccard similarity of co-expression across different clusters. Nodes are coloured depending if there was ChIP-seq signal enrichment for cdx4 in their promoter regions.

*(continued)*

**Figure 4.6:** (**B**) Overrepresented core promoter features for genes featured in cdx4 network. All core promoter features represented by a red dot are significantly enriched in comparison to all genes in this dataset. For this analysis, I used Promoter Ontology algorithm with promoter annotations from 30% epiboly.

---

involved in the regulation of mesoderm differentiation. It is indispensable for the formation of the tail structure. Hong et al. showed how much is required for zebrafish neural crest development and that its knockdown causes severe craniofacial abnormalities and defects in the eye (Hong, Tsang, and Dawid 2008). *vox* and *ved* are ventrally expressed homeobox genes that mediate Wnt signalling to restrict the organiser domain (Shimizu et al. 2005). Surprisingly, *cdx4* is involved in the function of all genes mentioned either as an upstream or downstream regulator of the above-mentioned processes.

*cdx4* is a member of caudal related homeobox transcription factor family and it is involved in anterior-posterior body patterning, development of extra-embryonic tissues and blood formation. It is well known for its ability to partially regulate *hox* genes by responding to retinoic acid, wnt and FGF. Additionally, *cdx4* is involved in determining the B-cell number and defining the placement of foregut organs (Kinkel et al. 2008). Davidson et al. showed that regulation of hox genes by *cdx4* is necessary for the determination of hematopoietic cell fate by rescuing blood deficiencies of cdx4 mutants by over-expression of certain hox genes (Davidson et al. 2003). Having such an important role in the regulation of numerous developmental processes explains why in this graph it has numerous interactions with other genes, but they are not of the highest similarity value. Often it happens that *cdx4* has lower similarity with a gene (like *snai1a*) which will than have much higher similarity values with peripheral genes. This could be explained by the fact that *cdx4* is expressed in a high number of Louvain clusters and therefore it is unlikely that a gene is going to be co-expressed with it in many of these clusters to obtain a high similarity value. The peripheral genes that are expressed in fewer clusters are then going to have higher similarity value, suggesting that these genes could already be committed to a more specific role in development.

Another interesting feature of this network is that two genes/transcripts don't have proper functional annotation. Gene annotation is a crucial element for learning precise insights into the biology and in this case development. Gene annotation is a laborious process which can require biological experiments and *in silico* analysis to devise the biological function of a gene product. Using network methods like this one we can devise expression patterns as well as co-expressing partners of these unannotated transcripts, leading to the testable hypothesis about their function.

Being based only on co-expression, this network may include many false positive interactions and indirect targets. To identify putative direct-binding targets of *cdx4* I used *cdx4* ChIP-seq data and screened if there were *cdx4*-binding peaks in promoter regions of these genes. ChIP-seq data I obtained is from the whole embryo at the Bud stage (10 hpf). I have identified 417 genes that had a binding peak of *cdx4* in their promoter. With this crude method, I have found *cdx4* peaks in promoters of only four genes in this network. This method was not sufficient to conclude which of these genes are targets of *cdx4*, as *cdx4* can regulate gene expression by binding to distal regulatory elements, and not only promoter regions. Also, I used ChIP-seq data from only one developmental stage and not the whole trajectory. Finally, some of these genes, like *wnt* genes, are directly regulating *cdx4* which we would not be able to verify with *cdx4* ChIP-seq data.

Although these genes are very diverse in their function, I analysed their promoter structure to see if they have any common promoter feature. When comparing promoter structure of genes in *cdx4* network to promoters from the entire co-expression matrix I find that these promoters are significantly enriched for the presence of GC-box and TATA-box along with downstream element III and CpG islands (Figure 4.6B). Again, enrichment of TATA-box in this dataset is likely to be artificially elevated since the background dataset is enriched for housekeeping genes and depleted of true tissue-specific genes.

## 4.4 Discussion

In the previous chapter, I defined a measure of gene expression complexity that accounted for dynamics of gene expression in levels, space and time. It showed that many developmental genes at the same time point are expressed in multiple tissues, at different levels of anatomical specificity. Following up on that work, in this chapter, I explored the dynamics of co-expression of developmental genes across all cell groups in which they are active. For this analysis, I obtained expression data from a method that enables interrogation of location and level of expression at the same time. I analysed scRNA-seq data from slightly fewer than 40,000 cells originating from a short period in development, from genome activation to the beginning of the somitogenesis (Farrell et al. 2018).

I showed that in the early stages of development, in particular, high, oblong and dome stage, cells cluster together suggesting that their transcriptomes are very similar regardless of where in embryo they come from. This result is in agreement with mRNA *in situ* hybridisation images coming from early developmental stages that did not show variability in their expression patterns. This observation could be explained by the fact that the zygotic genome in that period is still dormant. Only after 70% epiboly stage cells start separating into distinct clusters that represent different cell types.

When analysing temporally restricted expression patterns, I showed that the majority of genes detected in this study are expressed in all 12 stages of development assayed and that about 2000 genes are expressed in less than three stages of this timecourse. Furthermore, when analysing the specificity of a gene for a particular cell cluster, genes show a bimodal distribution with the majority of genes being expressed in all of the cell clusters, while other genes are found active in only one cluster. Genes active in all clusters represent housekeeping genes and the developmental genes.

Considering developmental genes are expressed in multiple tissues across the embryo where they commit cells to their lineages. I analysed whether these genes have a similar

pattern of activity across different tissues, or their expression is context-specific. For each cell cluster, I identified modules of co-expressed genes. Co-expressed genes are often controlled by the same regulatory program or they are functionally related. Genes that are most often found to be co-expressed with many genes age housekeeping genes. I also showed that these genes, across different cell groups, will be co-expressing with the same genes. This result is in accordance with (Saha et al. 2017) who reconstructed co-expression networks from 16 bulk RNA-seq samples and identified that hubs in these networks were strongly enriched for genes involved in RNA processing.

On the other hand, I showed that genes with the highest level of uncertainty of co-expression across different cell groups are developmental genes. This observation could suggest that although two developmental genes are expressed in the same cell, their interaction is context-specific and that they have distinct regulatory elements regulating expression in non-co-expressed tissues. Finally, I presented an example of a co-expression network around *cdx4* gene and its most frequent co-expressing partners. *cdx4* is most frequently co-expressed with other developmental genes, although their similarity of co-expression is significantly lower than in the case of housekeeping genes. This observation suggests that developmental genes, like *cdx4*, co-express with many other developmental genes only in a subset of tissues and that genes they are co-expressed with are more frequently co-expressed with other genes.

In this chapter, I have used WGCNA to identify modules of co-expressed genes. WGCNA infers modules of co-expression by calculating pair-wise correlation coefficients for all pairs of genes. However, correlation coefficients are not able to capture the non-linear relationship between variables which could be particularly important when analysing expression dynamics across development. A set of methods that are potentially able to model non-linear relationships between gene expression patterns exist. Some of those method are based on Bayesian statistics (Xiao et al. 2016; Friedman et al. 2000). These methods generate a network of co-expressed genes where conditional dependencies between the patterns of gene

expression are modelled. They require a lot of computational time to define networks and these methods are not able to account for mutually dependant genes like feedback loops. Another set of methods uses tree-based prediction models to define co-expressed genes. The idea of these methods is that genes which are co-expressed have predictive power to predict the expression pattern of co-expressed genes. For each gene (target), its expression pattern is predicted using expression patterns of all other genes by using methods like random forest (Huynh-Thu et al. 2010). The variable importance when predicting target's gene expression is used as an indication of the extent of co-expression. These methods are faster to compute than Bayesian methods and are able to infer non-linear co-expression.

Even though inferring co-expression by using correlations is limited because it is not able to define non-linear relationships between genes these methods have been often used to identify key genes in the biological pathways (Guo et al. 2019; Di et al. 2019; Maertens et al. 2018). Even in the field of scRNA-seq, Namboori et al. and Luo et al. have applied WGCNA to detect groups of co-expressed genes (Namboori et al. 2019; Luo et al.2015).

# 5 Discussion

Uncovering principles governing coordinated gene expression across space and time is crucial for a proper understanding of development of multicellular organisms. In this thesis, I explored the dynamics of gene expression in the embryonic development of zebrafish and its relationship with core promoter structure. First, I presented a novel method for overrepresentation analysis of core promoter elements for a group of genes. This method allows us to gain additional insights into how are groups of genes regulated. Next, I have summarised mRNA *in situ* hybridisation data from zebrafish development to define different modes of spatio-temporal dynamics of gene expression in development. Together with expression levels obtained from RNA-seq data for developmental time course, I have defined a novel measure of gene expression complexity. Using this measure, I showed that developmentally important genes are the most complexly expressed genes. Finally, by using single-cell data I have explored co-expression of genes across different groups of cells. In particular, for developmental genes expressed in multiple tissues, I showed that these genes show great variability in similarity of co-expression across different tissues. The observed variability could be attributed to the differences in the regulatory landscape of these genes.

## 5.1 Promoter Ontology

In Chapter 2, I introduced a novel method for identification of overrepresented core promoter features for a group of genes. This method starts from CAGE-seq data, to define promoters centred on dTSS. Next, it uses a default collection of predefined PolII core promoter features from the JASPAR database to annotate promoters defined by CAGE-seq and performs overrepresentation analysis. The main purpose of this method is to compare core promoter composition of different gene groups.

Using PromoterOntology, I annotated core promoters derived from zebrafish developmental timecourse as well as human cell line and primary tissue samples. This allowed

me to identify differences in core promoter composition of all active genes before and after zygotic genome activation. In addition, I showed differences in the composition of core promoters between human and zebrafish genes. I also showed how orthologous genes between zebrafish and human, when split based on the number of copies in these genomes, have distinct core promoter structure. Namely, orthologous genes that remained in a single copy in both genomes are strongly depleted of TATA-box and GC-box, while they are enriched for the presence of CpG islands. On the other hand, orthologous genes that are present in multiple copies in both human and zebrafish genome are depleted of CpG islands and are significantly enriched for TATA-box motif.

I have explored functional gene groups with the aim to identify characteristic promoter structure. For example, I showed that genes involved in telomere organisation are strongly enriched for both TATA-box and CpG islands. Finally, I showed that gene groups with similar biological function are more likely to also have similar core promoter features than a random gene group. Furthermore, I explored cases where this is not the trend.

