



# Bioinformatic Reconstruction of Gene Regulatory Networks Controlling EMT and Mesoderm Formation

Jinhua Liu

Dissertation zur Erlangung des Grades  
eines Doktors der Naturwissenschaften (Dr. rer. nat.)  
am Fachbereich Mathematik und Informatik  
der Freien Universität Berlin

Berlin, 2019

Erstgutachter: Prof. Dr. Martin Vingron

Zweitgutachter: Prof. Dr. Bernhard Herrmann

Tag der Disputation: 22.08.2019

## Acknowledgments

Upon completion of my PhD thesis, my deepest gratitude goes firstly to Professor Bernhard Herrmann. I would like to thank him for offering me the chance to study and work in his department, for being my supervisor and supporting me with valuable ideas, discussions and advice. I appreciate his time and patience to help me get into the field of developmental genetics and expand my knowledge.

I am very grateful to Professor Martin Vingron. I made my final decision to pursue my PhD when I was visiting his group years ago. I sincerely thank him for giving me the opportunity to enroll at Freie Universität, for being my mentor of my project, and helping me through meetings and suggestions.

I would like to offer my special thanks to my colleagues and collaborators Pavel Tsaytler and Frederic Koch. I highly appreciate their time for discussions and suggestions to my project and for proofreading my thesis. Dr. Pavel Tsaytler has performed all wet lab experiments and provided the data. I am grateful to him for his important contribution. Advice given by my colleague Jesse Veenliet has been of great help. My thanks are extended to him and to Gaby Bläß who has worked on the experimental part. I also thank the members from the sequencing group for performing the high-throughput sequencing assays.

I would like to thank my friends and my colleagues who have helped me and shared their experience and wisdom with me. They helped me out through many troubles I encountered during my PhD life as an international student.

Last but not least, I am deeply grateful to my family for their continuous love and support.

Jinhua Liu

2019-03-07 Berlin





## Abstract

Embryonic development is a complex multi-stage process, which at the gene expression level requires precise control by gene regulatory networks (GRNs). At each stage of pattern formation and organogenesis, during the transition of precursor cells to their descendants, various sets of signaling molecules and transcription factors (TFs) activate or repress their target genes to determine distinct cell fates. Misregulation of developmental pathways may cause severe diseases or lethality, while their ectopic activation in the adult organism often results in oncogenic transformation. It is therefore of great importance to decode the transcription factors and understand how they interact and form GRNs controlling developmental processes.

Mesoderm formation is vital for embryo development. It occurs during gastrulation and depends on the process of epithelial-mesenchymal transition (EMT). In vertebrates, mesoderm gives rise to various tissues, such as axial skeleton, skeletal muscle, heart, kidney, smooth muscles, blood vessels and blood. A plethora of studies has been focused on characterizing the genes that regulate the development of mesoderm. Signaling pathways including WNT, BMP and FGF, along with transcription factors such as Smads, Eomes and T have been reported to play fundamental roles in this process. However, the comprehensive mechanistic characterization of the mesodermal GRNs is still lacking.

This study aims at constructing a global gene regulatory network, which describes transcriptional regulatory events occurring dynamically during the course of mesoderm formation in the mouse. We demonstrated that *in vitro* mesodermal differentiation of mouse embryonic stem cells mimics mesoderm formation *in vivo*, and therefore chose it as a model system. Firstly, by combining ChIP-seq and RNA-seq techniques, I reconstructed GRNs mediated by the essential mesodermal TFs Smads, Eomes and T. Next, to build global dynamic GRN orchestrating EMT and mesoderm formation, time-series gene expression and TF-target datasets were integrated. The latter was obtained by an original method of discovering functionally active TFs from ATAC-seq data, followed by their association with putative target genes. Combing this method with a bioinformatical tool based on hidden Markov model allowed me to identify groups of co-expressed genes from time-series transcriptome data and predict TFs that regulate their expression.

The predictive power of this approach was validated by comparing its output with the Smads, Eomes and T datasets, demonstrating that it correctly assigned these TFs to their targets. Using this unbiased approach, novel candidate mesodermal TFs and target genes of previously known TFs were identified. This study expands our understanding of genetic regulation mechanisms underlying EMT and mesoderm formation in the mouse and provides a list of novel potential mesoderm regulators for future in-depth characterization. This bioinformatical approach thus is promising in future studies designed to characterize the molecular mechanism underlying specific developmental processes.



# Contents

|          |   |           |
|----------|---|-----------|
| <b>1</b> | <b>Introduction</b> .....   | <b>1</b>  |
| 1.1      | Transcription Factors and Transcriptional Regulation.....                                       | 1         |
| 1.2      | Chromatin Structure and Transcription.....  | 2         |
| 1.3      | Mesoderm Formation <i>in Vivo</i> .....   | 3         |
| 1.3.1    | Mouse Embryogenesis.....  | 3         |
| 1.3.2    | Blastulation.....   | 3         |
| 1.3.3    | Implantation and Axis Formation.....  | 6         |
| 1.3.4    | Gastrulation.....   | 6         |
| 1.3.5    | Mesoderm Formation through EMT during Gastrulation.....   | 6         |
| 1.4      | Master Regulators of Mesoderm Formation.....  | 10        |
| 1.5      | Studies of <i>in Vitro</i> Mesoderm Formation.....  | 13        |
| 1.6      | Gene Regulatory Networks.....   | 14        |
| <b>2</b> | <b>Experimental Material and Methods</b> .....  | <b>17</b> |
| 2.1      | Mesodermal Differentiation <i>In Vitro</i> .....  | 17        |
| 2.2      | Chromatin Immunoprecipitation (ChIP).....   | 17        |
| 2.3      | Chromatin Immunoprecipitation Followed by Sequencing (ChIP-seq).....                            | 19        |
| 2.4      | RNA Sequencing (RNA-seq).....   | 20        |
| 2.5      | Assay for Transposase-accessible Chromatin using Sequencing (ATAC-seq).....                     | 21        |
| 2.6      | Datasets for Analysis.....  | 23        |
| <b>3</b> | <b>Computational Methods</b> .....  | <b>24</b> |
| 3.1      | Correlations: Pearson vs. Spearman.....   | 24        |
| 3.2      | Fisher’s Exact Test for Enrichment Analysis.....  | 25        |
| 3.3      | Correcting for Multiple Testing.....  | 26        |
| 3.4      | Maximum Likelihood Estimation.....  | 27        |
| 3.5      | Expectation Maximization Algorithm.....   | 28        |
| 3.6      | Hidden Markov Model (HMM) and Corresponding Algorithms.....                                     | 29        |
| 3.7      | “Dynamic Regulatory Events Miner (DREM)” Based on Input-Output Hidden Markov Model (IOHMM)..... | 37        |
| 3.8      | Read Mapping: Hash Table or Burrows-Wheeler Transform.....                                      | 42        |
| 3.9      | ChIP-seq Analysis.....  | 44        |

|          |  |            |
|----------|--|------------|
| 3.9.1    | Peak Calling.....  | 44         |
| 3.9.2    | Association of Peaks to Genes.....   | 47         |
| 3.10     | RNA-seq Analysis.....  | 48         |
| 3.11     | ATAC-seq Analysis.....   | 48         |
| 3.12     | Motif Analysis.....  | 50         |
| 3.13     | Gene Ontology (GO) Term Enrichment Analysis.....   | 50         |
| 3.14     | Software.....  | 52         |
| <b>4</b> | <b>Results .....</b>   | <b>53</b>  |
| 4.1      | Time-series Transcriptome Analysis of Mesoderm Formation <i>in Vitro</i> .....             | 53         |
| 4.1.1    | Differential Gene Expression Analysis and Clustering.....                                  | 53         |
| 4.1.2    | Sub-Cluster Analysis.....  | 58         |
| 4.2      | Gene Regulation by Transcription Factors Smads, Eomes and T during Mesoderm Formation..... | 62         |
| 4.2.1    | Smads.....   | 63         |
| 4.2.2    | Eomes.....   | 67         |
| 4.2.3    | T .....  | 70         |
| 4.3      | Reconstruction of the Dynamic Regulatory Network Underlying Mesoderm Formation.....        | 73         |
| 4.3.1    | Inferring TF targets from ATAC-seq Data.....   | 74         |
| 4.3.2    | Reconstructing the Dynamic Regulatory Network Controlling Mesoderm Formation.....          | 79         |
| 4.3.3    | Validation of the Gene Regulatory Network .....  | 93         |
| <b>5</b> | <b>Discussion .....</b>  | <b>97</b>  |
| <b>6</b> | <b>Bibliography .....</b>  | <b>110</b> |
| <b>7</b> | <b>Appendices .....</b>  | <b>119</b> |
| A.       | Supplementary Figures .....  | 119        |
| B.       | Supplementary Tables.....  | 133        |
| C.       | Supplementary Notes.....   | 138        |
| D.       | List of Figures .....  | 141        |
| E.       | List of Tables.....  | 143        |
| F.       | Abbreviations .....  | 144        |

# 1 Introduction

## 1.1 Transcription Factors and Transcriptional Regulation

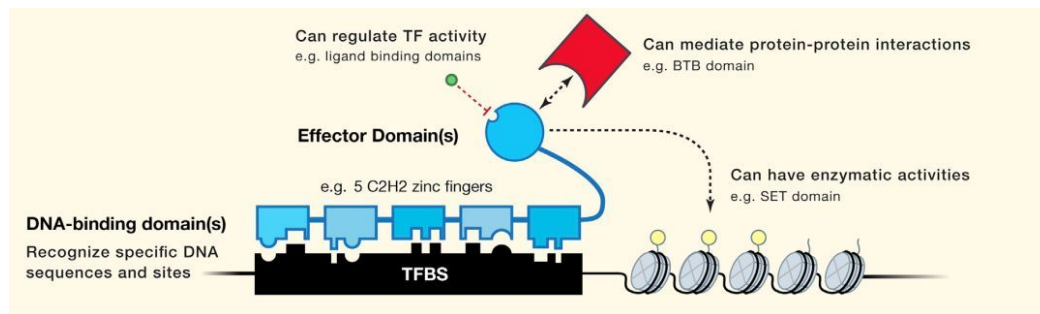
Transcription factors (TFs) are proteins which can bind specific DNA sequences and regulate the process of gene transcription. By determining when and where to switch specific genes on and off, TFs control the differentiation of cells and the formation of tissues<sup>1,2</sup>. Mutations in TFs or TF-binding sequences may lead to dysregulation of gene expression and are related to a diverse set of diseases.

TFs have caught the interest of researchers for more than three decades<sup>3</sup>. As more TFs are identified, their structures and the mechanisms underlying their functions get better understood. Typically, a TF contains a DNA-binding domain and multiple effector domains. The DNA-binding domain is necessary for a TF to recognize and bind to a specific DNA sequence. This binding sequence, in eukaryotic cells, is usually located at the promoter or enhancer regions. An effector domain can either bind ligands to regulate TF activity via external signals or mediate protein-protein interactions to recruit cofactors (Figure 1.1)<sup>2</sup>. Upon binding to the DNA, TFs can directly recruit RNA polymerase II (Pol II). However, in eukaryotic cells, most of the TFs are shown to function by firstly recruiting cofactors<sup>2,4</sup> and influence the activity of Pol II indirectly. The DNA sequences bound by distinct TFs are specified by their corresponding motifs. A motif of a given TF is the sequence pattern determined based on the binding sites which this TF prefers to occupy. Motifs can be used to scan the genomic regions to search for the candidate binding sites for specific TFs. A number of tools for motif discovery and analysis have been developed, including MEME, Homer, etc.<sup>5</sup>

TFs regulate transcription by either activating or repressing gene expression, and therefore they are defined as transcriptional activators or repressors. As mentioned above, TFs can directly or indirectly recruit Pol II to activate transcription. To repress transcription, a TF can block the DNA binding sites of Pol II or other activators. Furthermore, depending on the cell type and environmental cues, a given TF can either be an activator or a repressor of the same gene. Unlike in prokaryotic cells where transcription is usually regulated by single proteins, eukaryotic transcriptional regulation is performed by the combination of multiple proteins. TFs can bind to multiple cofactors in different ways. The multitude of the TFs'

combinations enriches the diversity of gene expression, thus fulfilling the higher demand of gene regulation in eukaryotic cells<sup>6</sup>.

Studies have shown that many diseases are linked to mutations in transcription factor binding sites, TFs and cofactors. These diseases range from developmental disorders, diabetes, cardiovascular disease to cancer. It is therefore very important to annotate all of the TFs and investigate how they function in different cell types<sup>7</sup>.



**Figure 1.1 The typical structure of a TF**

Typically, a TF contains a DNA-binding domain, which is needed to recognize a specific DNA sequence and then bind to it, and multiple effector domains. Figure taken from Lambert *et al.* (2018)<sup>2</sup>.

## 1.2 Chromatin Structure and Transcription

In eukaryotic cells, DNA is packed by histones into repeating structural units called nucleosomes. Each nucleosome contains 147 bp of DNA wrapped  $\sim 1.7$  turns around an octamer of the histone proteins H2A, H2B, H3 and H4<sup>8</sup>. Those proteins can be replaced by histone variants or modified by enzymes to add or remove post-translational modifications, through which the architecture of nucleosome changes to open or close the chromatin<sup>9</sup>. Open chromatin is the prerequisite for RNA polymerase binding to DNA. Likewise, most of the TFs, with the exception of pioneer factors, can only bind to open DNA regions<sup>10</sup>.

The accessibility of chromatin can be detected by different assays. DNase-seq (DNase I hypersensitive sites sequencing)<sup>11</sup>, FAIRE-seq (Formaldehyde assisted isolation of regulatory elements sequencing)<sup>12</sup> and ATAC-seq (Assay for transposase-accessible chromatin using sequencing)<sup>13</sup> can directly isolate open genomic regions, while MNase-seq (Direct sequencing following MNase digestion)<sup>14</sup> locates nucleosomes and thereby indirectly assesses the DNA accessibility. Among these methods, ATAC-seq is the most recently established one. The

open regions detected by it can be used for motif analysis to find enriched motifs and corresponding TFs<sup>15</sup>.

### **1.3 Mesoderm Formation *in Vivo***

#### **1.3.1 Mouse Embryogenesis**

Embryogenesis refers to the process of embryo formation. In mouse it can be generally divided into successive stages including: blastulation, implantation, axis formation and gastrulation. Mesoderm formation happens during gastrulation.