A key feature of this method is that it relies on the CAGE-seq-defined TSS location. Although this provides an increased power to detect true-positive core promoter features, it limits the applicability of this method to genomes and cell types for which CAGE-seq or equivalent data is available. For samples that do not have CAGE-seq data, the most similar available CAGE-seq dataset could be used, or TSSs could be predicted by using computational methods like The Markov Chain Promoter Prediction Server (McPromoter) (Ohler 2006). McPromoter uses a hidden Markov model to predict TSSs by incorporating states that represent different promoter elements and nucleotide composition of the promoter region. However, these alternative methods will omit some TSSs since they will either reflect promoterome of a different sample or will be trained on a limited number of promoter features that might not identify all active TSSs.

Overrepresentation testing in Promoter Ontology uses Fisher's exact test that com-

pares the number of features in the sample to background. Some of the core promoter features have very low frequency in the population of active genes (less than 5%). These features in a small sample might, just by a chance, be reported as enriched, which might not be biologically important. Conversely, for features that are abundant in the population (above 85%), they are seldom going to be significantly enriched, since a small sample size would need to have almost all promoters with that feature to be significantly enriched.

This method defines PWM-based features by performing a sequence scan, where many reported hits are false-positive (Wasserman and Sandelin 2004). In Promoter Ontology, I have implemented a procedure that helps to discard the majority of these events, but despite that, there could still be some remaining false-positives.

Other groups developed methods for detection of core promoter features or for the detection of overrepresented motifs in a sequence. The Elements Navigation Tool (ElemeNT) is a web server for prediction of core promoter elements and their combinations (Sloutskin et al. 2015). This method reports core promoter features identified in a sequence, but it does not utilise CAGE-seq to precisely define TSSs, nor does it support analysis of gene groups as a whole. oPOSSUM is a web-based system for identifying over-represented combinations of motifs in a provided list of genes (Ho Sui et al. 2007). It is aimed at identifying overrepresented transcription factor motifs in co-expressed genes.

Promoter Ontology is publicly available R package that can be accessed ('PromoterOntology Github Repository', 2019). The package consists of multiple functions that:

- analyse CAGE-seq data and create a set of promoters centred on the dominant TSS;
- annotate promoters by a set of user-defined features (here, if a user does not provide a list of features, by default JASPAR PolII collection of PWM motifs is going to be used);
- do a quality check of all reported hits and creates statistics that help in determining the optimal threshold for calling true hits;

- perform enrichment analysis for provided core promoter features and creates modifiable figures;
- compare core promoter composition of multiple gene groups

The package also supports the addition of features for the analysis such as histone mark or methylation information.

Promoter Ontology could be a useful tool for the identification of additional promoter classes and their equivalent architectures in other organisms.

## 5.2 Spatio-temporal complexity of gene expression

In Chapter 3, I proposed a novel measure of the complexity of gene expression that incorporates information on the localisation, timing and the levels of gene expression in embryonic development. Defining such a measure was challenging, because two principal methods for studying gene expression patterns provide only a fraction of necessary information. Microscopy-based approaches like mRNA *in situ* hybridization can assay one or a few genes per experiment. Additionally, they do not provide an accurate measure of expression levels. Conversely, methods based on RNA sequencing provide a way to study expression on the genome-wide level, but since they require high amounts of starting material, it is challenging to obtain spatially resolved expression information.

I utilised annotation from mRNA *in situ* hybridisation assays to learn about localisation of gene expression. To be able to score and compare the localisation of different genes, I developed a measure of anatomical specificity that defines, for each anatomical structure used by ZFIN, how precisely its localisation is defined. I showed that, by using anatomical specificity, it is possible to distinguish different gene groups (tissue-specific, housekeeping or developmental) and that genes with the most dynamic profile of anatomical specificity in development are known developmentally important genes. Anatomical specificity can be used to identify specific periods in development when changes in the localisation of gene expression

occurred. Finally, by using anatomical specificity to predict features extracted from RNA-seq experiments, I devised a measure of gene expression complexity. Anatomical specificity is a more stable measure of tissue-specificity than $\tau$ index since anatomical specificity is created from available anatomical annotations and not derived from comparison of expression levels across various tissues.

Although this measure was able to show that developmentally important genes have the highest complexity of gene expression, it has its own set of potential limitations. Firstly, the basis of this measure are annotations of mRNA *in situ* hybridization. Although I used all publicly available annotations from ZFIN; however, this database is not exhaustive and annotations for many genes are lacking. A great majority of genes included in ZFIN are assayed by fewer than 5 publications, which does not provide a reliable amount of information to study dynamics of gene expression localisation in development. This resulted in a small subset of genes, mainly involved in the development and the neuronal system which are reliably represented. Also, this dataset is vulnerable to mistakes in manual curation of *in situ* hybridization images.

All localisation annotations were then scored for the specificity of localisation by using anatomical specificity. Anatomical specificity is built as a directed acyclic graph of anatomical structures, where the position in the hierarchy defines how specific anatomical term is. This only gives relative information on the size/complexity of anatomical structures, and there could be the case that two structures have the same anatomical specificity score, but their sizes are vastly different.

Finally, when predicting RNA-seq extracted features using anatomical annotations, with given parameters, I obtained a limited recall of values being predicted. With the increase in the recall, a more robust and reliable measure of complexity could be obtained. The recall could be increased by including interaction of the factors into random forest algorithm or by using more sophisticated prediction algorithms.

Other researchers were also interested in spatio-temporal patterns of gene expression in development. Kruse et al developed a method that can extract time, location and levels of expression (Kruse et al. 2016). Tomo-seq is based on cryosectioning of frozen samples in three dimensions and each slice is then sequenced. It is possible to computationally reconstruct the original sample in 3D and quantify transcriptome for each intersection of slices. This method has been applied to zebrafish embryos at three different developmental stages, identifying genes with spatially restricted expression pattern (Junker et al. 2014). The results from this study are a valuable resource for finding differentially expressed genes in particular slices of the embryo, however, this method provides lower spatial resolution than microscopy-based techniques, since it is limited by the thickness of the slice. Also, 3D reconstructed images can contain artefacts (Junker et al. 2014).

Early single-cell gene expression atlases tried to resolve spatio-temporal expression dynamics by assigning single cells to their location of origin (Satija et al. 2015). Satija et al. used *in situ* hybridisation expression pattern from a limited set of landmark genes to define the localisation of cells from single-cell experiments. This method is dependant on landmark gene annotations and, although it can define major cell types, it does not have sufficient power to precisely define anatomical structures within anatomical systems.

The results from Anatomical Specificity analysis are publicliy available through the website ('Zfexpress - Anatomical Specificity Interactive Web Session', 2019). This dataset provides a rich resource of gene expression localisation information that could assist fruther, more specific studies. This website provides multiple functionalities as shown in the tabs on Figure A.2:

- Retrieving dynamics of anatomical specificity during developmental time course for a gene of interest. Here, it is possible to select range of time points for which a graph will be generated, and change the scoring method for anatomical specificity.
- Users interested in particular developmental structure can retrieve genes active in that

structure at a particular time point and search through annotations supporting those reports.

- Retrieving original ZFIN anotations of gene expression and finding which publications provided particular expression evidence.
- Comparison of anatomical specificity dynamics for two genes as shown in Figure A.3. This feature can be particularly useful for researchers who want to find a gene with the most similar expression localisation pattern to another gene.

## 5.3 Patterns of gene co-expression across developmental cell groups

A limited number of cell signalling pathways are crucial for the regulation of a variety of developmental processes. It is still unclear how so few pathways provide specific cell communication required for the development of distinct tissues. For a better understanding of these processes, it is necessary to understand the spatio-temporal distribution and the interaction of these pathways.

In Chapter 3 I showed that developmentally important genes have the most dynamic spatio-temporal profiles of gene expression. Anatomical specificity revealed that these genes, at the same time, can be expressed broadly in one anatomic system, and restricted in another. Also, Tomancak et al. showed that genes with the same tissue-specific pattern of gene expression are rare and limited to adult organs (Tomancak et al. 2007). Despite the differences in global expression patterns, developmental genes sharing a subset of their expression profile could be jointly contributing to the regulation of developmental processes.

In Chapter 4, I identified patterns of spatio-temporal gene co-expression in development. In particular, I investigated co-expression patterns of developmental genes across different cell groups to explore if those genes, in cells in which they are both active, are co-expressing. Co-expression, in turn, could suggest that these genes share regulatory pro-

grammes or that they are functionally related. scRNA-seq data is particularly useful since, along with the dynamics of expression of a single gene, it allows to identify other genes with a similar expression pattern in that cell, or groups of cells.

I showed that transcriptomes of embryonic cells in early developmental stages are very similar, and there is no distinction in transcriptomes of cells from different embryo structures. This observation is in agreement with ZFIN mRNA *in situ* hybridisation annotations, which in early stages, used very few anatomical structures to describe gene expression. Additionally, when analysing the specificity of a gene for a particular cell cluster, genes show a bimodal distribution, with the majority of genes being expressed in all of the cell clusters, while other genes are found active in only one cluster. I tested if genes that are specifically active in one group of cells share core promoter features with other specifically active genes. I showed that different groups of genes uniquely expressed in different cell groups have distinct core promoter composition. Genes uniquely expressed in one cell group could be either specifically expressed in the single developmental stage or are tissue-specific genes, specific for terminally differentiated cell types. This diversity could explain the distinct core promoters across different cell clusters.

Next, I analysed if developmental genes that are expressed in multiple tissues have a similar pattern of co-expression across tissues, or if their expression is context-specific. Unsurprisingly, modules of co-expressed genes in each cell cluster. I showed that housekeeping genes have the highest number of co-expressing partners. These genes also have the most uniform profile of co-expressing partners across cell groups. A similar result was previously published in co-expression networks from 16 bulk RNA-seq samples (Saha et al. 2017). However, in that study, the tissues analysed were human adult tissues from postmortem donors, which limits overlap of co-expression patterns across different tissues since most developmental genes will not show dynamic expression profile in that context.

Contrarily, I demonstrated that developmental genes have the highest level of

uncertainty of co-expression across different cell groups. The uncertainty could suggest that, despite being expressed in the same tissue, co-expression of developmental genes is context-specific. To illustrate this observation, I presented a co-expression network around *cdx4* gene and its most frequent co-expressing partners. This network is enriched for developmental genes, however, their similarity of co-expression is significantly lower than observed in the network of housekeeping genes. Co-expression partners of *cdx4* often have low co-expression similarity with *cdx4* while at the same time, they are more frequently co-expressed with a small set of genes from this network. This observation could suggest that developmental genes, like *cdx4*, co-express with many genes where they regulate their expression and commit a cell to specific cell fate, but once cells are committed, they are not required anymore.

A common problem in co-expression networks is that some genes found to be co-expressed could be false-positives. To alleviate this problem, SCENIC uses PWM sequence scan to identify putative targets (Aibar et al. 2017). Only genes with significant enrichment for sequence motif are retained in the network construction step. PWM match is still only an indirect proof that the gene is targeted by a transcription factor since many sequence scan hits turn out to be false positives. In this analysis, I used ChIP-seq data from the whole embryo to identify which genes in the network are direct targets of *cdx4*. However, not finding ChIP-seq binding in a promoter of a gene does not directly exclude gene as a downstream target of a TF, since ChIP-seq experiment only shows a snapshot of regulatory events. Many genes co-expressed with *cdx4* in the network I presented are genes involved in *wnt* pathway, suggesting that some of these developmental genes act upstream of *cdx4*, which is their downstream target (Pilon et al. 2006).

When identifying modules of co-expressed genes, soft-clustering determines a similarity score for each module. Finally, when a gene is assigned to the co-expression module its similarity score can be extracted, along with the assignment. Instead of using binary readout of module assignment, potentially more resolution would be gained if the similarity score was

used in downstream analysis. In addition, instead of using the correlation of gene expression to infer co-expression, more power would be gained by using more complex statistics like the measure of dependency based on mutual information (MI) (Butte and Kohane 2000).

Despite the limitations of using correlations of gene expression to infer co-expression networks, this method has been widely applied to identify hub genes in biological pathways and disease (Guo et al. 2019; Di et al. 2019; Maertens et al. 2018). More recently, WGCNA was applied to scRNA-seq datasets too (Namboori et al. 2019; Xue et al. 2013; Luo et al. 2015).

Unfortunately, many scRNA-seq assays are analysing an isolated tissue sample, or are of limited sensitivity due to noise or the limited number of cells assayed. More recently, a novel framework that analyses co-expression across multiple cell groups and resolutions has been published (Hie et al. 2019). Using co-expression, this study is able to infer networks from multiple studies and in that way increase power for interpreting new biological phenomena.

## 5.4 Future directions

A common topic throughout this thesis was the characterisation of the dynamics of gene expression during the time course instead of analysing gene activity at individual time points. The information obtained from time course data provides unique insights into gene expression and its regulation.

I have shown that developmental genes have elaborate expression patterns throughout time, space and magnitude of their expression. However, key questions about regulation of expression repertoire are still unanswered. In particular, coordination and dynamics of regulation from distal regulatory elements, as well as details of their interaction with promoters of targeted genes, are not known.

One of the limitations when studying these questions using bulk transcriptome and epigenome assays is that developmental genes are expressed in multiple tissues at different

levels of specificity, which causes bulk data to average all signals from different parts of the embryo and therefore mask individual regulatory programmes. Recently developed assays for analysis of transcriptome and epigenome from single cells could help uncover the dynamics of gene regulation of this group of genes.

In particular, single-cell assay for transposase-accessible chromatin using sequencing (ATAC-seq) recently identified open chromatin states in *Drosophila* (Cusanovich et al. 2018). DNA methylation was also assayed in single cells (Clark et al. 2017). Finally, chromatin immunoprecipitation followed by sequencing (ChIP-seq) was used in single cells to discover different chromatin states in cell subpopulations (Rotem et al. 2015). However, single-cell methods are still suffering from the reduced spatial resolution of isolated cells. Often, a small group of marker genes is used to identify cell types (Satija et al. 2015) but even this method is limiting proper understanding of spatial dynamics.

Using anatomical specificity measure, I would identify a developmental period when a particular gene significantly changes its anatomical specificity in one of the anatomical systems, and characterise changes in epigenome in the stages just before the expression change. To increase the power of the analysis, I would compare all genes that show a similar change in anatomical specificity.

For example, many genes whose expression is spatially restricted, initially start expressing in larger anatomical structures and then gradually reduce its expression to the final expression localisation (just like in the example of *cdh5* gene in Figure 3.2 in Chapter 3). This gradual decrease in expression localisation may be caused by the recruitment of additional repressive marks or loss of activating expression signal in some areas of the embryo. By analysing changes in gene regulation in this period of gene expression, I would be able to detect repressive inputs.

Since in case of developmental genes, again, changes in the localisation of expression are not confined to a single anatomical system, I would use single-cell embryonic data to

identify changes in their regulatory profile across different cell populations and correlate it to changes in anatomical specificity for each of those cell populations. In this case, identifying cell groups from single-cell assays by only a limited number of marker genes would be sufficient, since as long as I know which anatomical system these cells belong to by using ZFIN database, I would be able to define a more precise localisation within that system.

By analysing changes in gene localisation across many different genes regulated by long-range regulation, potential patterns of co-regulation would emerge.

# A Appendices

## A.1 ZFIN Zebrafish developmental stages

**Table A.1:** Description of standard ZFIN developmental stages for zebrafish

| Stage order | Period | Stage | Begins - hours post fertilisation |
|---|---|---|---|
| 1 | **Zygote** (0 - 0.75 h) | 1-cell | 0.00 h |
| 2 | **Cleavage**(0.75 - 2.25 h) | 2-cell | 0.75 h |
| 3 | | 4-cell | 1.00 h |
| 4 | | 8-cell | 1.25 h |
| 5 | | 16-cell | 1.50 h |
| 6 | | 32-cell | 1.75 h |
| 7 | | 64-cell | 2.00 h |
| 8 | **Blastula**(2.25 - 5.25 h) | 128-cell | 2.25 h |
| 9 | | 256-cell | 2.50 h |
| 10 | | 512-cell | 2.75 h |
| 11 | | 1k-cell | 3.00 h |
| 12 | | High | 3.33 h |
| 13 | | Oblong | 3.66 h |
| 14 | | Sphere | 4.00 h |
| 15 | | Dome | 4.33 h |
| 16 | | 30%-epiboly | 4.66 h |
| 17 | **Gastrula**(5.25 - 10.33 h) | 50%-epiboly | 5.25 h |
| 18 | | Germ-ring | 5.66 h |
| 19 | | Shield | 6.00 h |
| 20 | | 75%-epiboly | 8.00 h |
| 21 | | 90%-epiboly | 9.00 h |

| Stage order | Period | Stage | Begins - hours post fertilisation |
|---|---|---|---|
| 22 | | Bud | 10.00 h |
| 23 | **Segmentation**(10.33 - 24 h) | 1-4 somites | 10.33 h |
| 24 | | 5-9 somites | 11.66 h |
| 25 | | 10-13 somites | 14 h |
| 26 | | 14-19 somites | 16 h |
| 27 | | 20-25 somites | 19 h |
| 28 | | 26+ somites | 22 h |
| 29 | **Pharyngula**(24 - 48 h) | Prim-5 | 24 h |
| 30 | | Prim-15 | 30 h |
| 31 | | Prim-25 | 36 h |
| 32 | | High-pec | 42 h |
| 33 | **Hatching**(48 - 72 h) | Long-pec | 48 h |
| 34 | | Pec-fin | 60 h |
| 35 | **Larval** | Protruding-mouth | 72 h |
| 36 | | Day 4 | 96 h |
| 37 | | Day 5 | 120 h |
| 38 | | Day 6 | 144 h |
| 39 | | Days 7-13 | 168 h |
| 40 | | Days 14-20 | 14 d |
| 41 | | Days 21-29 | 21 d |
| 42 | **Juvenile** | Days 30 -44 | 30 d |
| 43 | | Days 45 - 89 | 45 d |
| 44 | **Adult** | Days 90 + | 90 d |

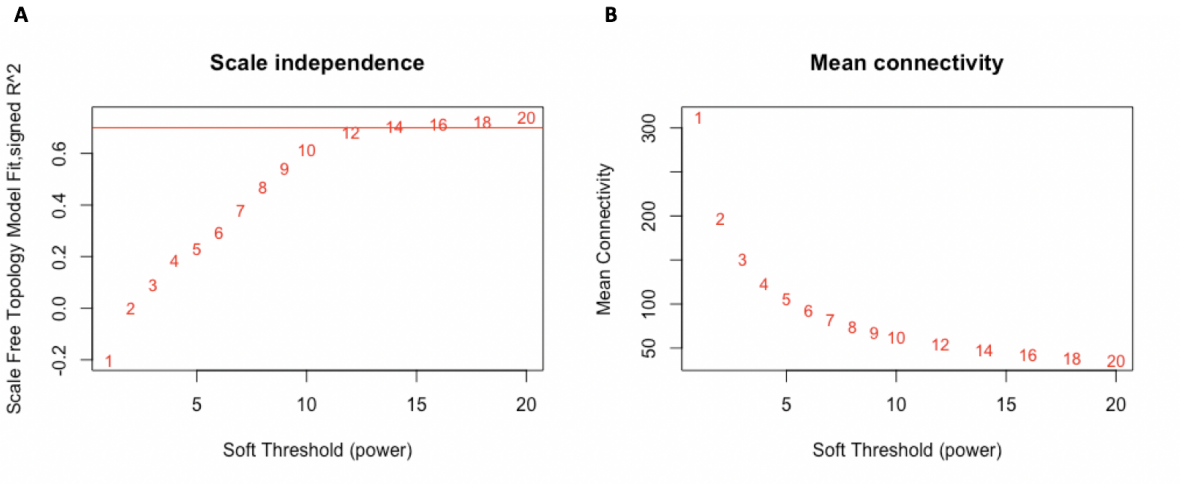


**Figure A.1: WGCNA module identification for the ZFIN expression table.**
(**A**) Scale-free topology fit ($R^2$ on the y-axis) for a range of proposed soft thresholding values from This plot reaches a saturation level around 12, which could be optimal threshold values. (**B**) Mean connectivity of the network with respect to the soft thresholding power. With the increase of power values, mean connectivity decreases.