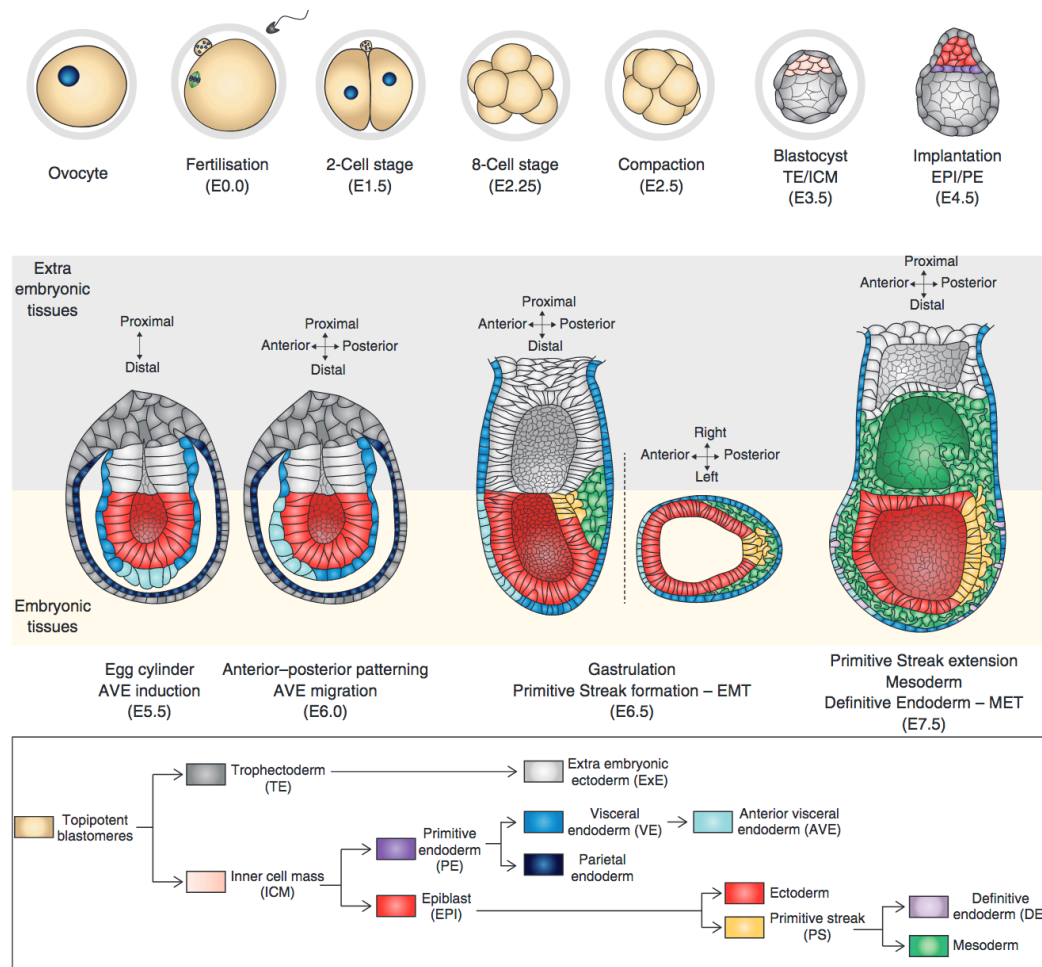
#### **1.3.2 Blastulation**

Embryo formation starts from the process of fertilization when the male and female gametes fuse into one cell called zygote (E0.0). Then the zygote starts to undergo consecutive cell divisions. The divisions up to eight cells are symmetric and the cells show no morphological difference. From the eight to sixteen cells stage, a process called compaction (E2.25) takes place. During compaction, the cells start to bind tightly and establish the communication with each other, resulting in the formation of morula at the sixteen cells stage. Compaction is the first step where the blastomeres show morphological differences. From fertilization to morula formation, the cell division increases the number of cells without increasing the size of the embryo. This special cell division stage is defined as cleavage. Subsequently, the morula develops into a blastocyst containing two cell lineages: outer cell layer trophoblast (TE) and the inner cells called inner cell mass (ICM). The ICM differentiates into epiblast (EPI) and primitive endoderm (PE) at E4.5 (Figure 1.2)<sup>16</sup>.

#### **Molecular Mechanisms of Cell Lineage Commitment during Blastulation**

Segregation of different cell fates is specified by transcriptional regulation. During embryogenesis, the cells are totipotent until the process of compaction. Then the transcription factors Pou5f1 (also known as Oct3/4) and Cdx2 take the key roles to mediate the differentiation from the compacted morula to the early stage of blastocyst with two cell layers TE and ICM. Pou5f1 and Cdx2 are co-expressed in all cells of morula. Later in some cells the expression of Cdx2 increases, which inhibits the expression of Pou5f1. On the contrary, in the inner cells Pou5f1 represses Cdx2. The reciprocal exclusive expression of Pou5f1 and Cdx2 leads to the formation of TE (larger outer cells) and ICM (smaller inner

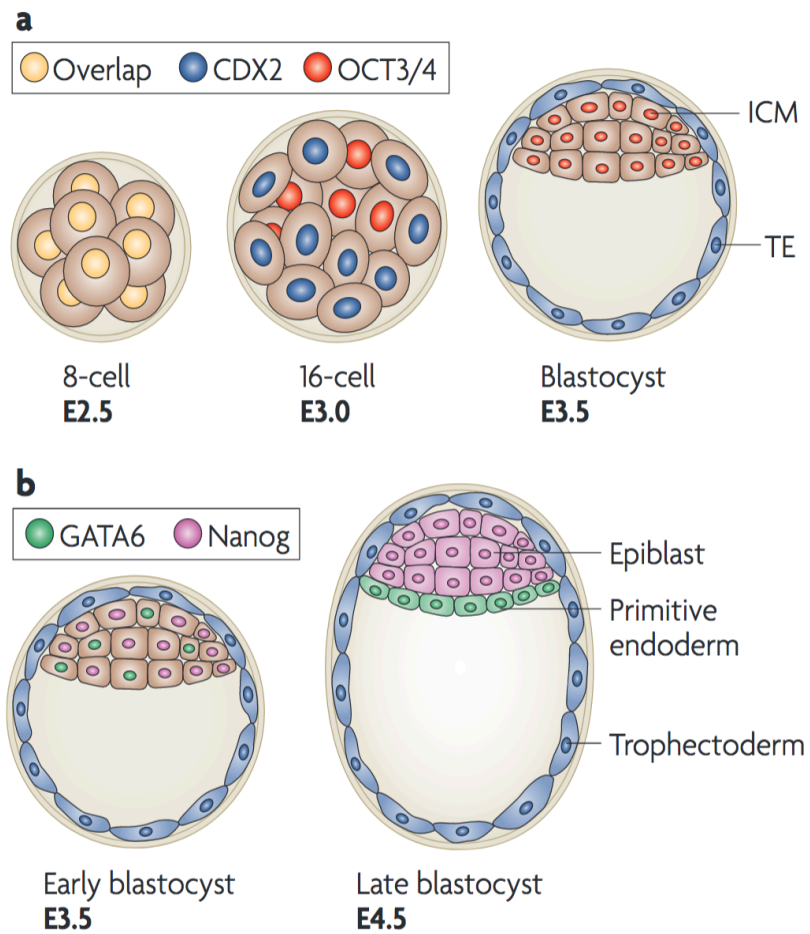
cells). Subsequently, at around E3.5 the transcription factors *Gata6* and *Nanog* start to be expressed in the ICM cells. The *Gata6*-positive cells develop into the primitive endoderm (PE) at the surface of ICM, while the *Nanog*-positive cells form the epiblast (EPI) (Figure 1.3)<sup>17</sup>. The EPI later generates the embryonic body.



**Figure 1.2 The process of mouse embryogenesis**

After fertilization, cells are totipotent until embryo compaction. Then, the blastocyst with two cell lineages forms: the outer cell layer trophoctoderm (TE, in grey) and the inner cells called inner cell mass (ICM, in beige). Later ICM is differentiated into epiblast (EPI, in red) and primitive endoderm (PE, in purple) which gives rise to the parietal endoderm (in dark blue) and the visceral endoderm (in blue). Visceral endoderm is the progenitor of anterior visceral endoderm (AVE, in light blue). At around E6.5, gastrulation occurs and three germ layers, ectoderm, mesoderm and endoderm, are formed. Figure taken from Nahaboo & Migeotte (2018)<sup>16</sup>.





**Figure 1.3 Cell lineage commitment during blastulation**

Pou5f1 (shown in the figure as Oct3/4, in red) and Cdx2 (in blue) are co-expressed in all cells of the morula. Later the increasing expression of Cdx2 inhibits the expression of Pou5f1 in some cells. On the contrary, in the inner cells Pou5f1 represses Cdx2. The reciprocal exclusive expression of these two TFs leads to the result that the outer larger cells form TE and the inner smaller cells form ICM. Subsequently at around E3.5, TFs Gata6 and Nanog start to express in the ICM cells. The Gata6-positive cells develop into the primitive endoderm (PE) at the surface of ICM, while the Nanog-positive cells form the epiblast (EPI). Figure taken from Arnold & Robertson (2009)<sup>17</sup>.

### **1.3.3 Implantation and Axis Formation**

After the late blastocyst with three cell lineages is established, the TE attaches to the uterine epithelium. The success of implantation is required for the embryo to receive nutrients from the mother. Around the time of implantation, the conceptus elongates along its proximal-distal (P-D) axis to form the “egg cylinder” embryo. This morphological change defines the embryonic pattern formation and is required for further establishment of different cell lineages. The extraembryonic ectoderm (ExE) derived from TE forms a layer of epithelial cells at the proximal site of the conceptus and combines with the distal epithelialized EPI. The PE differentiates into the parietal endoderm and the visceral endoderm, which belong to extraembryonic tissues. The parietal endoderm in the end develops into part of the parietal yolk sac, while the visceral endoderm is the progenitor of anterior visceral endoderm (AVE). Following the establishment of P-D axis, the anterior-posterior (A-P) axis is established along the process of AVE migration from distal embryo to the side of the prospective anterior (Figure 1.2).

### **1.3.4 Gastrulation**

During gastrulation, three germ layers differentiate from one single cell layer of the epiblast (Figure 1.2). Those three germ layers—ectoderm, mesoderm and definitive endoderm—are the progenitors of all embryonic body structures. The ectoderm forms the nervous system (brain, spinal cord and peripheral nervous system), epidermis and many sensory organs. The mesoderm gives rise to muscles, bones, connective tissues, blood, heart, kidney and reproductive system. The definitive endoderm produces the gastrointestinal tract and its future derivatives.

### **1.3.5 Mesoderm Formation through EMT during Gastrulation**

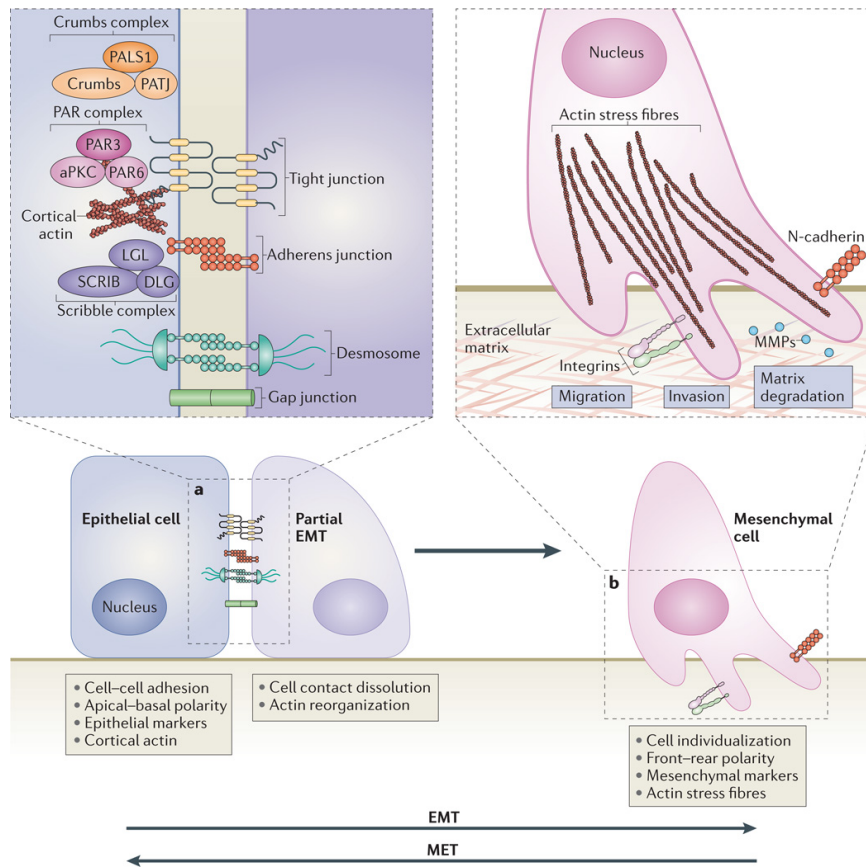
During gastrulation, mesoderm and definitive endoderm are formed through the process of epithelial-mesenchymal transition (EMT)<sup>17</sup>. Epithelial cells are tightly connected and polarized. EMT is the biological process characterized by the loss of the epithelial cells' connections and polarity and their subsequent ability to migrate as mesenchymal cells. This process is involved in cell differentiation, tissue formation, wound healing and cancer<sup>18</sup>.

As shown in Figure 1.4, tight junctions, adherens junctions, desmosomes and gap junctions keep the epithelial cells adhere to each other. The Crumbs, partitioning defective (PAR) and Scribble (SCRIB) complexes are involved in maintaining cell polarity. EMT is accompanied by the disruption of those complexes and junctions and the rearrangement of the original cytoskeleton. This results in the formation of mesenchymal cells, which upon maturation acquire migration ability.

EMT is a complex process, which involves many TFs and signaling pathways. During EMT, the epithelial markers (Claudins, Occludin, E-cadherin, Desmoplakin, ZO1) get repressed while the expression of mesenchymal markers (Fibronectin, Vitronectin, N-cadherin) is up-regulated. The main EMT TFs include the SNAIL and ZEB families. SNAIL directly inhibits the expression of Claudin, Occludin and E-cadherin thereby destroying cell-cell connection. On the other hand, it activates the expression of mesenchymal marker N-cadherin. ZEB factors also repress E-cadherin and activate N-cadherin, as well as up-regulate matrix metalloproteinases (MMPs) which help to enable cell invasion. The signaling pathways involved in EMT include TGF $\beta$ , FGF, HGF, EGF, WNT and NOTCH<sup>18</sup>.