## A.2 ZFexpress website containing results of Anatomical Specificity analysis

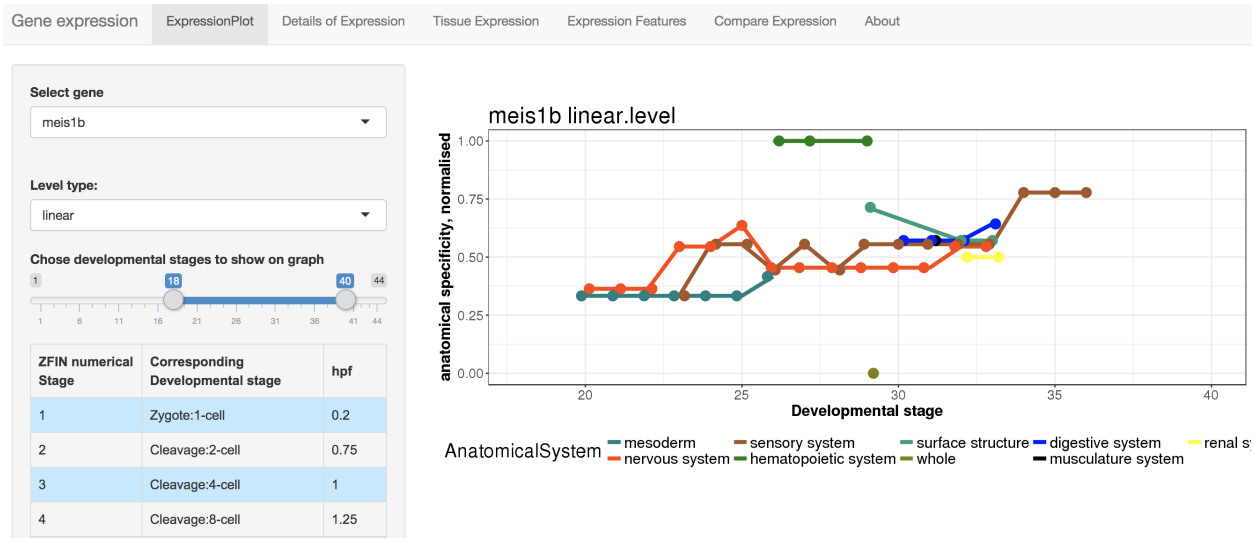


**Figure A.2: The snapshot of ZFexpress website with a graph of anatomical specificity.**
Results from Anatomical Specificity analysis of ZFIN database are provided in a website that can be used to retrieve patterns of gene expression in development and look for genes active in particular anatomical structures.

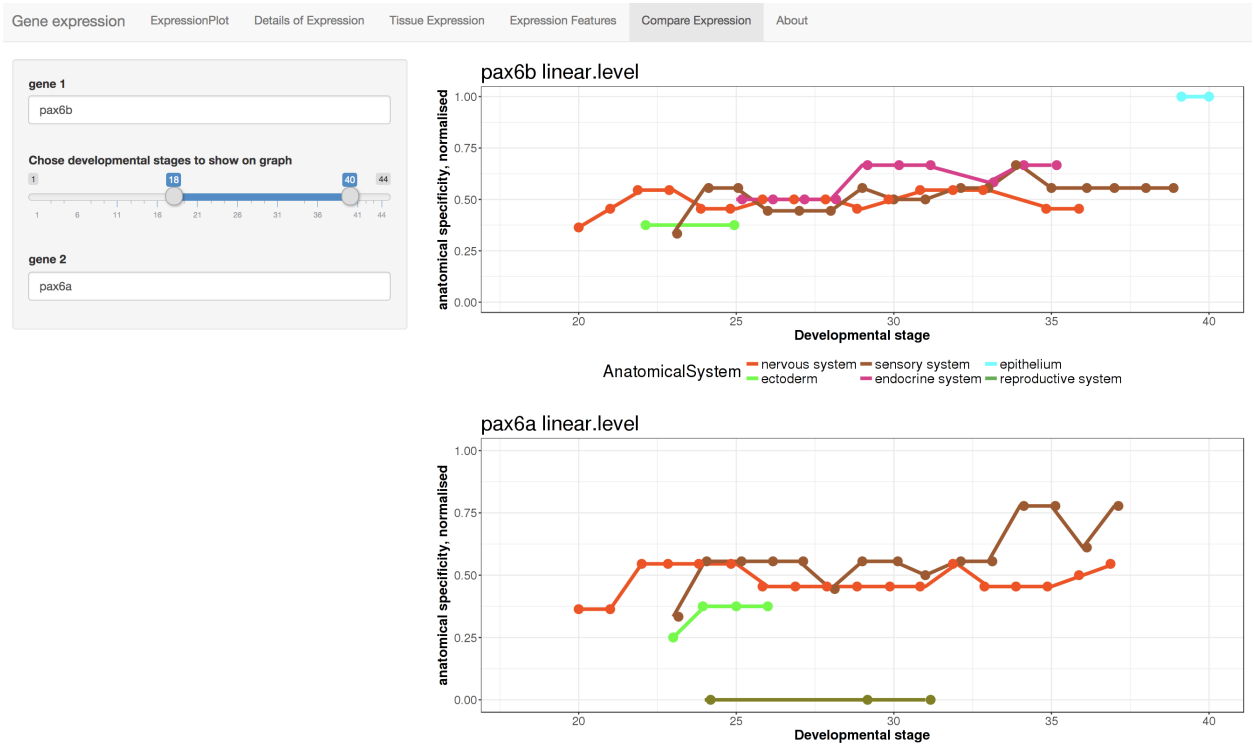


**Figure A.3: The snapshot of ZFexpress website - comparison of gene expression patterns.**
Results from Anatomical Specificity analysis of ZFIN database can also be used to compare and query expression patterns of two genes.

# References

Aibar, Sara, Carmen Bravo González-Blas, Thomas Moerman, Vân Anh Huynh-Thu, Hana Imrichova, Gert Hulselmans, Florian Rambow, et al. 2017. "SCENIC: Single-Cell Regulatory Network Inference and Clustering." *Nat. Methods* 14 (11): 1083–6.

Akalin, Altuna, David Fredman, Erik Arner, Xianjun Dong, Jan Christian Bryne, Harukazu Suzuki, Carsten O Daub, Yoshihide Hayashizaki, and Boris Lenhard. 2009. "Transcriptional Features of Genomic Regulatory Blocks." *Genome Biol.* 10 (4): R38.

Anders, Simon, Paul Theodor Pyl, and Wolfgang Huber. 2015. "HTSeq–a Python Framework to Work with High-Throughput Sequencing Data." *Bioinformatics* 31 (2): 166–69.

Armant, Olivier, Martin März, Rebecca Schmidt, Marco Ferg, Nicolas Diotel, Raymond Ertzer, Jan Christian Bryne, et al. 2013. "Genome-Wide, Whole Mount in Situ Analysis of Transcriptional Regulators in Zebrafish Embryos." *Dev. Biol.* 380 (2): 351–62.

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

Barabási, Albert-László, and Zoltán N Oltvai. 2004. "Network Biology: Understanding the Cell's Functional Organization." *Nat. Rev. Genet.* 5 (2): 101–13.

Bejerano, Gill, Michael Pheasant, Igor Makunin, Stuart Stephen, W James Kent, John S Mattick, and David Haussler. 2004. "Ultraconserved Elements in the Human Genome." *Science* 304 (5675): 1321–5.

Benjamini, Y, and Y Hochberg. 1995. "Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing." *J. R. Stat. Soc.*

Bernstein, Bradley E, Tarjei S Mikkelsen, Xiaohui Xie, Michael Kamal, Dana J Huebert, James Cuff, Ben Fry, et al. 2006. "A Bivalent Chromatin Structure Marks Key Developmental Genes in Embryonic Stem Cells." *Cell* 125 (2): 315–26.

Binns, David, Emily Dimmer, Rachael Huntley, Daniel Barrell, Claire O'Donovan, and Rolf Apweiler. 2009. "QuickGO: A Web-Based Tool for Gene Ontology Searching." *Bioinformatics* 25 (22): 3045–6.

Bolger, Anthony M, Marc Lohse, and Bjoern Usadel. 2014. "Trimmomatic: A Flexible Trimmer for Illumina Sequence Data." *Bioinformatics* 30 (15): 2114–20.

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 Dev.* 10 (6): 711–24.

Butte, A J, and I S Kohane. 2000. "Mutual Information Relevance Networks: Functional Genomic Clustering Using Pairwise Entropy Measurements." *Pac. Symp. Biocomput.*, 418–29.

Calle-Mustienes, Elisa de la, Cármen Gloria Feijóo, Miguel Manzanares, Juan J Tena, Elisa Rodrı'guez-Seguel, Annalisa Letizia, Miguel L Allende, and José Luis Gómez-Skarmeta. 2005. "A Functional Survey of the Enhancer Activity of Conserved Non-Coding Sequences from Vertebrate Iroquois Cluster Gene Deserts." *Genome Res.* 15 (8): 1061–72.

Cao, Junyue, Malte Spielmann, Xiaojie Qiu, Xingfan Huang, Daniel M Ibrahim, Andrew J Hill, Fan Zhang, et al. 2019. "The Single-Cell Transcriptional Landscape of Mammalian Organogenesis." *Nature*, February, 1.

Carlson, Marc, Seth Falcon, Herve Pages, and Nianhua Li. 2015. "GO. Db: A Set of Annotation Maps Describing the Entire Gene Ontology." *R Package Version* 3 (0): 568.

Carninci, Piero, Albin Sandelin, Boris Lenhard, Shintaro Katayama, Kazuro Shimokawa, Jasmina Ponjavic, Colin A M Semple, et al. 2006. "Genome-Wide Analysis of Mammalian Promoter Architecture and Evolution." *Nat. Genet.* 38 (6): 626–35.

Carninci, P, T Kasukawa, S Katayama, J Gough, M C Frith, N Maeda, R Oyama, et al. 2005. "The Transcriptional Landscape of the Mammalian Genome." *Science* 309 (5740):

1559–63.

Choy, John S, Sijie Wei, Ju Yeon Lee, Song Tan, Steven Chu, and Tae-Hee Lee. 2010. "DNA Methylation Increases Nucleosome Compaction and Rigidity." *J. Am. Chem. Soc.* 132 (6): 1782–3.

Ciccarese, Paolo, Marco Ocana, Leyla Jael Garcia Castro, Sudeshna Das, and Tim Clark. 2011. "An Open Annotation Ontology for Science on Web 3.0." *J. Biomed. Semantics* 2 Suppl 2 (May): S4.

Clark, Stephen J, Sébastien A Smallwood, Heather J Lee, Felix Krueger, Wolf Reik, and Gavin Kelsey. 2017. "Genome-Wide Base-Resolution Mapping of DNA Methylation in Single Cells Using Single-Cell Bisulfite Sequencing (scBS-seq)." *Nat. Protoc.* 12 (3): 534–47.

Collings, Clayton K, Peter J Waddell, and John N Anderson. 2013. "Effects of DNA Methylation on Nucleosome Stability." *Nucleic Acids Res.* 41 (5): 2918–31.

Crow, Megan, Anirban Paul, Sara Ballouz, Z Josh Huang, and Jesse Gillis. 2016. "Exploiting Single-Cell Expression to Characterize Co-Expression Replicability." *Genome Biol.* 17 (May): 101.

Cusanovich, Darren A, Andrew J Hill, Delasa Aghamirzaie, Riza M Daza, Hannah A Pliner, Joel B Berletch, Galina N Filippova, et al. 2018. "A Single-Cell Atlas of in Vivo Mammalian Chromatin Accessibility." *Cell* 174 (5): 1309–1324.e18.

Davidson, Alan J, Patricia Ernst, Yuan Wang, Marcus P S Dekens, Paul D Kingsley, James Palis, Stanley J Korsmeyer, George Q Daley, and Leonard I Zon. 2003. "Cdx4 Mutants Fail to Specify Blood Progenitors and Can Be Rescued by Multiple Hox Genes." *Nature* 425 (6955): 300–306.

Deng, Wensheng, and Stefan G E Roberts. 2005. "A Core Promoter Element Downstream of the TATA Box That Is Recognized by TFIIB." *Genes Dev.* 19 (20): 2418–23.

De Robertis, E M, J Larraı'n, M Oelgeschläger, and O Wessely. 2000. "The Establishment of Spemann's Organizer and Patterning of the Vertebrate Embryo." *Nat. Rev. Genet.* 1 (3): 171–81.

Dezso, Zoltán, Yuri Nikolsky, Evgeny Sviridov, Weiwei Shi, Tatiana Serebriyskaya, Damir Dosymbekov, Andrej Bugrim, et al. 2008. "A Comprehensive Functional Analysis of Tissue Specificity of Human Gene Expression." *BMC Biol.* 6 (November): 49.

Di, Yu, Dongshan Chen, Wei Yu, and Lei Yan. 2019. "Bladder Cancer Stage-Associated Hub Genes Revealed by WGCNA Co-Expression Network Analysis." *Hereditas* 156 (January): 7.

Dickmeis, Thomas, and Ferenc Muller. 2005. "The Identification and Functional Characterization of Conserved Regulatory Elements in Developmental Genes." *Brief. Funct. Genomic. Proteomic.* 3 (4): 1–19.

Dixon, Jesse R, Siddarth Selvaraj, Feng Yue, Audrey Kim, Yan Li, Yin Shen, Ming Hu, Jun S Liu, and Bing Ren. 2012. "Topological Domains in Mammalian Genomes Identified by Analysis of Chromatin Interactions." *Nature* 485 (7398): 376–80.

Dobin, Alexander, Carrie A Davis, Felix Schlesinger, Jorg Drenkow, Chris Zaleski, Sonali Jha, Philippe Batut, Mark Chaisson, and Thomas R Gingeras. 2013. "STAR: Ultrafast Universal RNA-seq Aligner." *Bioinformatics* 29 (1): 15–21.

Down, Thomas A, Casey M Bergman, Jing Su, Tim J P Hubbard, and E Davydov. 2007. "Large-Scale Discovery of Promoter Motifs in Drosophila Melanogaster." *PLoS Comput. Biol.* 3 (1): e7.

Eidsaa, Marius, Lisa Stubbs, and Eivind Almaas. 2017. "Comparative Analysis of Weighted Gene Co-Expression Networks in Human and Mouse." *PLoS One* 12 (11): e0187611.

Eisenberg, Eli, and Erez Y Levanon. 2013. "Human Housekeeping Genes, Revisited." *Trends Genet.* 29 (10): 569–74.

Ellingsen, Staale, Mary A Laplante, Melanie König, Hiroshi Kikuta, Tomasz Furmanek, Erling A Hoivik, and Thomas S Becker. 2005. "Large-Scale Enhancer Detection in the Zebrafish Genome." *Development* 132 (17): 3799–3811.

Falcon, S, and R Gentleman. 2007. "Using GOstats to Test Gene Lists for GO Term Association." *Bioinformatics* 23 (2): 257–58.

Farrell, Jeffrey A, Yiqun Wang, Samantha J Riesenfeld, Karthik Shekhar, Aviv Regev, and Alexander F Schier. 2018. "Single-Cell Reconstruction of Developmental Trajectories During Zebrafish Embryogenesis." *Science* 360 (6392).

Feng, Jianxing, Tao Liu, and Yong Zhang. 2011. "Using MACS to Identify Peaks from ChIP-Seq Data." In *Current Protocols in Bioinformatics*, Chapter 2:Unit 2.14. Hoboken, NJ, USA: John Wiley & Sons, Inc.

FitzGerald, Peter C, David Sturgill, Andrey Shyakhtenko, Brian Oliver, and Charles Vinson. 2006. "Comparative Genomics of Drosophila and Human Core Promoters." *Genome Biol.* 7 (7): R53.

Flores, Maria Vega, Christopher J Hall, Kathryn E Crosier, and Philip S Crosier. 2010. "Visualization of Embryonic Lymphangiogenesis Advances the Use of the Zebrafish Model for Research in Cancer and Lymphatic Pathologies." *Dev. Dyn.* 239 (7): 2128–35.

Forrest, Alistair R R, Hideya Kawaji, Michael Rehli, J Kenneth Baillie, Michiel J L de Hoon, Timo Lassmann, Masayoshi Itoh, et al. 2014. "A Promoter-Level Mammalian Expression Atlas." *Nature* 507 (7493): 462–70.

Friedman, N, M Linial, I Nachman, and D Pe'er. 2000. "Using Bayesian Networks to Analyze Expression Data." *J. Comput. Biol.* 7 (3-4): 601–20.

Futschik, Matthias E. 2007. "Introduction to Mfuzz Package and Its Graphical User Interface."

Genomics, Chromium 10x. 2017. "Transcriptional Profling of 1.3 Million Brain Cells with the Chromium Single Cell 3' Solution."

Girard, Mathilde, and Michel Goossens. 2006. "Sumoylation of the SOX10 Transcription Factor Regulates Its Transcriptional Activity." *FEBS Lett.* 580 (6): 1635–41.

GTEx Consortium. 2013. "The Genotype-Tissue Expression (GTEx) Project." *Nat. Genet.* 45 (6): 580–85.

Gu, Weifeng, Heng-Chi Lee, Daniel Chaves, Elaine M Youngman, Gregory J Pazour, Darryl Conte Jr, 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.

Guo, Yuhe, Junjie Ma, Lanyan Xiao, Jiali Fang, Guanghui Li, Lei Zhang, Lu Xu, Xingqiang Lai, Guanghui Pan, and Zheng Chen. 2019. "Identification of Key Pathways and Genes in Different Types of Chronic Kidney Disease Based on WGCNA." *Mol. Med. Rep.* 20 (3): 2245–57.

Haberle, Vanja, Alistair R R Forrest, Yoshihide Hayashizaki, Piero Carninci, and Boris Lenhard. 2015. "CAGEr: Precise TSS Data Retrieval and High-Resolution Promoterome Mining for Integrative Analyses." *Nucleic Acids Res.* 43 (8): e51.

Haberle, Vanja, and Boris Lenhard. 2016. "Promoter Architectures and Developmental Gene Regulation." *Semin. Cell Dev. Biol.*, January.

Haberle, Vanja, Nan Li, Yavor Hadzhiev, Charles Plessy, Christopher Previti, Chirag Nepal, Jochen Gehrig, et al. 2014. "Two Independent Transcription Initiation Codes Overlap on Vertebrate Core Promoters." *Nature* 507 (7492): 381–85.

Haberle, Vanja, and Alexander Stark. 2018. "Eukaryotic Core Promoters and the Functional Basis of Transcription Initiation." *Nat. Rev. Mol. Cell Biol.*, June.

Hariharan, N, and R P Perry. 1990. "Functional Dissection of a Mouse Ribosomal

Protein Promoter: Significance of the Polypyrimidine Initiator and an Element in the TATA-box Region." *Proc. Natl. Acad. Sci. U. S. A.* 87 (4): 1526–30.

Hie, Brian, Hyunghoon Cho, Bryan Bryson, and Bonnie Berger. 2019. "Coexpression Uncovers a Unified Single-Cell Transcriptomic Landscape." *bioRxiv*.

Hishiki, T, S Kawamoto, S Morishita, and K Okubo. 2000. "BodyMap: A Human and Mouse Gene Expression Database." *Nucleic Acids Res.* 28 (1): 136–38.

Hong, Sung-Kook, Michael Tsang, and Igor B Dawid. 2008. "The Mych Gene Is Required for Neural Crest Survival During Zebrafish Development." *PLoS One* 3 (4): e2029.

Hoskins, Roger A, Jane M Landolin, James B Brown, Jeremy E Sandler, Hazuki Takahashi, Timo Lassmann, Charles Yu, et al. 2011. "Genome-Wide Analysis of Promoter Architecture in Drosophila Melanogaster." *Genome Res.* 21 (2): 182–92.

Ho Sui, Shannan J, Debra L Fulton, David J Arenillas, Andrew T Kwon, and Wyeth W Wasserman. 2007. "oPOSSUM: Integrated Tools for Analysis of Regulatory Motif over-Representation." *Nucleic Acids Res.* 35 (Web Server issue): W245–52.

Hu, Peng, Mingli Liu, Dong Zhang, Jinfeng Wang, Hongbo Niu, Yimeng Liu, Zhichao Wu, et al. 2015. "Global Identification of the Genetic Networks and Cis-Regulatory Elements of the Cold Response in Zebrafish." *Nucleic Acids Res.* 43 (19): 9198–9213.

Huynh-Thu, Vân Anh, Alexandre Irrthum, Louis Wehenkel, and Pierre Geurts. 2010. "Inferring Regulatory Networks from Expression Data Using Tree-Based Methods." *PLoS One* 5 (9).

Ibarra-Soria, Ximena, Wajid Jawaid, Blanca Pijuan-Sala, Vasileios Ladopoulos, Antonio Scialdone, David J Jörg, Richard C V Tyser, et al. 2018. "Defining Murine Organogenesis at Single-Cell Resolution Reveals a Role for the Leukotriene Pathway in Regulating Blood Progenitor Formation." *Nat. Cell Biol.* 20 (2): 127–34.

Ito, M, C X Yuan, H J Okano, R B Darnell, and R G Roeder. 2000. "Involvement of the TRAP220 Component of the TRAP/SMCC Coactivator Complex in Embryonic Development and Thyroid Hormone Action." *Mol. Cell* 5 (4): 683–93.

Jin, Suk-Won, Wiebke Herzog, Massimo M Santoro, Tracy S Mitchell, Julie Frantsve, Benno Jungblut, Dimitris Beis, et al. 2007. "A Transgene-Assisted Genetic Screen Identifies Essential Regulators of Vascular Development in Vertebrate Embryos." *Dev. Biol.* 307 (1): 29–42.