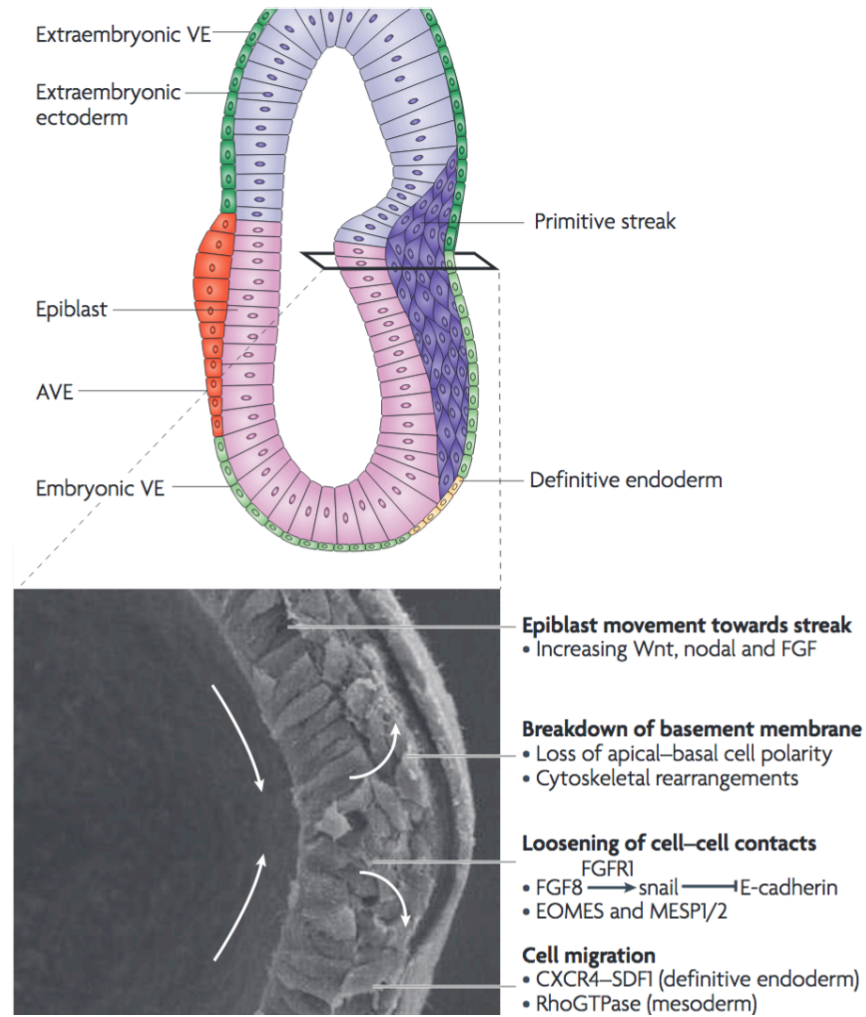
At around E6.0, the epiblast cells converge at the proximal junction of ExE and EPI to form the primitive streak (PS), which can be identified as the first sign of gastrulation.

As shown in Figure 1.2 and Figure 1.5, mesoderm and definitive endoderm are formed through the process of EMT from epithelial cells of the primitive streak<sup>17</sup>. Depending on the time and site of the cells migrating from the primitive streak, different cell fates are initiated. Several essential mesodermal TFs (e.g., Smads, Eomes, T, and Mesp1/2) and signaling pathways (e.g., FGF and WNT) have been identified<sup>19,20</sup>.



**Figure 1.4 Cellular mechanism of EMT**

Epithelial cells are attached to each other via tight junctions, adherens junctions, desmosomes and gap junctions. The Crumbs, partitioning defective (PAR) and Scribble (SCRIB) complexes maintain their polarity. Upon EMT, these complexes and junctions are disrupted, followed by the rearrangement of the cytoskeleton. Upon maturation, the newly formed mesenchymal cells acquire motility. Figure taken from Lamouille *et al.* (2014)<sup>18</sup>.



**Figure 1.5 Mesoderm formation through EMT during gastrulation**

Mesoderm formation is a result of EMT. When epiblast cells move towards primitive streak, the increasing WNT, Nodal and FGF signals change the cell behaviour. Cells in the primitive streak lose apical-basal polarity and obtain the ability to migrate. The whole process involves many regulators. E-cadherin is downregulated via a FGF signaling cascade, allowing cells to migrate. Eomes has been shown to be a Nodal target to influence EMT. Mesp1/2 are also required for EMT. Figure taken from Arnold & Robertson (2009)<sup>17</sup>.

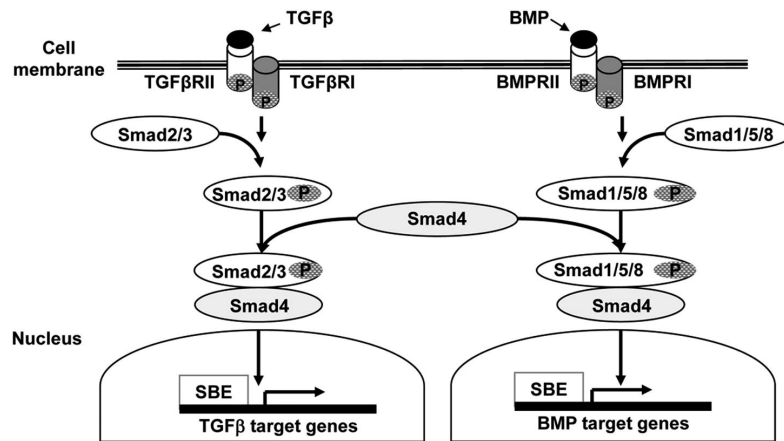
## 1.4 Master Regulators of Mesoderm Formation

### Smads

Smads are the main intracellular mediators of TGF $\beta$  family signaling and are required for mesoderm formation<sup>21–23</sup>. The SMAD signaling cascade is initiated by binding of a ligand to the transmembrane receptor, which phosphorylates receptor-mediated Smads (R-Smads), namely Smad 1, 2, 3, 5, 8. Then the phosphorylated R-Smads combine with Smad4 to form a complex which is transported to the nucleus to regulate the expression of the downstream target genes<sup>24–26</sup> (Figure 1.6). Smad2/3 mediate TGF $\beta$  and Nodal signaling, while Smad1/5/8 mediate BMP signaling.

Smad2 and Smad3, which can be activated by the receptors of TGF $\beta$  ligands, are expressed from the blastocyst stage in early mouse embryo. Their common and distinct functions have been widely studied. The phenotypes of *Smad2* mutant mice are very different from *Smad3* mutant mice. *Smad2* mutant embryos exhibit early patterning abnormalities<sup>27</sup>, while *Smad3* mutant embryos are viable<sup>28</sup>. Double homozygous mutation of Smad2 and Smad3 results in failure of mesoderm formation<sup>21</sup>. The high expression level of *Nodal/Smad2/3* is required for the specification of definitive endoderm, node and notochord<sup>21</sup>. In a study of human embryonic stem (ES) cells differentiation, Smad2/3 and  $\beta$ -catenin were shown to bind to the same regions in PS genes, and their direct interaction mediated PS gene activation<sup>29</sup>. A global identification of Smad2 targets via ChIP-seq in zebrafish showed that the Smad2-mediated transcriptional network is conserved in vertebrate mesoderm and endoderm<sup>30</sup>. Mullen *et al.* suggested that the differential Smad2/3 binding sites are determined by cell-type-specific master TFs<sup>31</sup>.

*Smad1* and *Smad5* were reported to exhibit genetic interaction and to function cooperatively to control the expression of BMP target genes in the early mouse embryo<sup>32</sup>. Genome-wide binding sites of Smad1/5 were detected by ChIP-seq in human endothelia cells, followed by the discovery of a GC-rich motif<sup>33</sup>. In a separate study, Smad1 was shown to share the same motif as Pou5f1/Sox2, which reflects the frequent co-binding of Smad1 with Pou5f1/Sox2<sup>34</sup>. It was also shown that in the context of human ES cell differentiation, interaction of Smad1 with T promotes mesoderm formation, while repressing endodermal differentiation<sup>35</sup>.



**Figure 1.6 Smads function as intracellular signaling mediators**

Smad2/3 mediate TGFβ signaling pathway. Smad1/5/8 mediate BMP signaling pathway. Smad4, which is involved in both pathways, binds to phosphorylated Smad2/3 in TGFβ pathway and to phosphorylated Smad1/5/8 in BMP pathway. Figure taken from Malkoski & Wang (2012)<sup>24</sup>.

## Eomes

Eomes is a member of the T-box TF family and plays important roles during gastrulation and trophoblast development in vertebrate embryos<sup>36,37</sup>. *Eomes*-null mouse embryos arrest at the stage of blastocyst. The functions of *Eomes* in cell lineage differentiation are strongly conserved in vertebrate systems<sup>38</sup>. In the mouse, Eomes has been shown to be essential for EMT, mesoderm formation and DE specification<sup>19</sup>. In the early zebrafish embryo, Eomes was shown to regulate all three germ layers<sup>30</sup>.

During mouse embryogenesis, *Eomes* expression in the embryo proper starts in the posterior part of the epiblast at embryonic day 5.75 (E5.75)<sup>36</sup>. During gastrulation, *Eomes* is expressed in the distal PS and *Eomes* expressing cells generate two distinct progenitor cells: the cranial and cardiac mesodermal progenitors and the progenitor cells of anterior primitive streak (APS) derivatives (definitive endoderm, node and notochord)<sup>37</sup>. *Eomes*-deficient mouse embryos fail to downregulate E-cadherin, blocking EMT and hence mesoderm formation<sup>19</sup>. *Eomes* cooperates with the *Nodal/Smad2/3* pathway, promoting delamination of nascent mesoderm<sup>19</sup>. *Eomes* is a marker gene of the earliest cardiac mesoderm. The master regulator of multipotent cardiovascular progenitor specification *Mesp1* can be directly activated by Eomes<sup>37,39,40</sup>. In addition to *Mesp1*, the other cardiac-specific markers including *Myf7*, *Myf2*,

*Nkx2.5*, *Myocardin* and *Mef2c* are also not expressed in *Eomes* null embryos<sup>37</sup>. Notably, the expression of *T* is not altered while *Eomes* is absent<sup>37</sup>.

## T

As the founder member of the T-box TF family, which play a central role in mesoderm formation<sup>41</sup>, the gene *Brachyury* (*T*) was identified by Dobrovolskaia-Zavadskaia in 1927 with the discovery that *T* mutation caused truncated tails in mice<sup>42,43</sup>. It was shown that mouse embryos with a homozygous *T* mutation die at about E10, lacking the allantois and failing to form a proper notochord, neural tube and somites<sup>44,45</sup>. The cloning of the *T* gene in 1990 was a big breakthrough after decades of classical experimental analysis, which has promoted the functional analysis of *T* on a molecular level<sup>46</sup>. *T* is expressed in nascent mesoderm, its progenitors and in the notochord<sup>47</sup>. *T* expression in neuro-mesodermal progenitors and the *T* mutant phenotype showed that *T* is essential for mesoderm formation in the trunk<sup>48</sup>. Studies performed on *Xenopus* and chick embryos showed that the expression patterns of *T* are conserved in vertebrates<sup>49,50</sup>. Studies by Smith *et al.* show that in *Xenopus* the mis-expression of *T* homologue can induce mesoderm formation at ectopic locations<sup>49,51</sup>.

The role of T as a TF was shown by Kispert and Herrmann<sup>52</sup>. They demonstrated that T protein binds to a consensus palindromic sequence and a monomer of T is sufficient for binding. Casey *et al.* showed that in *Xenopus* T can regulate expression of *eFGF* by binding to a half-site of the palindromic sequence<sup>53</sup>. A systematic study of T function in the *Ciona* notochord, where T is notochord-specific, shows that T regulates most of its target genes via non-palindromic binding sites<sup>54</sup>.

The gene regulatory functions of *T* in embryonic development have been extensively studied. *T* is a direct target of Wnt3a, a signal expressed in the PS and required for paraxial mesoderm formation in mouse. It was suggested that Wnt3a modulates the determination of mesodermal and neural cell fates via *T*<sup>55</sup>. The study by Schulte-Merker and Smith suggests that *T* and FGF form a regulatory loop, in which T activates a member of FGF while FGF maintains *T*'s expression<sup>56</sup>. The following studies showed that loss of *Fgf4* and *Fgf8* in presomitic mesoderm progenitors results in severe down-regulation of *T*<sup>57</sup>. In addition, *Fgfr1* functions in mesoderm cell fate specification by positively regulating *T* and *Tbx6* expression<sup>58</sup>, suggesting that *T* is an important downstream gene of the FGF signaling pathway. The modern experimental techniques, like ChIP-chip (Chromatin immunoprecipitation



combined with microarray), ChIP-seq (Chromatin immunoprecipitation followed by sequencing) and RNA-seq (RNA sequencing), facilitated the prediction of genome-wide direct targets of *T*. Morley *et al.* performed ChIP-chip and identified targets of *T* ortholog No tail (*Ntl*) in zebrafish. They showed that *Ntl* directly regulates the notochord-expressed gene *flh* and other TFs. A gene regulatory network describing mesoderm formation in zebrafish was assembled by using this ChIP-chip data<sup>59</sup>. In *Xenopus*, *T* was shown by ChIP-seq to have around 5500 binding sites. Comparison of *T* and Eomes binding revealed that the T-box proteins are recruited to the same genomic sites and their collaboration ensures the correct determination of mesoderm and neural tissues<sup>60</sup>. *T* targets were identified in differentiating human ES cells and it was discovered that the expression of *T* target genes depended on the cellular environment and differential interaction of *T* with Smad1 or Smad2/3 signaling<sup>35,61</sup>. Lolas *et al.* showed that *T* forms a feedback loop with *Foxa2* and *Sox17* to direct cell lineage commitment during streak formation<sup>62</sup>. A recent study of our group demonstrated that in differentiated mouse ES cells *T* directly targets many key regulators, including *Wnt3a*, *Fgf8*, *Tbx6*, *Msgn1* and *Sox2*<sup>48</sup>.