Junker, Jan Philipp, Emily S Noël, Victor Guryev, Kevin A Peterson, Gopi Shah, Jan Huisken, Andrew P McMahon, Eugene Berezikov, Jeroen Bakkers, and Alexander van Oudenaarden. 2014. "Genome-Wide RNA Tomography in the Zebrafish Embryo." *Cell* 159 (3): 662–75.

Juven-Gershon, Tamar, Jer-Yuan Hsu, Joshua Wm Theisen, and James T Kadonaga. 2008. "The RNA Polymerase II Core Promoter - the Gateway to Transcription." *Curr. Opin. Cell Biol.* 20 (3): 253–59.

Kadonaga, James T. 2012. "Perspectives on the RNA Polymerase II Core Promoter." *Wiley Interdiscip. Rev. Dev. Biol.* 1 (1): 40–51.

Kellerer, Susanne, Silke Schreiner, C Claus Stolt, Stefanie Scholz, Michael R Bösl, and Michael Wegner. 2006. "Replacement of the Sox10 Transcription Factor by Sox8 Reveals Incomplete Functional Equivalence." *Development* 133 (15): 2875–86.

Khan, Aziz, Oriol Fornes, Arnaud Stigliani, Marius Gheorghe, Jaime A Castro-Mondragon, Robin van der Lee, Adrien Bessy, et al. 2018. "JASPAR 2018: Update of the Open-Access Database of Transcription Factor Binding Profiles and Its Web Framework." *Nucleic Acids Res.* 46 (D1): D260–D266.

Kinkel, Mary D, Stefani C Eames, Martha R Alonzo, and Victoria E Prince. 2008. "Cdx4 Is Required in the Endoderm to Localize the Pancreas and Limit Beta-Cell Number."

*Development* 135 (5): 919–29.

Kodzius, Rimantas, Miki Kojima, Hiromi Nishiyori, Mari Nakamura, Shiro Fukuda, Michihira Tagami, Daisuke Sasaki, et al. 2006. "CAGE: Cap Analysis of Gene Expression." *Nat. Methods* 3 (3): 211–22.

Kolde, Raivo. 2012. "Pheatmap: Pretty Heatmaps." *R Package Version* 61 (926): 915.

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

Kruse, F, J P Junker, A van Oudenaarden, and J Bakkers. 2016. "Tomo-Seq: A Method to Obtain Genome-Wide Expression Data with Spatial Resolution." *Methods Cell Biol.* 135 (February): 299–307.

Kutach, A K, and J T Kadonaga. 2000. "The Downstream Promoter Element DPE Appears to Be as Widely Used as the TATA Box in Drosophila Core Promoters." *Mol. Cell. Biol.* 20 (13): 4754–64.

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.

Lampugnani, M G, and E Dejana. 1997. "Interendothelial Junctions: Structure, Signalling and Functional Roles." *Curr. Opin. Cell Biol.* 9 (5): 674–82.

Langfelder, Peter, and Steve Horvath. 2008. "WGCNA: An R Package for Weighted Correlation Network Analysis." *BMC Bioinformatics* 9 (1): 559.

Langfelder, Peter, Bin Zhang, and Steve Horvath. 2008. "Defining Clusters from

a Hierarchical Cluster Tree: The Dynamic Tree Cut Package for R." *Bioinformatics* 24 (5): 719–20.

Langmead, Ben, and Steven L Salzberg. 2012. "Fast Gapped-Read Alignment with Bowtie 2." *Nat. Methods* 9 (4): 357–59.

Lee, Ju Yeon, and Tae-Hee Lee. 2012. "Effects of DNA Methylation on the Structure of Nucleosomes." *J. Am. Chem. Soc.* 134 (1): 173–75.

Lee, Miler T, Ashley R Bonneau, Carter M Takacs, Ariel A Bazzini, Kate R DiVito, Elizabeth S Fleming, and Antonio J Giraldez. 2013. "Nanog, Pou5f1 and SoxB1 Activate Zygotic Gene Expression During the Maternal-to-Zygotic Transition." *Nature* 503 (7476): 360–64.

Leichsenring, M, J Maes, R Mossner, W Driever, and D Onichtchouk. 2013. "Pou5f1 Transcription Factor Controls Zygotic Gene Activation in Vertebrates." *Science* 341 (6149): 1005–9.

Lenhard, Boris, Albin Sandelin, and Piero Carninci. 2012. "Metazoan Promoters: Emerging Characteristics and Insights into Transcriptional Regulation." *Nat. Rev. Genet.* 13 (4): 233–45.

Levine, Michael, and Robert Tjian. 2003. "Transcription Regulation and Animal Diversity." *Nature* 424 (6945): 147–51.

Liang, Hsiao-Lan, Chung-Yi Nien, Hsiao-Yun Liu, Mark M Metzstein, Nikolai Kirov, and Christine Rushlow. 2008. "The Zinc-Finger Protein Zelda Is a Key Activator of the Early Zygotic Genome in Drosophila." *Nature* 456 (7220): 400–403.

Lifton, R P, M L Goldberg, R W Karp, and D S Hogness. 1978. "The Organization of the Histone Genes in Drosophila Melanogaster: Functional and Evolutionary Implications." *Cold Spring Harb. Symp. Quant. Biol.* 42 Pt 2: 1047–51.

Lim, Chin Yan, Buyung Santoso, Thomas Boulay, Emily Dong, Uwe Ohler, and James T Kadonaga. 2004. "The MTE, a New Core Promoter Element for Transcription by RNA Polymerase II." *Genes Dev.* 18 (13): 1606–17.

Louder, Robert K, Yuan He, José Ramón López-Blanco, Jie Fang, Pablo Chacón, and Eva Nogales. 2016. "Structure of Promoter-Bound TFIID and Model of Human Pre-Initiation Complex Assembly." *Nature* 531 (7596): 604–9.

Love, Michael, Simon Anders, and Wolfgang Huber. 2014. "Differential Analysis of Count Data–the DESeq2 Package." *Genome Biol.* 15 (550): 10–1186.

Lundin, M, J O Nehlin, and H Ronne. 1994. "Importance of a Flanking AT-rich Region in Target Site Recognition by the GC Box-Binding Zinc Finger Protein MIG1." *Mol. Cell. Biol.* 14 (3): 1979–85.

Luo, Yuping, Volkan Coskun, Aibing Liang, Juehua Yu, Liming Cheng, Weihong Ge, Zhanping Shi, et al. 2015. "Single-Cell Transcriptome Analyses Reveal Signals to Activate Dormant Neural Stem Cells." *Cell* 161 (5): 1175–86.

Maaten, Laurens van der, and Geoffrey Hinton. 2008. "Visualizing Data Using t-SNE." *J. Mach. Learn. Res.* 9 (Nov): 2579–2605.

Mack, Katya L, Megan Phifer-Rixey, Bettina Harr, and Michael W Nachman. 2019. "Gene Expression Networks Across Multiple Tissues Are Associated with Rates of Molecular Evolution in Wild House Mice." *Genes* 10 (3).

Macosko, Evan Z, Anindita Basu, Rahul Satija, James Nemesh, Karthik Shekhar, Melissa Goldman, Itay Tirosh, et al. 2015. "Highly Parallel Genome-Wide Expression Profiling of Individual Cells Using Nanoliter Droplets." *Cell* 161 (5): 1202–14.

Maertens, Alexandra, Vy Tran, Andre Kleensang, and Thomas Hartung. 2018. "Weighted Gene Correlation Network Analysis (WGCNA) Reveals Novel Transcription Factors Associated with Bisphenol a Dose-Response." *Front. Genet.* 9 (November): 508.

Mathelier, Anthony, Oriol Fornes, David J Arenillas, Chih-Yu Chen, Grégoire Denay, Jessica Lee, Wenqiang Shi, et al. 2016. "JASPAR 2016: A Major Expansion and Update of the Open-Access Database of Transcription Factor Binding Profiles." *Nucleic Acids Res.* 44 (D1): D110–5.

Mähler, Niklas, Jing Wang, Barbara K Terebieniec, Pär K Ingvarsson, Nathaniel R Street, and Torgeir R Hvidsten. 2017. "Gene Co-Expression Network Connectivity Is an Important Determinant of Selective Constraint." *PLoS Genet.* 13 (4): e1006402.

Mifsud, Borbala, Filipe Tavares-Cadete, Alice N Young, Robert Sugar, Stefan Schoenfelder, Lauren Ferreira, Steven W Wingett, et al. 2015. "Mapping Long-Range Promoter Contacts in Human Cells with High-Resolution Capture Hi-C." *Nat. Genet.* 47 (6): 598–606.

Mortazavi, Ali, Brian A Williams, Kenneth McCue, Lorian Schaeffer, and Barbara Wold. 2008. "Mapping and Quantifying Mammalian Transcriptomes by RNA-Seq." *Nat. Methods* 5 (7): 621–28.

Mullins, M C, M Hammerschmidt, P Haffter, and C Nüsslein-Volhard. 1994. "Large-Scale Mutagenesis in the Zebrafish: In Search of Genes Controlling Development in a Vertebrate." *Curr. Biol.* 4 (3): 189–202.

Mungall, Christopher J, Carlo Torniai, Georgios V Gkoutos, Suzanna E Lewis, and Melissa A Haendel. 2012. "Uberon, an Integrative Multi-Species Anatomy Ontology." *Genome Biol.* 13 (1): R5.

Müller, Ferenc, and Làszlò Tora. 2009. "TBP2 Is a General Transcription Factor Specialized for Female Germ Cells." *J. Biol.* 8 (11): 97.

Namboori, Seema C, Patricia Thomas, Ryan Ames, Lawrence O Garrett, Craig R G Willis, Lawrence W Stanton, and Akshay Bhinge. 2019. "Single Cell Transcriptomics Identifies Master Regulators of Dysfunctional Pathways in SOD1 ALS Motor Neurons."

*bioRxiv*.

Nepal, Chirag, Yavor Hadzhiev, Christopher Previti, Vanja Haberle, Nan Li, Hazuki Takahashi, Ana Maria M Suzuki, et al. 2013. "Dynamic Regulation of the Transcription Initiation Landscape at Single Nucleotide Resolution During Vertebrate Embryogenesis." *Genome Res.* 23 (11): 1938–50.

Nes, Johan van, Wim de Graaff, Franck Lebrin, Markus Gerhard, Felix Beck, and Jacqueline Deschamps. 2006. "The Cdx4 Mutation Affects Axial Development and Reveals an Essential Role of Cdx Genes in the Ontogenesis of the Placental Labyrinth in Mice." *Development* 133 (3): 419–28.

Nikolov, D B, H Chen, E D Halay, A A Usheva, K Hisatake, D K Lee, R G Roeder, and S K Burley. 1995. "Crystal Structure of a TFIIB-TBP-TATA-element Ternary Complex." *Nature* 377 (6545): 119–28.

Nobrega, Marcelo A, Ivan Ovcharenko, Veena Afzal, and Edward M Rubin. 2003. "Scanning Human Gene Deserts for Long-Range Enhancers." *Science* 302 (5644): 413.

Nowick, Katja, Tim Gernat, Eivind Almaas, and Lisa Stubbs. 2009. "Differences in Human and Chimpanzee Gene Expression Patterns Define an Evolving Network of Transcription Factors in Brain." *Proc. Natl. Acad. Sci. U. S. A.* 106 (52): 22358–63.

Nüsslein-Volhard, C. 1994. "Of Flies and Fishes." *Science* 266 (5185): 572–74.

Odom, Duncan T, Nora Zizlsperger, D Benjamin Gordon, George W Bell, Nicola J Rinaldi, Heather L Murray, Tom L Volkert, et al. 2004. "Control of Pancreas and Liver Gene Expression by HNF Transcription Factors." *Science* 303 (5662): 1378–81.

Ohler, Uwe. 2006. "Identification of Core Promoter Modules in Drosophila and Their Application in Accurate Transcription Start Site Prediction." *Nucleic Acids Res.* 34 (20): 5943–50.

Ohler, Uwe, Guo-Chun Liao, Heinrich Niemann, and Gerald M Rubin. 2002. "Computational Analysis of Core Promoters in the Drosophila Genome." *Genome Biol.* 3 (12): RESEARCH0087.

Ohler, Uwe, David A Wassarman, B Ahsan, T L Saito, S Hashimoto, K Muramatsu, M Tsdua, et al. 2010. "Promoting Developmental Transcription." *Development* 137 (1): 15–26.

Paik, Elizabeth J, Shaun Mahony, Richard M White, Emily N Price, Anthony Dibiase, Bilguujin Dorjsuren, Christian Mosimann, Alan J Davidson, David Gifford, and Leonard I Zon. 2013. "A Cdx4-Sall4 Regulatory Module Controls the Transition from Mesoderm Formation to Embryonic Hematopoiesis." *Stem Cell Reports* 1 (5): 425–36.

Pandey, Shristi, Karthik Shekhar, Aviv Regev, and Alexander F Schier. 2018. "Comprehensive Identification and Spatial Mapping of Habenular Neuronal Types Using Single-Cell RNA-Seq." *Curr. Biol.* 28 (7): 1052–1065.e7.

Parry, Trevor J, Joshua W M Theisen, Jer-Yuan Hsu, Yuan-Liang Wang, David L Corcoran, Moriah Eustice, Uwe Ohler, and James T Kadonaga. 2010. "The TCT Motif, a Key Component of an RNA Polymerase II Transcription System for the Translational Machinery." *Genes Dev.* 24 (18): 2013–8.

Pennacchio, Len A, Nadav Ahituv, Alan M Moses, Shyam Prabhakar, Marcelo A Nobrega, Malak Shoukry, Simon Minovitsky, et al. 2006. "In Vivo Enhancer Analysis of Human Conserved Non-Coding Sequences." *Nature* 444 (7118): 499–502.

Pilon, Nicolas, Karen Oh, Jean-René Sylvestre, Nathalie Bouchard, Joanne Savory, and David Lohnes. 2006. "Cdx4 Is a Direct Target of the Canonical Wnt Pathway." *Dev. Biol.* 289 (1): 55–63.

Piras, Vincent, Masaru Tomita, and Kumar Selvarajoo. 2014. "Transcriptome-Wide Variability in Single Embryonic Development Cells." *Sci. Rep.* 4 (November): 7137.

Ponjavic, Jasmina, Boris Lenhard, Chikatoshi Kai, Jun Kawai, Piero Carninci, Yoshihide Hayashizaki, and Albin Sandelin. 2006. "Transcriptional and Structural Impact of TATA-initiation Site Spacing in Mammalian Core Promoters." *Genome Biol.* 7 (8): R78.

'PromoterOntology Github Repository.', 2019 https://github.com/Dunjanik/PromoterOntology. Accessed September 30, 2019.

Rach, Elizabeth A, Deborah R Winter, Ashlee M Benjamin, David L Corcoran, Ting Ni, Jun Zhu, and Uwe Ohler. 2011. "Transcription Initiation Patterns Indicate Divergent Strategies for Gene Regulation at the Chromatin Level." *PLoS Genet.* 7 (1): e1001274.

Rach, Elizabeth A, Hsiang-Yu Yuan, William H Majoros, Pavel Tomancak, and Uwe Ohler. 2009. "Motif Composition, Conservation and Condition-Specificity of Single and Alternative Transcription Start Sites in the Drosophila Genome." *Genome Biol.* 10 (7): R73.

Regev, Aviv, Sarah A Teichmann, Eric S Lander, Ido Amit, Christophe Benoist, Ewan Birney, Bernd Bodenmiller, et al. 2017. "The Human Cell Atlas." *Elife* 6 (December).

Reiner, Anat, Daniel Yekutieli, and Yoav Benjamini. 2003. "Identifying Differentially Expressed Genes Using False Discovery Rate Controlling Procedures." *Bioinformatics* 19 (3): 368–75.

Revelle, William, and Maintainer William Revelle. 2007. "The Psych Package."

Roepcke, Stefan, Degui Zhi, Martin Vingron, and Peter F Arndt. 2006. "Identification of Highly Specific Localized Sequence Motifs in Human Ribosomal Protein Gene Promoters." *Gene* 365 (January): 48–56.

Rotem, Assaf, Oren Ram, Noam Shoresh, Ralph A Sperling, Alon Goren, David A Weitz, and Bradley E Bernstein. 2015. "Single-Cell ChIP-seq Reveals Cell Subpopulations Defined by Chromatin State." *Nat. Biotechnol.* 33 (11): 1165–72.

Ruzicka, Leyla, Yvonne M Bradford, Ken Frazer, Douglas G Howe, Holly Paddock,

Sridhar Ramachandran, Amy Singer, et al. 2015. “ZFIN, the Zebrafish Model Organism Database: Updates and New Directions.” *Genesis* 53 (8): 498–509.

Saha, Ashis, Yungil Kim, Ariel D H Gewirtz, Brian Jo, Chuan Gao, Ian C McDowell, GTEx Consortium, Barbara E Engelhardt, and Alexis Battle. 2017. “Co-Expression Networks Reveal the Tissue-Specific Regulation of Transcription and Splicing.” *Genome Res.* 27 (11): 1843–58.

Sandelin, Albin, Wynand Alkema, Pär Engström, Wyeth W Wasserman, and Boris Lenhard. 2004. “JASPAR: An Open-access Database for Eukaryotic Transcription Factor Binding Profiles.” *Nucleic Acids Res.* 32 (suppl_1): D91–D94.

Sandelin, Albin, Peter Bailey, Sara Bruce, Pär G Engström, Joanna M Klos, Wyeth W Wasserman, Johan Ericson, and Boris Lenhard. 2004. “Arrays of Ultraconserved Non-Coding Regions Span the Loci of Key Developmental Genes in Vertebrate Genomes.” *BMC Genomics* 5 (1): 99.

Sandelin, Albin, Piero Carninci, Boris Lenhard, Jasmina Ponjavic, Yoshihide Hayashizaki, and David A Hume. 2007. “Mammalian RNA Polymerase II Core Promoters: Insights from Genome-Wide Studies.” *Nat. Rev. Genet.* 8 (6): 424–36.

Satija, Rahul, Jeffrey A Farrell, David Gennert, Alexander F Schier, and Aviv Regev. 2015. “Spatial Reconstruction of Single-Cell Gene Expression Data.” *Nat. Biotechnol.* 33 (5): 495–502.

Schier, A F. 2004. “Nodal Signaling During Gastrulation.” *See Ref. 303a*, 491–504.

Schier, Alexander F, and William S Talbot. 2005. “Molecular Genetics of Axis Formation in Zebrafish.” *Annu. Rev. Genet.* 39: 561–613.

Schug, Jonathan, Winfried-Paul Schuller, Claudia Kappen, J Michael Salbaum, Maja Bucan, and Christian J Stoeckert Jr. 2005. “Promoter Features Related to Tissue Specificity as Measured by Shannon Entropy.” *Genome Biol.* 6 (4): R33.

Shi, Junchao, Qi Chen, Xin Li, Xiudeng Zheng, Ying Zhang, Jie Qiao, Fuchou Tang, Yi Tao, Qi Zhou, and Enkui Duan. 2015. "Dynamic Transcriptional Symmetry-Breaking in Pre-Implantation Mammalian Embryo Development Revealed by Single-Cell RNA-seq." *Development* 142 (20): 3468–77.

Shimizu, Takashi, Young-Ki Bae, Osamu Muraoka, and Masahiko Hibi. 2005. "Interaction of Wnt and Caudal-Related Genes in Zebrafish Posterior Body Formation." *Dev. Biol.* 279 (1): 125–41.

Sloutskin, Anna, Yehuda M Danino, Yaron Orenstein, Yonathan Zehavi, Tirza Doniger, Ron Shamir, and Tamar Juven-Gershon. 2015. "ElemeNT: A Computational Tool for Detecting Core Promoter Elements." *Transcription* 6 (3): 41–50.

Smale, S T. 1997. "Transcription Initiation from TATA-less Promoters Within Eukaryotic Protein-Coding Genes." *Biochim. Biophys. Acta* 1351 (1-2): 73–88.

Smale, Stephen T, and David Baltimore. 1989. "The 'Initiator' as a Transcription Control Element." *Cell* 57 (1): 103–13.

Smedley, Damian, Syed Haider, Steffen Durinck, Luca Pandini, Paolo Provero, James Allen, Olivier Arnaiz, et al. 2015. "The BioMart Community Portal: An Innovative Alternative to Large, Centralized Data Repositories." *Nucleic Acids Res.* 43 (W1): W589–98.

Solnica-Krezel, L, A F Schier, and W Driever. 1994. "Efficient Recovery of ENU-induced Mutations from the Zebrafish Germline." *Genetics* 136 (4): 1401–20.

Sonawane, Abhijeet Rajendra, John Platig, Maud Fagny, Cho-Yi Chen, Joseph Nathaniel Paulson, Camila Miranda Lopes-Ramos, Dawn Lisa DeMeo, John Quackenbush, Kimberly Glass, and Marieke Lydia Kuijjer. 2017. "Understanding Tissue-Specific Gene Regulation." *Cell Rep.* 21 (4): 1077–88.