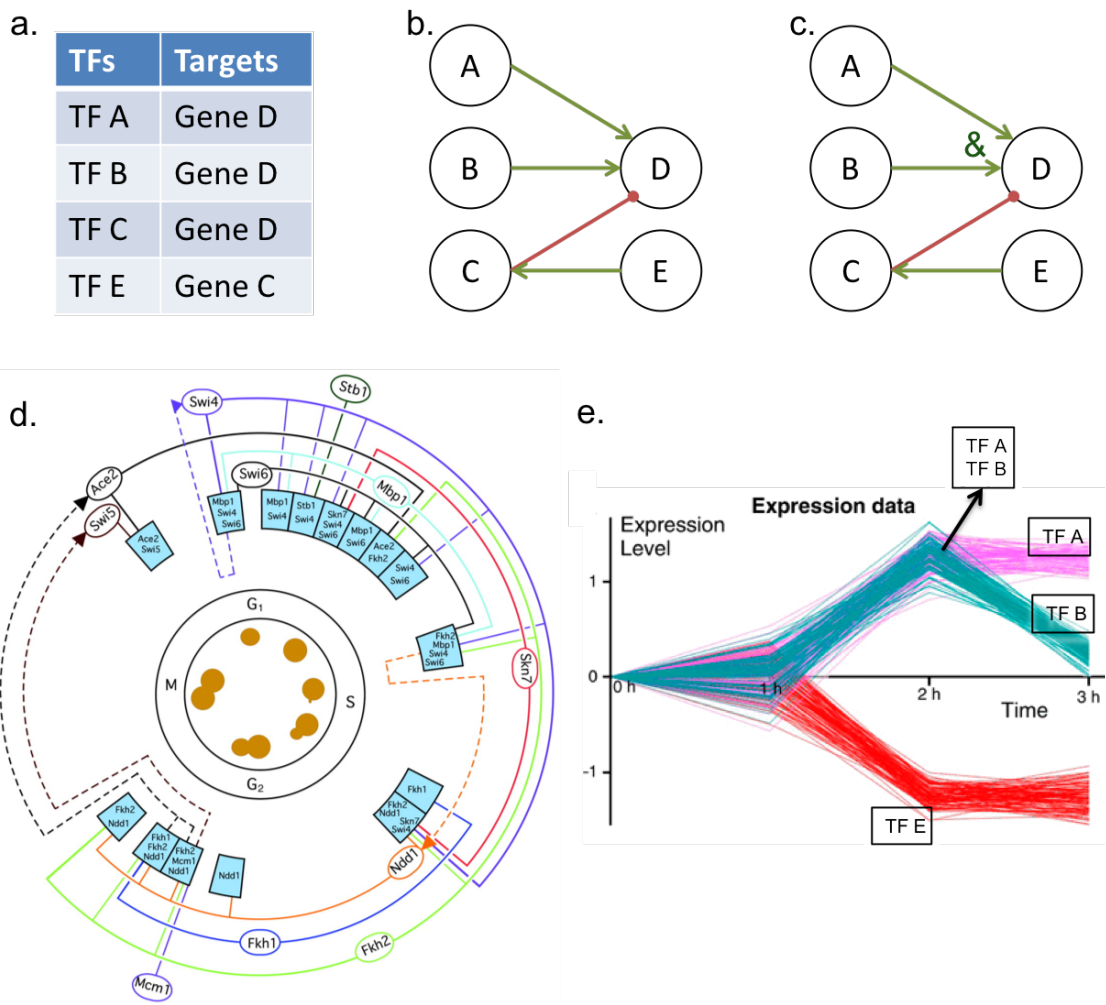
### 1.5 Studies of *in Vitro* Mesoderm Formation

The studies of *in vitro* ES cell differentiation help to understand the mechanisms of embryogenesis and to exploit the therapeutic potential of ES cells. Advantages of *in vitro* differentiation studies include the ability to produce enough material for large scale experimental studies and to precisely score developmental stages.

ES cells are pluripotent and can be used to generate all embryonic tissues. The initiation of primitive streak (PS) is essential for mesoderm formation. Modelling of the PS development showed that WNT and TGF $\beta$  signaling are required *in vitro* for the induction of PS from mouse ES cells<sup>63</sup>. It was reported that Nodal/Activin A can initiate TGF $\beta$  signaling and induce PS-like cells. In addition, for mesoderm induction *in vitro*, Nodal/Activin A direct the nascent mesoderm toward axial mesoderm and mesendoderm.<sup>64</sup> It was shown that short-term BMP4 treatment initiates mesoderm formation in human ES cells<sup>65</sup>. In this study we generated mesodermal cells by growing mouse ES cells on fibronectin in medium containing Bmp4.

## 1.6 Gene Regulatory Networks

Gene regulatory networks (GRNs), which can involve signaling pathways, gene regulators and target genes, are built to understand and visualize how a cell responds to internal or external stimuli<sup>66</sup>. Schilitt and Brazma suggested to categorize GRNs by increasing details and complexities into the following four classes<sup>67</sup>(Figure 1.7, a-d). The first class, parts lists, comprises a collection, organization and description of the associated elements involved in the networks, e.g., TFs, promoters and TF binding sites. Such lists can be represented as a table or database. The second class, topology models, connects the different elements to show their interactions and relationships as wiring diagrams. The third class, control logic models, on top of the topology models explains the rules/logic of interactions between all elements, e.g., how the regulators collaborate when they regulate the same gene. The fourth class, dynamic models, captures the dynamic changes of gene regulatory events over time. These four categories describe a gene regulatory event from different perspectives. For a fixed number of elements in a network, each category can explain the regulatory program in more detail than the previous one<sup>67</sup>. In 2007, Ernst *et al.* presented a new type of GRN model called “global temporal map”<sup>66</sup>. This model is dynamic over time. As opposed to the dynamic model of the previous approach (class four), it globally describes the transcriptional regulatory events causing the observed time-series gene expression patterns and the TFs controlling these events<sup>66</sup> (Figure 1.7, e).



**Figure 1.7 Different types of gene regulatory network (GRN) models**

(a). Example of a small parts list, which collects, organizes and describes the elements of a network. (b). A graph based on the known elements of a network, where nodes represent genes and edges denote the relationships between genes. A, B and C can all bind to D. The relationships are as follows: A and B activate D, C inhibits D, E activates C. (c). Example of network logics. Gene D is activated if A and B are both bound, but not C. (d). Example of a dynamic model. A transcriptional regulatory network over the time of the yeast cell cycle (stages G1, S, G2, M). (e). Example of the “global temporal map”, describing globally the transcriptional regulatory events causing the observed time-series gene expression patterns and the TFs controlling these events. The split at the time point 2h is induced by TF A and B. Figures d and e are taken from Lee *et al.* (2002)<sup>68</sup> and Ernst *et al.* (2007)<sup>66</sup>.

The desired complexity of a GRN is based on the purpose of the study. If the goal is to explain how one specific TF regulates the downstream genes at one specific time point, a static GRN model is generally sufficient. Technologies such as ChIP-seq and RNA-seq followed by the appropriate bioinformatic analysis can define the downstream genes of one specific TF. When many TFs are involved in the studied process and the datasets are limited, the static GRNs can be constructed using computational predictions. To evaluate the relevance of two variables in a network, the pairwise association methods, e.g., Pearson correlation<sup>69</sup>, mutual information<sup>70</sup> and partial correlation<sup>71</sup>, can be used. Given a threshold of the association score, the logic behind the pairwise association methods is that variables with high association scores are relevant as edges in the network. In addition, Bayesian networks have also been suggested as an effective method to construct GRNs<sup>72,73</sup>.

To model dynamic time-series GRNs as shown in Figure 1.7 d, previous studies used methods including Boolean network model<sup>74</sup>, Petri nets<sup>75,76</sup> and difference equation model<sup>77</sup>. For the modelling of GRNs such as the one depicted in Figure 1.7 e, which integrate time-series gene expression data and TF-gene interactions, the computational method developed by Ernst *et al.* can be used<sup>66</sup>.

## 2 Experimental Material and Methods

All wet lab experiments were performed by Dr. Pavel Tsaytler.

### 2.1 Mesodermal Differentiation *In Vitro*

For mesoderm differentiation, mouse embryonic stem cells (mESCs) were firstly separated from feeders. Then the single cell suspension of feeder-free mESCs was plated on square plates in 5  $\mu$ l drops using multichannel pipette. The square plates were kept upside down in the incubator for 12 hours to allow single cells to form aggregates. The aggregates were then transferred to fibronectin-coated plates and treated with Bmp4. The Bmp4 containing medium was refreshed daily. For each experiment, the cells were collected at the appropriate time points.

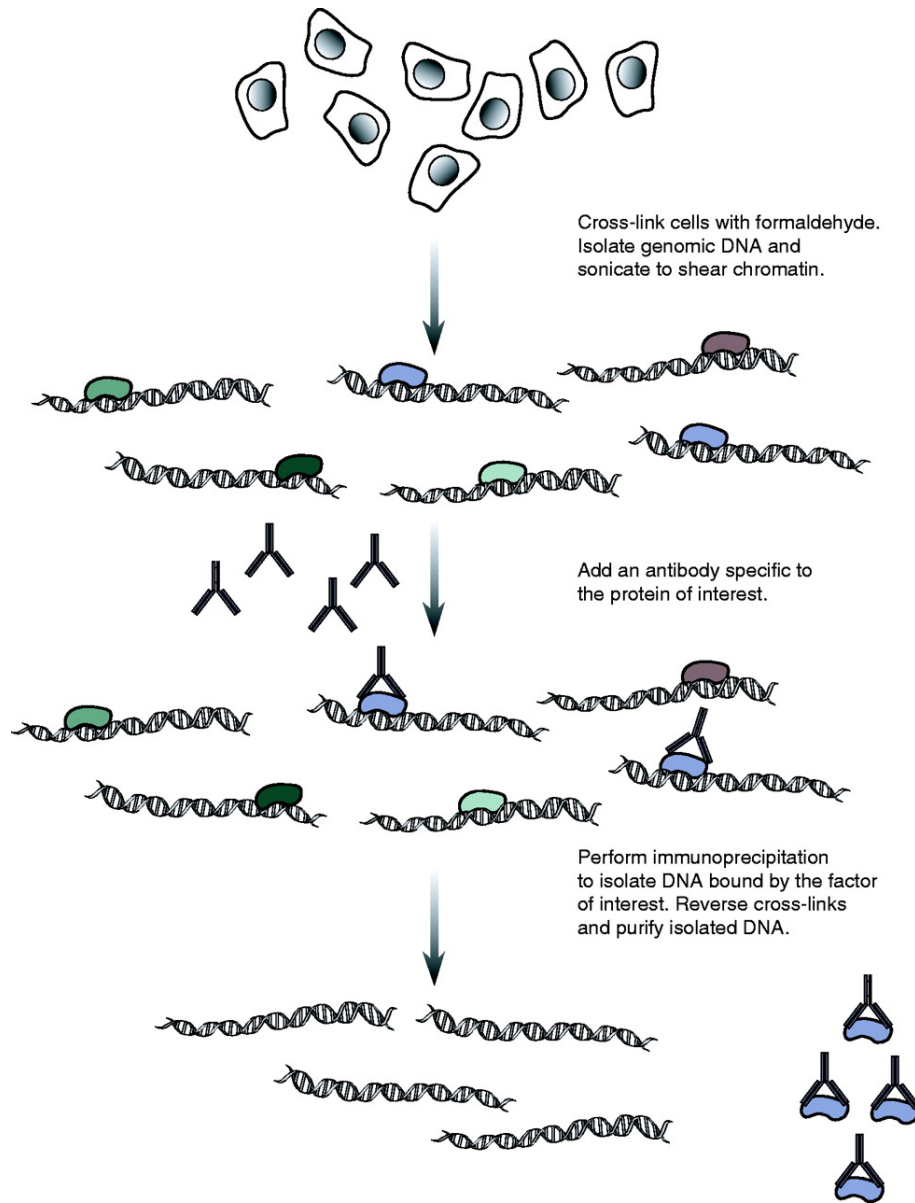
### 2.2 Chromatin Immunoprecipitation (ChIP)

Chromatin immunoprecipitation is an experimental method to study the interactions of DNA and proteins in the cell. It aims to determine the DNA binding sites of a specific protein (e.g., whether a TF binds to the gene of interest). In addition, ChIP can also be used to identify the locations of histone modifications. Figure 2.1 shows the steps of a ChIP experiment. Firstly, genomic DNA and DNA-binding proteins are cross-linked usually by using formaldehyde. Then the cells are lysed, and the DNA is sheared to small fragments with desired length by sonication. With the antibody specific to the protein of interest, the DNA fragments bound by this protein are enriched by immunoprecipitation. The DNA fragments which are not bound to this specific protein are washed away. The final step is to purify isolated DNA fragments by reversing cross-links and removing the proteins. Polymerase chain reaction (PCR) is usually used afterwards to determine the enrichment of specific DNA fragments.

For the ChIP experiment, a control sample which is called “Input” is usually generated. The preparation of “Input” follows the same procedure without the immunoprecipitation step, thus representing the whole genome.

The purified DNA fragments can be identified on a larger scale by combining ChIP with other techniques. ChIP-chip combines ChIP with microarray hybridization, where the target

genes are detected by enrichment on the microarray<sup>78</sup>. ChIP-seq combines ChIP with high-throughput sequencing, by which all of the target DNA fragments get sequenced and determined on a genome-wide scale.