Stormo, G D, T D Schneider, L Gold, and A Ehrenfeucht. 1982. "Use of the 'Perceptron' Algorithm to Distinguish Translational Initiation Sites in E. Coli." *Nucleic Acids*

*Res.* 10 (9): 2997–3011.

Su, Andrew I, Tim Wiltshire, Serge Batalov, Hilmar Lapp, Keith A Ching, David Block, Jie Zhang, et al. 2004. "A Gene Atlas of the Mouse and Human Protein-Encoding Transcriptomes." *Proc. Natl. Acad. Sci. U. S. A.* 101 (16): 6062–7.

Subramanian, Aravind, Pablo Tamayo, Vamsi K Mootha, Sayan Mukherjee, Benjamin L Ebert, Michael A Gillette, Amanda Paulovich, et al. 2005. "Gene Set Enrichment Analysis: A Knowledge-Based Approach for Interpreting Genome-Wide Expression Profiles." *Proc. Natl. Acad. Sci. U. S. A.* 102 (43): 15545–50.

Sun, Yujia, Chung-Yi Nien, Kai Chen, Hsiao-Yun Liu, Jeff Johnston, Julia Zeitlinger, and Christine Rushlow. 2015. "Zelda Overcomes the High Intrinsic Nucleosome Barrier at Enhancers During Drosophila Zygotic Genome Activation." *Genome Res.* 25 (11): 1703–14.

Tadros, Wael, and Howard D Lipshitz. 2009. "The Maternal-to-Zygotic Transition: A Play in Two Acts." *Development* 136 (18): 3033–42.

Takahashi, Kazutoshi, and Shinya Yamanaka. 2016. "A Decade of Transcription Factor-Mediated Reprogramming to Pluripotency." *Nat. Rev. Mol. Cell Biol.* 17 (3): 183–93.

Tan, G. 2015. "JASPAR2016: Data Package for JASPAR 2016."

Tan, Ge, and Boris Lenhard. 2016. "TFBSTools: An R/Bioconductor Package for Transcription Factor Binding Site Analysis." *Bioinformatics* 32 (10): 1555–6.

Tang, Fuchou, Catalin Barbacioru, Yangzhou Wang, Ellen Nordman, Clarence Lee, Nanlan Xu, Xiaohui Wang, et al. 2009. "mRNA-Seq Whole-Transcriptome Analysis of a Single Cell." *Nat. Methods* 6 (5): 377–82.

Tarifeño-Saldivia, Estefania, Arnaud Lavergne, Alice Bernard, Keerthana Padamata, David Bergemann, Marianne L Voz, Isabelle Manfroid, and Bernard Peers. 2017. "Transcriptome Analysis of Pancreatic Cells Across Distant Species Highlights Novel Important

Regulator Genes." *BMC Biol.* 15 (1): 21.

Teles, Jose, Cristina Pina, Patrik Edén, Mattias Ohlsson, Tariq Enver, and Carsten Peterson. 2013. "Transcriptional Regulation of Lineage Commitment–a Stochastic Model of Cell Fate Decisions." *PLoS Comput. Biol.* 9 (8): e1003197.

"The Zebrafish Information Network - Gene Expression Data." 2019. *Https://Zfin.org/Action/Expression/Search*, September.

Thisse, Bernard, Vincent Heyer, Aline Lux, Violaine Alunni, Agnès Degrave, Iban Seiliez, Johanne Kirchner, Jean-Paul Parkhill, and Christine Thisse. 2001. "Expression of the Zebrafish Genome During Embryogenesis (Nih R01 Rr15402)." *ZFIN Direct Data Submission.*

Thisse, Christine, and Bernard Thisse. 2008. "High-Resolution in Situ Hybridization to Whole-Mount Zebrafish Embryos." *Nat. Protoc.* 3 (1): 59–69.

Tomancak, Pavel, Benjamin P Berman, Amy Beaton, Richard Weiszmann, Elaine Kwan, Volker Hartenstein, Susan E Celniker, and Gerald M Rubin. 2007. "Global Analysis of Patterns of Gene Expression During Drosophila Embryogenesis." *Genome Biol.* 8 (7): R145.

Trapnell, Cole, Adam Roberts, Loyal Goff, Geo Pertea, Daehwan Kim, David R Kelley, Harold Pimentel, Steven L Salzberg, John L Rinn, and Lior Pachter. 2012. "Differential Gene and Transcript Expression Analysis of RNA-seq Experiments with TopHat and Cufflinks." *Nat. Protoc.* 7 (3): 562–78.

Trinklein, Nathan D, Shelley Force Aldred, Sara J Hartman, Diane I Schroeder, Robert P Otillar, and Richard M Myers. 2004. "An Abundance of Bidirectional Promoters in the Human Genome." *Genome Res.* 14 (1): 62–66.

Van Slyke, Ceri E, Yvonne M Bradford, Monte Westerfield, Melissa A Haendel, D M Medeiros, J G Crump, L Solnica-Krezel, et al. 2014. "The Zebrafish Anatomy and Stage Ontologies: Representing the Anatomy and Development of Danio Rerio." *J. Biomed. Semantics* 5 (1): 12.

Vastenhouw, Nadine L, Wen Xi Cao, and Howard D Lipshitz. 2019. "The Maternal-to-Zygotic Transition Revisited." *Development* 146 (11).

Visel, Axel, Simon Minovitsky, Inna Dubchak, and Len A Pennacchio. 2007. "VISTA Enhancer Browser–a Database of Tissue-Specific Human Enhancers." *Nucleic Acids Res.* 35 (Database issue): D88–92.

Wagner, Daniel E, Caleb Weinreb, Zach M Collins, James A Briggs, Sean G Megason, and Allon M Klein. 2018. "Single-Cell Mapping of Gene Expression Landscapes and Lineage in the Zebrafish Embryo." *Science* 360 (6392): 981–87.

Wasserman, Wyeth W, and Albin Sandelin. 2004. "Applied Bioinformatics for the Identification of Regulatory Elements." *Nat. Rev. Genet.* 5 (4): 276–87.

White, Richard J, John E Collins, Ian M Sealy, Neha Wali, Christopher M Dooley, Zsofia Digby, Derek L Stemple, et al. 2017. "A High-Resolution mRNA Expression Time Course of Embryonic Development in Zebrafish." *Elife* 6 (November).

Wickham, Hadley. 2006. "Ggplot: An Implementation of the Grammar of Graphics in R, 2006." *R Package Version 0. 4. 0*, January.

Wolpert, L. 1994. "Do We Understand Development?" *Science* 266 (5185): 571–72.

Woolfe, Adam, Debbie K Goode, Julie Cooke, Heather Callaway, Sarah Smith, Phil Snell, Gayle K McEwen, and Greg Elgar. 2007. "CONDOR: A Database Resource of Developmentally Associated Conserved Non-Coding Elements." *BMC Dev. Biol.* 7 (August): 100.

Woolfe, Adam, Martin Goodson, Debbie K Goode, Phil Snell, Gayle K McEwen, Tanya Vavouri, Sarah F Smith, et al. 2004. "Highly Conserved Non-Coding Sequences Are Associated with Vertebrate Development." Edited by Sean Eddy. *PLoS Biol.* 3 (1): e7.

Xiao, Fei, Lin Gao, Yusen Ye, Yuxuan Hu, and Ruijie He. 2016. "Inferring Gene

Regulatory Networks Using Conditional Regulation Pattern to Guide Candidate Genes." *PLoS One* 11 (5): e0154953.

Xue, Zhigang, Kevin Huang, Chaochao Cai, Lingbo Cai, Chun-Yan Jiang, Yun Feng, Zhenshan Liu, et al. 2013. "Genetic Programs in Human and Mouse Early Embryos Revealed by Single-Cell RNA Sequencing." *Nature* 500 (7464): 593–97.

Yang, Robert Y, Jie Quan, Reza Sodaei, Francois Aguet, Ayellet V Segrè, John A Allen, Thomas A Lanz, et al. 2018. "A Systematic Survey of Human Tissue-Specific Gene Expression and Splicing Reveals New Opportunities for Therapeutic Target Identification and Evaluation." *bioRxiv*.

Yu, Guangchuang, Fei Li, Yide Qin, Xiaochen Bo, Yibo Wu, and Shengqi Wang. 2010. "GOSemSim: An R Package for Measuring Semantic Similarity Among GO Terms and Gene Products." *Bioinformatics* 26 (7): 976–78.

Yu, Guangchuang, Li-Gen Wang, Yanyan Han, and Qing-Yu He. 2012. "ClusterProfiler: An R Package for Comparing Biological Themes Among Gene Clusters." *OMICS* 16 (5): 284–87.

Yuan, Kai, Charles A Seller, Antony W Shermoen, and Patrick H O'Farrell. 2016. "Timing the Drosophila Mid-Blastula Transition: A Cell Cycle-Centered View." *Trends Genet.* 32 (8): 496–507.

Zabidi, Muhammad A, Cosmas D Arnold, Katharina Schernhuber, Michaela Pagani, Martina Rath, Olga Frank, and Alexander Stark. 2015. "Enhancer-Core-Promoter Specificity Separates Developmental and Housekeeping Gene Regulation." *Nature* 518 (7540): 556–59.

Zacchigna, Serena, Carmen Ruiz de Almodovar, and Peter Carmeliet. 2008. "Similarities Between Angiogenesis and Neural Development: What Small Animal Models Can Tell Us." *Curr. Top. Dev. Biol.* 80: 1–55.

'Zfexpress - Anatomical Specificity Interactive Web Session.', 2019 http://zfexpress.

genereg.net:3838/dunja/AnatomicalSpecificity/. Accessed September 30, 2019.

Zhang, H, and J-K Zhu. 2012. "Active DNA Demethylation in Plants and Animals." *Cold Spring Harb. Symp. Quant. Biol.* 77 (November): 161–73.

Zhang, Yong, Zarmik Moqtaderi, Barbara P Rattner, Ghia Euskirchen, Michael Snyder, James T Kadonaga, X Shirley Liu, and Kevin Struhl. 2009. "Intrinsic histone-DNA Interactions Are Not the Major Determinant of Nucleosome Positions in Vivo." *Nat. Struct. Mol. Biol.* 16 (8): 847–52.

Zhang, Yuqing, Sylvia Bonilla, Leelyn Chong, and Yuk Fai Leung. 2013. "Irx7, a Smarca4-Regulated Gene for Retinal Differentiation, Regulates Other Genes Controlled by Smarca4 in Zebrafish Retinas." *Gene Expr. Patterns* 13 (5-6): 177–82.