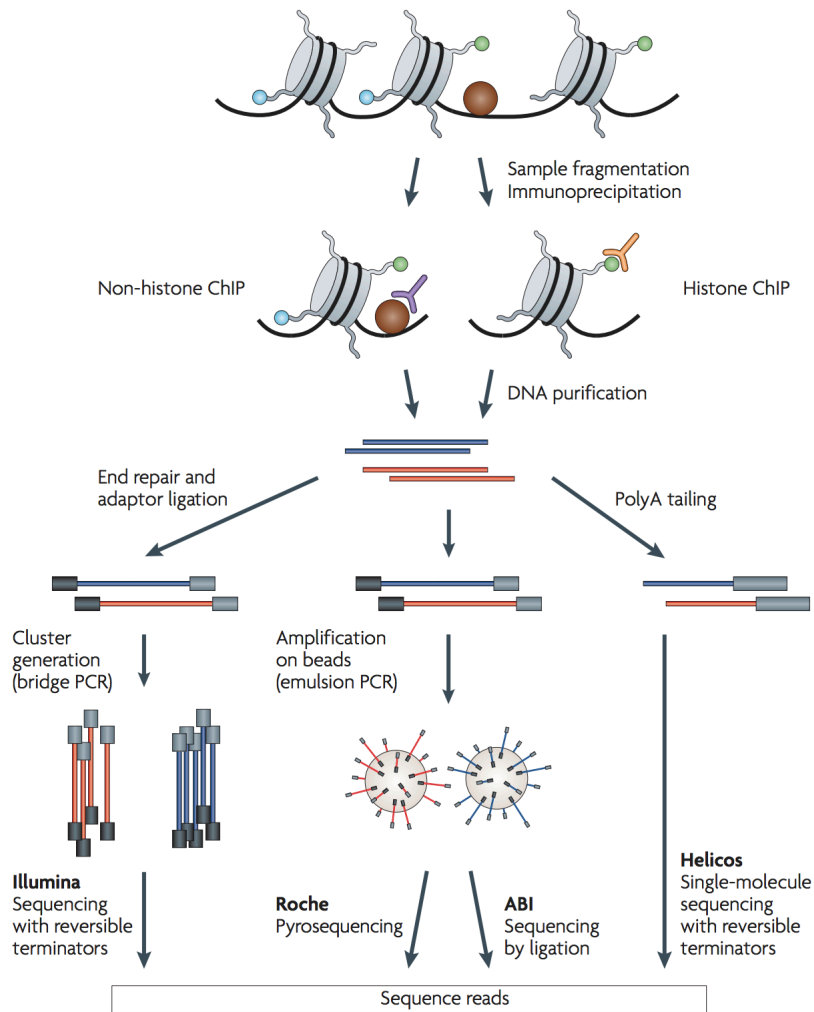


**Figure 2.1 Overview of the chromatin immunoprecipitation (ChIP) experiment**  
(1) Formaldehyde cross-linking. (2) Chromatin sonication. (3) Adding protein-specific antibody. (4) Immunoprecipitation of chromatin. (5) Purifying DNA fragments. Figure taken from Hoffman *et al.* (2009)<sup>183</sup>.

### 2.3 Chromatin Immunoprecipitation Followed by Sequencing (ChIP-seq)

The ChIP-seq method, first reported in 2007, combines ChIP with high-throughput sequencing<sup>79</sup>. High-throughput sequencing is also known as next-generation sequencing (NGS) to make a distinction from previously existing methods, such as Sanger sequencing. The significant improvement of NGS technologies is that they allow for massively parallel reactions, with millions sequencing reactions that can be performed at the same time<sup>80</sup>. NGS technologies include Illumina sequencing, Roche 454 sequencing, SOLiD sequencing and Ion Torrent: Proton/PGM sequencing. For different technologies, the experimental steps are common, including library preparation, sequencing reactions and data output. The difference lies in the details of sequencing approaches. Illumina sequencing, which is the dominant method currently, takes the approach of “sequencing-by-synthesis”. There are billions of fragments attached to a solid surface. For each of them, the polymerase enzyme can add only one single base to the growing complementary strand because each base is modified with a terminator which will block further polymerization. The terminator has a fluorescent label (four colors or two colors chemistry) which can be detected. All templates are detected simultaneously for each cycle of sequencing reactions and then the terminators are removed to continue the polymerization. The reversible terminators are the key for this method since they allow the reactions to synchronize. In the end, all of the images combined show the A, T, G, C order for each sequence. NGS is quicker than the old methods. For example, Sanger sequencing separates the processes of chemical reaction and signal detection while NGS usually combines them. In addition, Sanger sequencing takes only one read each time while NGS is massively parallel.

ChIP-seq has many advantages, including high genome coverage, high throughput and decreasing cost of sequencing, so it has become an important tool to study gene regulation. Figure 2.2 illustrates the sequencing strategies for a ChIP-seq experiment<sup>81</sup>. Following the purification of DNA fragments, they are assayed for NGS to identify the sequences associated with the TF. Further analysis of these fragments will reveal the genomic locations and potential target genes of this TF across the entire genome.



**Figure 2.2 Overview of the ChIP-seq process**

DNA and DNA-binding proteins are cross-linked and enriched by ChIP. Purified DNA can be sequenced with NGS technique by different platforms. ChIP-seq can be used for either histones or TFs. Figure taken from Park (2009)<sup>81</sup>.

## 2.4 RNA Sequencing (RNA-seq)

RNA-seq aims to obtain the transcriptome information using NGS. The transcriptome is the complete set of RNA molecules and their relative quantities in a sample. Knowing the transcriptome is important to understand the functional elements in cells at a given condition. It is an effective way to analyze the molecular mechanisms of development and disease<sup>82</sup>.



The most common method to perform RNA-seq is Illumina NGS and the steps are as mentioned in section 2.2: (A) library preparation, (B) sequencing, and (C) data output. The library for RNA-seq can be the cDNA fragments converted from total RNA or selected population, like poly(A)+, depending on the research purpose.

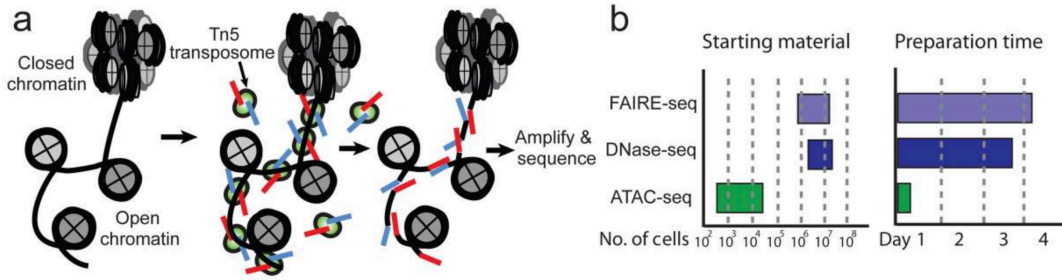
Compared to DNA microarray<sup>83</sup>, which is a hybridization-based approach for studying transcriptome, RNA-seq has significant advantages. Firstly, the sequences detected by microarrays are based on the probes for hybridization, so microarrays cannot detect novel genes, which is not a limitation in RNA-seq. Secondly, RNA-seq allows to determine expression levels of all detected transcripts, while microarrays are not sensitive to genes very lowly or highly expressed<sup>82</sup>.

## **2.5 Assay for Transposase-accessible Chromatin using Sequencing (ATAC-seq)**

The ATAC-seq procedure was published in 2013 by Buenrostro *et al.*<sup>13</sup>. It combines the usage of Tn5 transposase and NGS, aiming to assess chromatin accessibility.

Eukaryotic DNA is hierarchically compacted into chromatin to fit in the nucleus. However, for transcriptional machinery to function, certain chromatin regions should be accessible<sup>84</sup>. Since information about the chromatin structure can shed light on the mechanisms of gene regulation, chromatin accessibility data is a valuable asset to study GRNs. The published methods to determine chromatin accessibility include DNase-seq<sup>85</sup>, FAIRE-seq<sup>86</sup> and ATAC-seq. ATAC-seq has two main benefits as compared to other methods. Firstly, it needs less cells (from 1 to 50000) than DNase-seq (50 million) and FAIRE-seq (1-50 million). Secondly, it requires fewer steps than the other two methods to perform, which is less time-consuming (Figure 2.3 b)<sup>13</sup>.

ATAC-seq takes advantage of a mutated hyperactive Tn5 transposase which effectively cuts exposed DNA regions and simultaneously integrates adapters into those regions<sup>87</sup>. The adapter-ligated fragments are then amplified by PCR for NGS (Figure 2.3 a). The NGS profile provides us insights into the structure of chromatin, such as chromatin accessibility and nucleosome positioning.



### Figure 2.3 ATAC-seq process and advantages

(a) Tn5 transposase (green) with adapters (red and blue) interacts with open chromatin regions, cuts exposed DNA fragments and integrates adapters to these fragments. (b) ATAC-seq needs less cells and time to perform, compared to the methods FAIRE-seq and DNase-seq. Figure taken from Buenrostro *et al.* (2013)<sup>13</sup>.

## 2.6 Datasets for Analysis

As shown in Table 2-1, the data generated by Dr. Pavel Tsaytler in this study include: time-series transcriptome RNA-seq for 10 time points (index: 1 to 10), Smads ChIP-seq and RNA-seq (wild-type (WT) vs. knockout (KO)) at day 2 (index: 11 to 14), Eomes ChIP-seq and RNA-seq at day 2 (index: 15, 16), T ChIP-seq and RNA-seq at day 3 (index: 17, 18), T/Eomes double knockout RNA-seq at day 3 (index: 19), and time-series ATAC-seq for 6 time points (index: 20 to 25).

**Table 2-1. Datasets from experiments**

| Index | Library Type     | ES Cell Type              | Treatment                                    |
|-------|------------------|---------------------------|--|
| 1     | RNA-seq          | wild-type                 | ES (2 replicates)                            |
| 2     | RNA-seq          | wild-type                 | Mesoderm differentiation 1 h (2 replicates)  |
| 3     | RNA-seq          | wild-type                 | Mesoderm differentiation 6 h (2 replicates)  |
| 4     | RNA-seq          | wild-type                 | Mesoderm differentiation 12 h (2 replicates) |
| 5     | RNA-seq          | wild-type                 | Mesoderm differentiation 1 d (2 replicates)  |
| 6     | RNA-seq          | wild-type                 | Mesoderm differentiation 2 d (2 replicates)  |
| 7     | RNA-seq          | wild-type                 | Mesoderm differentiation 3 d (2 replicates)  |
| 8     | RNA-seq          | wild-type                 | Mesoderm differentiation 4 d (2 replicates)  |
| 9     | RNA-seq          | wild-type                 | Mesoderm differentiation 5 d (2 replicates)  |
| 10    | RNA-seq          | wild-type                 | Mesoderm differentiation 6 d (2 replicates)  |
| 11    | ChIP-seq/P-Smad1 | wild-type                 | Mesoderm differentiation 2 d                 |
| 12    | ChIP-seq/Smad2/3 | wild-type                 | Mesoderm differentiation 2 d                 |
| 13    | RNA-seq          | wild-type (Smad4 control) | Mesoderm differentiation 2 d                 |
| 14    | RNA-seq          | Smad4 knockout            | Mesoderm differentiation 2 d                 |
| 15    | ChIP-seq/Eomes   | wild-type                 | Mesoderm differentiation 2 d                 |
| 16    | RNA-seq          | Eomes knockout            | Mesoderm differentiation 2 d                 |
| 17    | ChIP-seq/T       | wild-type                 | Mesoderm differentiation 3 d                 |
| 18    | RNA-seq          | T knockout                | Mesoderm differentiation 3 d                 |
| 19    | RNA-seq          | T/Eomes knockout          | Mesoderm differentiation 3 d                 |
| 20    | ATAC-seq         | wild-type                 | ES (2 replicates)                            |
| 21    | ATAC-seq         | wild-type                 | Mesoderm differentiation 1 d (2 replicates)  |
| 22    | ATAC-seq         | wild-type                 | Mesoderm differentiation 2 d (2 replicates)  |
| 23    | ATAC-seq         | wild-type                 | Mesoderm differentiation 3 d (2 replicates)  |
| 24    | ATAC-seq         | wild-type                 | Mesoderm differentiation 4 d (2 replicates)  |
| 25    | ATAC-seq         | wild-type                 | Mesoderm differentiation 5 d (2 replicates)  |

## 3 Computational Methods

### 3.1 Correlations: Pearson vs. Spearman

Correlations are commonly used to measure the association between two variables. For example, correlations were used in this study for estimating the reproducibility of replicates and for clustering. The correlation coefficient, ranging from -1 to 1, indicates the strength and the direction of the relationship. The most common method, Pearson correlation<sup>69</sup>, measures linear relationships, while the distribution-free method Spearman's rank correlation<sup>88</sup> measures monotonic relationships.

#### Pearson correlation coefficient

The Pearson correlation coefficient is calculated as the quotient of the covariance of two variables divided by the product of their standard deviations:

$$\rho_{X,Y} = \frac{\text{cov}(X,Y)}{\sigma_X \sigma_Y} \quad (1)$$

This formula is used for a population to get the population correlation coefficient represented by  $\rho$ . When applied to a sample to calculate the sample correlation coefficient represented by  $\gamma$ , the covariances and variances are estimated based on a sample, in which case the formula for  $\gamma$  is

$$\gamma = \frac{\sum_i^n (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (2)$$

where  $n$  is the sample size of  $\{x_1, \dots, x_n\}$  and  $\{y_1, \dots, y_n\}$ ,  $i \in n$ ,  $\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$ .

#### Spearman's rank correlation coefficient

The Spearman's rank correlation coefficient is the Pearson correlation coefficient of the ranked variables. At the first step all the variables are ranked, and then the Pearson correlation coefficient is calculated with the ranked variables.

## Comparison of the Pearson and Spearman correlation methods

The correlation coefficient of both methods ranges from -1 to 1. A positive value means the correlation between two variables is positive, otherwise it is negative (anti-correlated). The absolute value of the correlation coefficient shows how strong the tendency is.

The difference between the two is that the Spearman correlation cannot be used to estimate whether two variables change at a constant rate. If one variable increases as the other increases, the Spearman correlation coefficient equals 1, but the Pearson correlation can be less than 1 as long as the increase rate of two variables is different. In practice, the Pearson correlation is used to answer whether the correlation is linear while the Spearman correlation is used only for the estimation of the monotonicity.

### 3.2 Fisher's Exact Test for Enrichment Analysis

High-throughput sequencing data analysis usually outputs large gene lists, which are hard to be functionally interpreted. Enrichment analysis is a traditional way to mine big data, with the idea to compare the genes of interest with the background to see whether a specific feature is enriched. For example, if 10% of the background genes and 50% of the selected genes share the same feature, this feature is very likely associated with the selected genes. Enrichment analysis allows to treat large gene lists with gene group-based view, rather than individual gene-based view<sup>89</sup>. It makes the process of biological interpretation for large gene lists more effective and convenient. The well-known statistical methods for enrichment analysis include Chi-square test, hypergeometric test and Fisher's exact test. Hypergeometric test is identical to one-tailed Fisher's exact test. Chi-square is not appropriate when the values are very small (any value in a contingency table less than 5), making Fisher's exact test the preferred choice in this case<sup>90</sup>.

Fisher's exact test is mainly used to examine the significance of the association between two categories within a  $2 \times 2$  contingency table (Table 3-1)<sup>91</sup>. In this table, the marginal totals (row and column totals) are assumed to be fixed. The cells are filled with observed values of different features. The question here is "how likely it is to obtain the observed contingency table, given the null hypothesis that there is no association between two categories". The significance (p-value) can be exactly calculated by summing up the probabilities of observing "a" and more extreme cases given the null hypothesis were true. To calculate the probability

of observing “ $a$ ”, since it is actually a repeated drawing event without replacement, the hypergeometric distribution is used to get the exact probability:

$$p = \frac{\binom{a+b}{a} \binom{c+d}{c}}{\binom{n}{a+c}} = \frac{(a+b)! (c+d)! (a+c)! (b+d)!}{a! b! c! d! n!} \quad (3)$$

**Table 3-1.  $2 \times 2$  contingency table**

The value “ $a$ ” is the count of the intersection of category 1 and 2. “ $b$ ” or “ $c$ ” refers to the count of the elements belonging to only one of the two categories. “ $d$ ” is for the rest which are not in both. The totals of rows and columns are called marginal totals.

|                  | Category 2 (Yes) | Category 2 (No) | Row Total           |
|------------------|------------------|-----------------|---------------------|
| Category 1 (Yes) | $a$              | $b$             | $a + b$             |
| Category 1 (No)  | $c$              | $d$             | $c + d$             |
| Column Total     | $a + c$          | $b + d$         | $a+b + c + d (= n)$ |

In practice, one-tailed Fisher’s exact test (equivalent of hypergeometric test) is usually performed to predict whether there is a positive or negative association between two categories. In this case, the same or more extreme situations compared to the observed data needed to be considered from only one direction to calculate the significance of the observed data.

In sections 4.2 and 4.3, the one-tailed Fisher’s exact test with “alternative hypothesis: true odds ratio is greater than 1” was used.

### 3.3 Correcting for Multiple Testing

Given a threshold of p-value 0.05 for a hypothesis test, there is 5% chance to obtain a false significant result which fits the null hypothesis. If the same test is run for thousands of hypotheses, the chance of getting false positive results will be highly increased. For NGS data analysis, the data size usually is large and multiple tests are inevitable. Adding up all false positives results in a high number, which produces the so-called multiple testing problem or multiple comparisons problem<sup>92</sup>.

To control the number of false positives, an approach developed by Benjamini and Hochberg in 1995 is mainly used. It is a method aiming to control the expected proportion of falsely rejected hypotheses, namely the false discovery rate (FDR)<sup>93</sup>. In a multiple comparisons test, if  $V$  and  $S$  indicate the values of false and true positives separately, the ratio of  $V/(V + S)$  is defined as the proportion of false discoveries represented by  $Q$ . The FDR is the expectation of  $Q$ :

$$FDR = E(Q) = E \{V/(V + S)\} = E(V/R) \quad (4)$$

The Benjamini-Hochberg procedure can be divided in two steps. Firstly, the p-values of all of the  $m$  hypotheses tested are ranked from the smallest  $p_1$  to the largest  $p_m$ . Secondly, to control FDR at level  $q$ , the largest  $i \in m$  is determined such that

$$p_i \leq \frac{i}{m} q \quad (5)$$

The discoveries with p-values ranking from  $p_1$  to  $p_i$  can be finally selected, which will statistically assure that FDR is controlled not higher than  $q$ .

### 3.4 Maximum Likelihood Estimation

Maximum likelihood estimation (MLE) is a method used in the situation when the statistical model is given while the parameters are unknown. MLE tries to obtain the parameters  $\theta$  that maximize the likelihood function  $\mathcal{L}(\theta; \mathbf{x}) = p(\mathbf{x}; \theta)$ , given the observations  $\mathbf{x}$ . The logic behind this method is to find the parameter values that make the observations most probable<sup>94</sup>.

This method defines  $\hat{\theta}$  which maximizes the likelihood function  $\mathcal{L}(\theta; \mathbf{x})$  as the maximum likelihood estimate:

$$\hat{\theta} = \underset{\theta}{\operatorname{argmax}} \mathcal{L}(\theta; \mathbf{x}) \quad (6)$$

In practice, log-likelihood  $\log\mathcal{L}(\theta; \mathbf{x})$  is usually used instead of  $\mathcal{L}(\theta; \mathbf{x})$ , which produces the same maximum likelihood estimate.

The general process of obtaining  $\hat{\theta}$  includes: (1) generate the likelihood function, (2) take the derivative of the likelihood function and set it equal to 0:  $\frac{d\log\mathcal{L}(\theta; \mathbf{x})}{d\theta} = 0$ , given continuous parameter space.

### 3.5 Expectation Maximization Algorithm

The expectation maximization (EM) algorithm is used to extend the usage of maximum likelihood estimation (MLE) to the cases where there are hidden variables<sup>95</sup>. Given the observed data set  $\mathbf{x}$  and the hidden data set  $\mathbf{z}$ , EM algorithm aims to obtain the parameters  $\theta$  that maximize the likelihood function  $\log\mathcal{L}(\theta; \mathbf{x}) = \sum_{\mathbf{x}} \log P(\mathbf{x}; \theta) = \sum_{\mathbf{x}} \log \sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}; \theta)$ .

The EM algorithm is based on Jensen's inequality for concave function (e.g.,  $\log(*)$ ), which states  $\log(E[\mathbf{y}]) \geq E[\log(\mathbf{y})]$  (Equality holds if and only if  $\mathbf{y}$  is constant), where  $\mathbf{y}$  is a random variable and  $E[\mathbf{y}]$  is the expectation of it<sup>95</sup>. Thus, for any probability distribution  $Q(\mathbf{z})$ , using the random variable  $\mathbf{y} = \frac{P(\mathbf{x}, \mathbf{z}; \theta)}{Q(\mathbf{z})}$ :

$$\log\left(\sum_{\mathbf{z}} P(\mathbf{x}, \mathbf{z}; \theta)\right) = \log\left(\sum_{\mathbf{z}} Q(\mathbf{z}) \frac{P(\mathbf{x}, \mathbf{z}; \theta)}{Q(\mathbf{z})}\right) \geq \sum_{\mathbf{z}} Q(\mathbf{z}) \log\left(\frac{P(\mathbf{x}, \mathbf{z}; \theta)}{Q(\mathbf{z})}\right) \quad (7)$$

where  $Q(\mathbf{z}) = P(\mathbf{z}|\mathbf{x}; \theta)$  makes the equality hold.

The EM algorithm is as follows: (1) initiate the parameters  $\theta^t$ , (2) the expectation step (E-step): construct a function  $g_t$ ,  $g_t(\theta) = \sum_{\mathbf{z}} P(\mathbf{z}|\mathbf{x}; \hat{\theta}^t) \log\left(\frac{P(\mathbf{x}, \mathbf{z}; \theta)}{P(\mathbf{z}|\mathbf{x}; \hat{\theta}^t)}\right)$ , which lower-bounds  $\log P(\mathbf{x}; \theta)$  [Eq. (7)], (3) the maximization step (M-step): determine new parameters  $\theta^{t+1}$  which is calculated as the maximum of  $g_t$ , (4) repeat E-step and M-step until the algorithm converges at the final optimized parameters<sup>95</sup>. In general, the idea behind EM algorithm is to assume the missing data, then the problem with incomplete data turns to using MLE to estimate parameters based on imputed complete data.

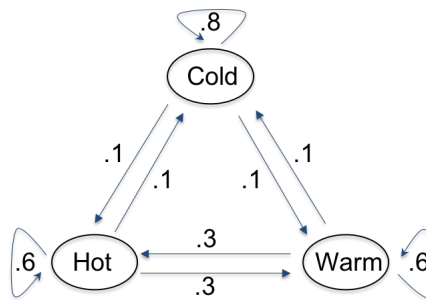


### 3.6 Hidden Markov Model (HMM) and Corresponding Algorithms

Hidden Markov model (HMM) is a statistical model, which with respect to time series data analysis describes the relationships of the observed events and the causal hidden states. Referred to the book written by Jurafsky and Martin<sup>96</sup>, this part introduces the algorithms for HMM.

#### Parameters of Hidden Markov Model

The hidden Markov model was developed in 1996 by Baum *et al.*<sup>97</sup> based on Markov chains<sup>98</sup>. A Markov chain describes a sequence of states and the corresponding transitions between different states, as shown in Figure 3.1. In a Markov chain, the initial probability distribution, which specifies the probabilities of the starting states, and the transition probabilities are required.



**Figure 3.1 A Markov chain for three different weather conditions**

The transition probabilities are shown. For example, the transition probabilities of “Hot” to “Hot”, “Warm” and “Cold” are 0.6, 0.3 and 0.1, respectively.

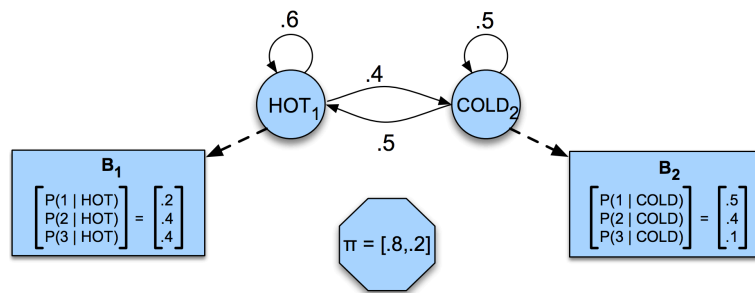
In a hidden Markov model, the states of interest are hidden. Rather, we observe a sequence of events caused by the states. When we have a set of  $N$  states  $Q = \{q_1, q_2, \dots, q_N\}$  and a sequence of  $T$  observations  $O = (o_1, o_2, \dots, o_T)$ , a hidden Markov model is defined by three parameters which can be represented as  $\lambda = (\pi, A, B)$ .  $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$  defines the initial probability distribution over all states, whereof  $\pi_i$  is the probability of the Markov chain starting in state  $i$ .  $A = [a_{ij}]_{N \times N}$  is the transition probability matrix, with  $a_{ij}$  representing the probability of changing from state  $i$  to  $j$ .  $B = b_i(o_t)$  is the sequence of observation likelihoods (emission probability matrix), representing the probability of

observing  $o_t$  from state  $i$ . Table 3-2 shows the components of HMM. Figure 3.2 shows a hidden Markov model<sup>96,99</sup>. In this example, there are two hidden states of weather conditions “HOT” and “COLD”. The observations are the numbers of ice creams to eat in either weather condition.  $\pi$ ,  $A$  and  $B$  are shown in the figure.

**Table 3-2. Components of a Hidden Markov model**

Table edited according to Daniel *et al.*<sup>96</sup>

| Components   | Property   |
|--|--|
| State space: $Q = \{q_1, q_2, \dots, q_N\}$                              | Set of $N$ states  |
| Observation sequence: $O = (o_1, o_2, \dots, o_T)$                       | Sequence of $T$ observations. $o_t$ is from all possible observations                        |
| Transition probabilities: $A = [a_{ij}]_{N \times N}$                    | $a_{ij}$ is the probability of changing from state $i$ to $j$ .<br>$\sum_{j=1}^N a_{ij} = 1$ |
| Emission probabilities: $B = b_i(o_t)$                                   | $b_i(o_t)$ is the probability of observing $o_t$ from state $i$                              |
| Initial probability distribution: $\pi = \{\pi_1, \pi_2, \dots, \pi_N\}$ | $\pi_i$ is the probability of starting in state $i$ . $\sum_{i=1}^N \pi_i = 1$               |



**Figure 3.2 An example of hidden Markov model**

A hidden Markov model representing the association of weather with the numbers of ice creams eaten in either weather. Figure taken from Jurafsky and Martin (2018)<sup>96,99</sup>.

### Two assumptions of HMM

HMM makes two important assumptions about the data while modeling (HMM in this thesis refers to first-order HMM). First, the probability for one specific state depends only on its previous state, which is called Markov assumption:

$$P(q_i | q_1, \dots, q_{i-1}) = P(q_i | q_{i-1}) \quad (8)$$

Second, each observation  $o_i$  depends only on the direct hidden state  $q_i$ , which is called independence assumption:

$$P(o_i|q_1, \dots, q_i, \dots, q_T, o_1, \dots, o_i, \dots, o_T) = P(o_i|q_i) \quad (9)$$

### Three fundamental questions regarding HMM

HMM is featured with three important questions to answer. Firstly, HMM can be used for evaluation. Given  $\lambda = (\pi, A, B)$  and an observation sequence of  $O = (o_1, o_2, \dots, o_T)$ , the likelihood of  $P(O|\lambda)$  can be determined. Secondly, given  $\lambda = (\pi, A, B)$  and an observation sequence of  $O = (o_1, o_2, \dots, o_T)$ , we can use HMM to find the most likely hidden sequence underlying the observations. Lastly, given an observation sequence of  $O = (o_1, o_2, \dots, o_T)$  and a set of states, the parameters  $\lambda = (\pi, A, B)$  can be predicted via machine learning. There are corresponding algorithms to address those questions in the following sections.

### Evaluation: The Forward Algorithm

Given a sequence of observations  $O = (o_1, o_2, \dots, o_T)$  and parameters  $\lambda = (\pi, A, B)$  of an HMM, to calculate the likelihood of observing  $O$  from a particular sequence of states  $Q$ , we can calculate the joint probability:

$$P(O, Q) = P(O|Q) \times P(Q) = \prod_{i=1}^T P(o_i|q_i) \times \prod_{i=1}^T P(q_i|q_{i-1}) \quad (10)$$

Then, to get the likelihood of observing  $O$  from all possible hidden states, we just need to sum up all joint probability results from all possible hidden state sequences:

$$P(O) = \sum_Q P(O, Q) = \sum_Q P(O|Q)P(Q) \quad (11)$$

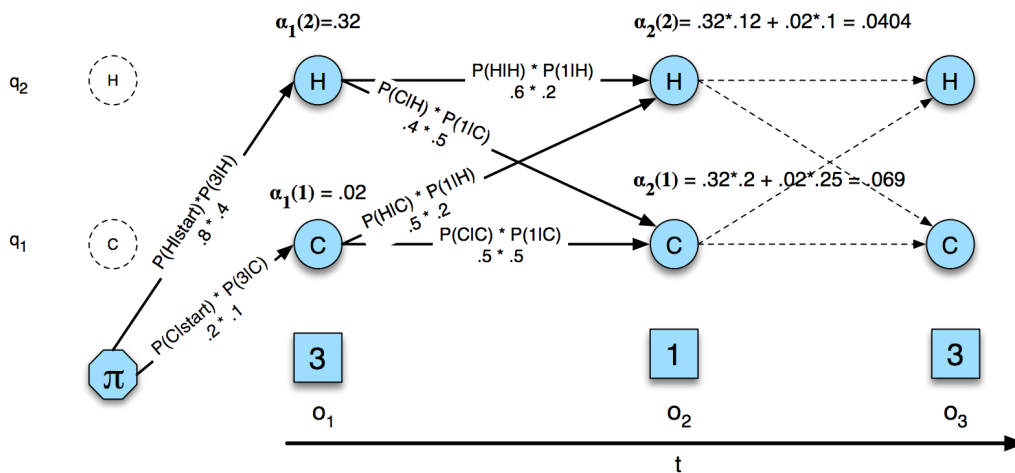
There is an obvious problem with the above method. It works well when the sample is small. If the state space  $N$  is big, the  $N^T$  possible hidden state sequences make it hard to compute. Instead, we use a dynamic programming algorithm, which is called the forward algorithm.

The forward algorithm calculates the final result in a recursive manner by utilizing a forward trellis. As shown in Figure 3.3, which is an example of forward algorithm for calculating the

likelihood of observing  $O = (3, 1, 3)$  given the hidden states H (hot) and C (cold), the forward algorithm recursively calculates the probability of each cell in the forward trellis  $\alpha_t(j)$ , representing the probability of being in state  $j$  after passing through the first  $t$  observations, by

$$\alpha_t(j) = \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} b_j(o_t) \quad (12)$$

Here,  $\alpha_{t-1}(i)$  is the previous forward path probability from the beginning to step  $t-1$ .  $a_{ij}$  is the transition probability of moving from previous state  $i$  at step  $t-1$  to current state  $j$  at step  $t$ .  $b_j(o_t)$  is the probability of observing  $o_t$  in the current state  $j$ . In the example,  $\alpha_2(1)$  is computed as the sum of  $\alpha_1(1) \times P(C|C) \times P(1|C)$  and  $\alpha_1(2) \times P(C|H) \times P(1|C)$ .



**Figure 3.3 An illustration of forward algorithm**

The forward algorithm using forward trellis to compute the likelihood of observing ice-cream events  $O = (3, 1, 3)$ , given the hidden states of weather H (Hot) and C (Cold). States are in circles, while observations are in squares.  $\alpha_t(j)$  at each cell represents the sum of probabilities of all paths reaching this cell. Figure taken from Jurafsky and Martin (2018)<sup>96</sup>.

Putting together the initialization and termination step, the forward algorithm to obtain the likelihood of observations in an HMM is summarized below.

(1) Initialization:

For each hidden state  $j$ :

$$\alpha_1(j) = \pi_j b_j(o_1) \quad (13)$$

(2) Recursion:

For  $t = 2$  to  $T$ :

For each hidden state  $j$ :

$$\alpha_t(j) = \sum_{i=1}^N \alpha_{t-1}(i) a_{ij} b_j(o_t) \quad (14)$$

(3) Termination:

$$P(O|\lambda) = \sum_{i=1}^N \alpha_T(i) \quad (15)$$

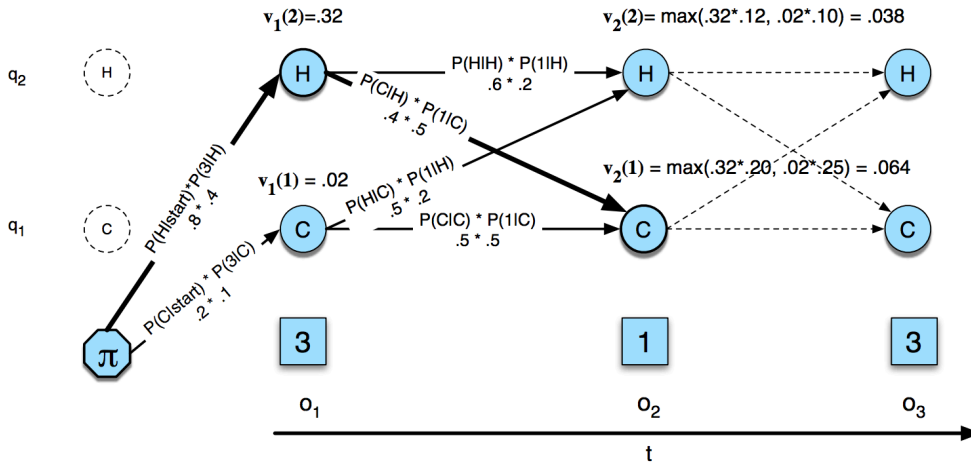
### Decoding the hidden states: Viterbi Algorithm

Given an observation sequence  $O = (o_1, o_2, \dots, o_T)$  and an HMM  $\lambda = (\pi, A, B)$ , using Viterbi algorithm<sup>100</sup>, we can find the most possible hidden state sequence.

The logic of Viterbi algorithm is almost the same as the forward algorithm. The difference is that at the recursion step, we take the max of the previous path probabilities instead of summation. As shown in Figure 3.4, which is the same example as for forward algorithm, each cell of the forward trellis,  $v_t(j)$ , is calculated by taking the most probable path heading to this cell,

$$v_t(j) = \max_{1 \leq i \leq N} v_{t-1}(i) a_{ij} b_j(o_t) \quad (16)$$

Here,  $v_{t-1}(i)$  is the pervious Viterbi path probability from the start to step  $t - 1$ .  $a_{ij}$  is the transition probability from previous state  $i$  to current state  $j$ .  $b_j(o_t)$  is the probability of seeing  $o_t$  in the current state  $j$ . In the example,  $v_2(1)$  is computed by taking the max of  $v_1(1) \times P(C|C) \times P(1|C)$  and  $v_1(2) \times P(C|H) \times P(1|C)$ .



**Figure 3.4 An example of calculating Viterbi path probability**

Given the observation sequence (in squares) of the numbers of ice-creams taken for three days  $O = (3, 1, 3)$  and two hidden states of weather H (Hot) and C (Cold) (in circles), the best path heading to the cell  $v_2(1)$  is the path with bold arrows and the value of  $v_2(1)$  is calculated as the max of  $v_1(1) \times P(C|C) \times P(1|C)$  and  $v_1(2) \times P(C|H) \times P(1|C)$ . Figure taken from Jurafsky and Martin (2018)<sup>96</sup>.

In the end, we want to get the most probable hidden state sequence underlying the observation sequence. The idea is to create a trace-back array at each step of getting  $v_t(j)$ . In Equation 16, we store the state  $i$  at  $t - 1$  which leads to the largest value for  $v_{t-1}(i) a_{ij} b_j(o_t)$ , namely  $\operatorname{argmax}_{1 \leq i \leq N} v_{t-1}(i) a_{ij} b_j(o_t)$ , in an array. After reaching the final step of getting the best score of  $\max_{1 \leq i \leq N} v_T(i)$ , we can trace back step by step and obtain the corresponding hidden state sequence which leads to this best score. Putting all steps together, the Viterbi algorithm is performed as below.

(1) Initialization:

For each hidden state  $j$ :

$$v_1(j) = \pi_j b_j(o_1) \quad (17)$$

$$bt_1(j) = 0 \quad (18)$$

(2) Recursion:

For  $t = 2$  to  $T$ :

For each hidden state  $j$ :

$$v_t(j) = \max_{1 \leq i \leq N} v_{t-1}(i) a_{ij} b_j(o_t) \quad (19)$$

$$bt_t(j) = \operatorname{argmax}_{1 \leq i \leq N} v_{t-1}(i) a_{ij} b_j(o_t) \quad (20)$$

(3) Termination:

$$\text{The best score: } P^* = \max_{1 \leq i \leq N} v_T(i) \quad (21)$$

$$\text{The start of backtrace: } q_T^* = \operatorname{argmax}_{1 \leq i \leq N} v_T(i) \quad (22)$$

### Learning the parameters: Baum-Welch Algorithm

Given an observation sequence  $O = (o_1, o_2, \dots, o_T)$ , the HMM parameters  $\lambda = (\pi, A, B)$  can be learned by using Baum-Welch algorithm, which is an instance of EM algorithm (section 3.5)<sup>101</sup>.

In Baum-Welch algorithm, the values for  $\lambda = (\pi, A, B)$  are randomly initialized firstly, then they are repeatedly updated until convergence. Each updating iteration has two steps: E-step and M- step ([www.cs.cmu.edu/~tbergkir/11711fa17/recitation4\\_notes.pdf](http://www.cs.cmu.edu/~tbergkir/11711fa17/recitation4_notes.pdf)).

**E-step:** Assume  $\lambda = (\pi, A, B)$  are known and compute  $\gamma_t(i)$  (the probability of being in state  $i$  at time  $t$ ) and  $\xi_t(i, j)$  (the probability of being in state  $i$  at time  $t$  and  $j$  at time  $t + 1$ ):

$$\gamma_t(i) = P(q_t = i | O, \lambda) \quad (23)$$

$$\xi_t(i, j) = P(q_t = i, q_{t+1} = j | O, \lambda) \quad (24)$$

To compute  $\gamma_t(i)$  and  $\xi_t(i, j)$ , we firstly define the backward probability  $\beta_t(i) = P(o_{t+1}, o_{t+2}, \dots, o_T | q_t = i, \lambda)$  which is the probability of seeing  $o_{t+1}, o_{t+2}, \dots, o_T$  given starting state  $i$  at time  $t$ .  $\beta_t(i)$  is calculated as:

$$\beta_t(i) = \sum_{j=1}^N \alpha_{ij} b_j(o_{t+1}) \beta_{t+1}(j), 1 \leq i \leq N, 1 \leq t < T \quad (25)$$

Now, knowing the forward probability  $\alpha_t(i)$  and the backward probability  $\beta_t(i)$ , we calculate

$$\gamma_t(i) = P(q_t = i | O, \lambda) = \frac{P(q_t = i, O | \lambda)}{P(O | \lambda)} = \frac{\alpha_t(i) \beta_t(i)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)} \quad (26)$$

$$\begin{aligned} \xi_t(i, j) &= P(q_t = i, q_{t+1} = j | O, \lambda) \\ &= \frac{P(q_t = i, q_{t+1} = j, O | \lambda)}{P(O | \lambda)} \end{aligned} \quad (27)$$



$$= \frac{\alpha_t(i) a_{ij} b_j(o_{t+1}) \beta_{t+1}(j)}{\sum_{j=1}^N \alpha_t(j) \beta_t(j)}$$

The denominators of  $\gamma_t(i)$  and  $\xi_t(i, j)$  are the same, which is the probability of seeing the observation  $O$  given the parameters  $\lambda$ .

**M-step:** When  $\gamma$  and  $\xi$  are known, we use MLE to estimate the updated  $\lambda$ :

$$\hat{\pi}_i = \gamma_1(i) \quad (28)$$

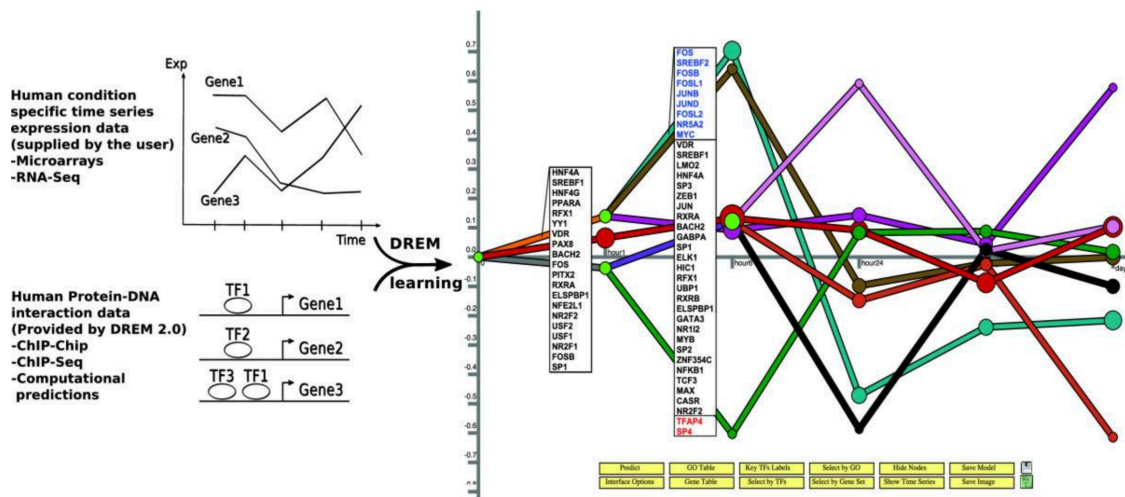
$$\hat{a}_{ij} = \frac{\text{\#times } j \text{ follows } i}{\text{\#times anything follows } i} = \frac{\sum_{t=1}^{T-1} \xi_t(i, j)}{\sum_{t=1}^{T-1} \gamma_t(i)} \quad (29)$$

$$\hat{b}_i(v_k) = \frac{\text{\#times } o \text{ is observed given } i}{\text{\#times anything is observed given } i} = \frac{\sum_{t=1}^T \delta_{(o_t, v_k)} \gamma_t(i)}{\sum_{t=1}^T \gamma_t(i)}, \quad (30)$$

$$\text{where } \delta_{(o_t, v_k)} = \begin{cases} 1, & \text{if } o_t = v_k \\ 0, & \text{if } o_t \neq v_k \end{cases}$$

### 3.7 “Dynamic Regulatory Events Miner (DREM)” Based on Input-Output Hidden Markov Model (IOHMM)

DREM is a software developed to integrate the time-series gene expression data and motif information (e.g., TF-targets information from CHIP-seq) to infer an annotated global temporal gene regulatory map<sup>66</sup>. In this map, the gene expression patterns are explained by transcriptional regulatory events and the corresponding transcription factors (Figure 3.5).



**Figure 3.5 An illustration of a DREM (version 2.0) analysis**

Left: Input data for DREM, including the time-series gene expression data and the motif information (TF-gene interactions). Right: The model acquired for the given 6 time points. TFs (in boxes) are predicted to annotate the map, showing when and at which paths the specific TFs are functioning. The colors blue and red indicate up- and down-regulated TFs as compared to time point 0, respectively. Figure taken from Schulz *et al.* (2012)<sup>102</sup>.

This method is based on input-output hidden Markov model (IOHMM)<sup>103</sup>, an extension of HMM. Analogous to HMM, IOHMM uses hidden states to cluster genes. Each hidden state is associated with a Gaussian output distribution of the gene expression values for one time point. IOHMM extends HMM by allowing for an additional input dataset, which is the motif information offered by users, to control the transition probabilities of genes from one hidden state to another<sup>66</sup>. In the first version of DREM, this input is the same for all time points. With the second version, the dynamic input can be used, meaning the input can be different for different time points. For calculating DREM models, this additional input of motif information is an optional feature. In this study, the second version of DREM was used and the use of the motif information was disabled in the model calculation.

Details of the DREM algorithm are shown below<sup>66</sup>:

### Parameters

Given the motif information is used for the training of the model  $M$ , the parameters include:

$n$  is the number of time points of the time-series expression data.

$H$  is a set of hidden states. Each hidden state  $h$ ,  $h \in H$ , is associated with one time point and a Gaussian output distribution  $f_h$ .

$\Theta$  is for the Gaussian output distributions. For each hidden state  $h$ , there is an element  $(\mu_h, \sigma_h) \in \Theta$  corresponding to the Gaussian distribution  $f_h$ , where  $\mu_h$  and  $\sigma_h$  refer to the mean and standard deviation respectively.

$E$  is the set of directed edges connecting hidden states  $H$  and are constrained to enforce a tree structure. Each hidden state is constrained to have a maximum of  $\gamma$  children. The root of a tree is formed by the state at the first time point. For any hidden state except for the ones at the last time point, there must be at least one following state transitioning from it. For any hidden state except for the state at the first time point, there must be exactly one state at the previous time point.

$\Psi$  controls transition probabilities between hidden states, given the motif information is used to train the model. If a state  $h$  ( $h \in H$ ) has two children  $a$  and  $b$ , meaning  $(h, a) \in E$ ,  $(h, b) \in E$ ,  $a \neq b$ , there is  $\psi_h$  ( $\psi_h \in \Psi$ ) which defines a logistic function<sup>104</sup> to calculate the transition probability of a gene from one state to another. For a gene  $g$  targeted by TFs  $x$ , the transition probability from state  $h$  to  $a$  is:

$$\frac{1}{1 + e^{-\psi_h(INT) - \sum_x \psi_h(x) \times I_g(x)}} \quad (31)$$

where  $\psi_h(INT)$  is the intercept parameter of the logistic function and  $\psi_h(x)$  is the corresponding element of  $\psi$  for each TF.  $I_g(x_i)$  ( $x_i \in x, I_g(x_i) = \{0, 1\}$ ) equals 1 if gene  $g$  is regulated by TF  $x_i$  and 0 otherwise.

### Likelihood Function

$o_g = (o_g(1), \dots, o_g(n-1))$  indicates the log ratio expression values of gene  $g$  at time points 1 to  $n-1$ , with the value at time point 0 as control.  $P(H_t = h_b | H_{t-1} = h_a, I_g)$ , where  $H_t$  refers to the hidden state variable at time  $t$  and  $I_g$  is the static input vector, is calculated as the transition probability of a gene  $g$  transitioning from state  $h_a$  at time  $t-1$  to  $h_b$  at time  $t$ . This probability is 1 when  $h_b$  is the only child of  $h_a$  and 0 when  $h_b$  is not a child of  $h_a$ . If  $h_a$  has more children, the transitions are depending on  $I_g$ .  $I_g$  is mapped to transition probabilities by a logistic function as described in previous part. The likelihood density  $\gamma$  for a gene set  $G$  with the model  $M$  is:

$$r(G|M) = \sum_{g \in G} \log \sum_{q \in Q} \prod_{t=1}^{n-1} f_{q(t)}(o_g(t)) \prod_{t=1}^{n-1} P(H_t = q(t) | H_{t-1} = q(t-1), I_g) \quad (32)$$

In the above equation,  $Q$  is the set of all paths of the constructed tree and each path connecting hidden states is with the length of  $n$  starting from the root. In a path  $q \in Q$ ,  $q(i)$  is the hidden state of this path at time point  $i$ . The product  $\prod_{t=1}^{n-1} f_{q(t)}(o_g(t))$  is the product of the output densities of the expression values with given hidden states. The other product  $\prod_{t=1}^{n-1} P(H_t = q(t) | H_{t-1} = q(t-1), I_g)$  is the product of transition probabilities of given hidden states. Here this equation considers  $I_g$ , but it will not be related to  $I_g$  when the motif information is not used for model training.

## Model Learning

The model learning process has the following steps:

Pseudocode\*:

1. Separate the gene set into a train set  $G_{train}$  and a test set  $G_{test}$ .
2. Initiate the tree structure  $(H, E)$  to be a single chain. Then perform parameter training and calculate the test score. The training is aimed to find the settings for  $\Psi$  and  $\Theta$  which maximize  $r(G_{train}|M)$ . The test score is  $r(G_{test}|M)$ .
3. If the test score improves do
  - a.  $(H', E') \leftarrow (H, E)$
  - b. For each hidden state,  $h$ , which can have another child
    - i. Temporarily add a single chain of hidden states from  $h$  to  $(H', E')$
    - ii. Train the temporary model from step 3.b.i
    - iii. let  $(H, E)$  be the model structure from step 3.b.i, if the score of  $r(G_{test}|M)$  is best found so far
  - c.  $(H', E') \leftarrow (H, E)$
  - d. For each hidden state,  $h$  in  $H'$ , which has a sibling in  $H'$ 
    - i. Temporarily remove  $h$  and all descendants from  $(H', E')$
    - ii. Train the temporary model from step 3.d.i
    - iii. let  $(H, E)$  be the model structure from step 3.d.i, if the score of  $r(G_{test}|M)$  is at least as good as the best so far
4. Randomly resplit train and test data.
5. Delete weakly supported paths, delay appropriate splits.
6. Train parameters of model using all genes.
7. Assign genes to paths using the Viterbi algorithm.
8. Remove any path with fewer than 5 genes.

\* edited based on the pseudocode published by Ernst *et al.*<sup>66</sup>

## Transcription Factor Scoring

After the tree structure is determined and the genes are assigned to the tree, TFs are assigned to the paths out of the splits to explain the bifurcation events in the time course. Given a TF  $f$ , a split  $S$  and a path  $A$  out of this split, the score of TF  $f$  for split  $S$  on path  $A$  is computed using the hypergeometric distribution:

$$\sum_{i=c_A}^{\min(c_S, n_A)} \frac{\binom{c_S}{i} \binom{n_S - c_S}{n_A - i}}{\binom{n_S}{n_A}} \quad (33)$$

where  $n_S$  is the total number of genes for split  $S$ , of which  $n_A$  genes are on path  $A$ , and  $c_S$  is the number of genes into the split regulated by TF  $f$ , of which  $c_A$  genes are on path  $A$ .

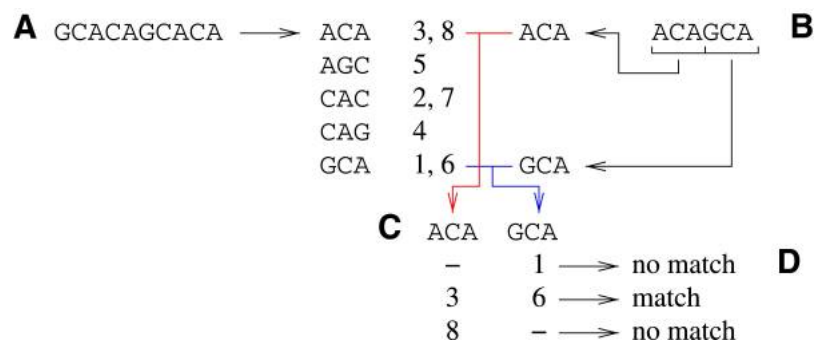
### 3.8 Read Mapping: Hash Table or Burrows-Wheeler Transform

Mapping reads to the reference genome as accurate as possible is a crucial step for the NGS data analysis. The process follows the sub-processes: building index of the reference genome, obtaining seeds from a query sequence, performing pairwise alignment (seeds vs. reference genome). Several mapping tools have been developed, including BWA<sup>105</sup>, Bowtie<sup>106</sup>, SOAP2<sup>107</sup> and SSAHA2<sup>108</sup>. The algorithms of index building tools can be grouped in two categories: hash table and Burrows-Wheeler Transform (BWT). One example of using hash table is shown in Figure 3.6. The disadvantage of this algorithm for read mapping is that it is time and internal memory consuming during computation.

BWT is shown in Figure 3.7. It was originally developed for data compression, so BWT-based indexing uses less memory than hashing algorithm. As shown in Figure 3.7 a, firstly, the reference sequence is transformed by adding a suffix \$ which is lexicographically less than A, C, G from the sequence. Then, the raw Burrows-Wheeler matrix “M” is constructed by taking turns to move one base from the last column to the first column each time for each row. The rows are further sorted lexicographically according to the columns to generate the transformed matrix “Mt”. The matrix “Mt” has three features: first, ordering the last (L) column lexicographically outputs the first (F) column; second, for each row, the base in column L is the one before the base in column F in the original sequence; third, for each base, the relative location in column L and F does not change, which means the first “a” in column L corresponds to the first “a” in column F. All of those features make it sufficient to keep only the last column of “Mt” as the index.

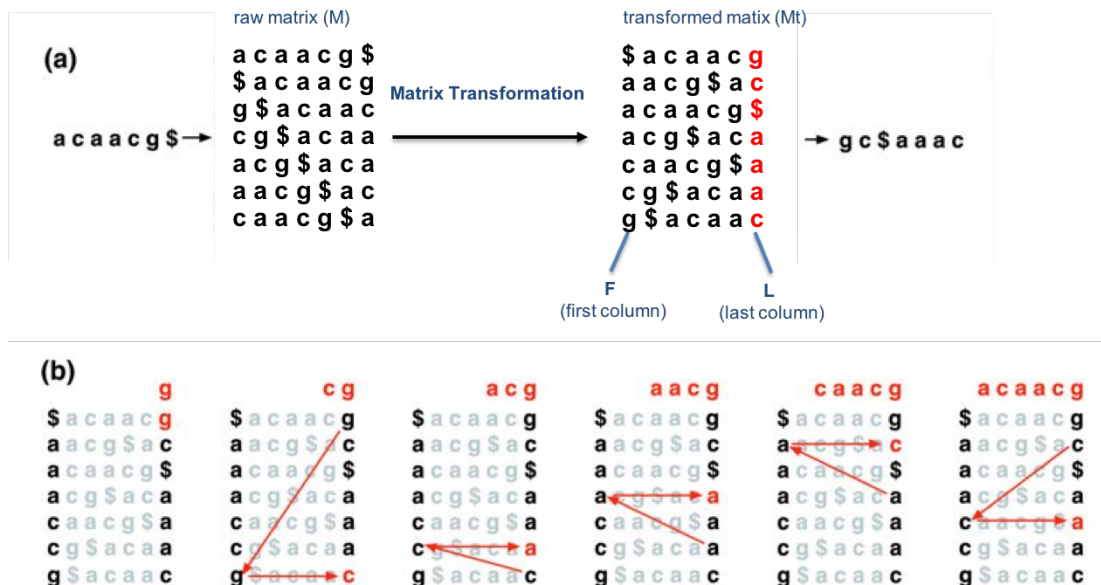
Knowing the last column of “Mt”, the first column can be inferred. Since the relative location of each base does not change from L to F, we can get the original sequence by doing “last first (LF) mapping” as shown in Figure 3.7 b. Similarly, when read mapping for a query sequence is performed by using the “LF mapping” algorithm, we will know whether the query sequence is matched (depending on how many mismatches are allowed) and its matching locations. Bowtie, which was used in this project, is based on BWT. It has the option *-v* to set the maximum number of mismatches for each query sequence. A maximum of two mismatches is allowed in this study. Another concern for read mapping is whether to consider the reads which are not uniquely mapped to the reference genome. Keeping uniquely mapped reads only may lead to the loss of some true binding sites, while keeping

multiple mapped reads may generate false-positives. In this study, the option *-m* in Bowtie (version 1.0.0) was used to keep only the reads uniquely mapped to the reference genome.



**Figure 3.6 Hashing algorithm**

(A) The genome is cut into overlapping 3-mers and their positions are stored in a hash table. (B) The query read is cut into 3-mers and compared to 3-mers from the reference genome. (C) Positions of each seed are sorted and compared to the other seeds. (D) The compatible positions are kept. Figure taken from Schbath *et al.* (2012)<sup>184</sup>.



**Figure 3.7 Burrows-Wheeler transform**

(a) The Burrows-Wheeler matrix and its transformation. (b) “last first (LF) mapping”. Figure modified from Langmead *et al.* (2009)<sup>106</sup>.

### 3.9 ChIP-seq Analysis

In this study, to characterize the gene regulatory networks mediated by Smads, Eomes and T, ChIP-seq experiments were performed for each of these TFs. By analyzing ChIP-seq data, the potential target genes across the entire genome for the corresponding TFs could be identified. Generally, the process for ChIP-seq data analysis can be divided into four steps: read mapping (section 3.8), peak calling, motif analysis and peak annotation including genomic distribution and GO enrichment analysis.

#### 3.9.1 Peak Calling

After the reads are mapped to the reference genome, the next step is to detect genome regions with significant enrichment of aligned reads. Those regions represent the DNA binding sites of a studied TF and will be associated with the genes that are potentially regulated by this factor. This step is performed using computational methods of peak calling.

The commonly used tool for ChIP-seq peak calling is MACS2, which is the latest version of MACS<sup>109</sup>. MACS2 has some improvements, but the underlying algorithm for peak calling is the same as in MACS. The workflow (Figure 3.8) and the related algorithms of MACS are described below<sup>110</sup>. The terms “tags” and “reads” are interchangeable in this session.

#### Removing Redundancy

To reduce the effects of the biases from PCR amplification and sequencing library preparation, MACS removes redundant reads and retains non-duplicated mapped reads.

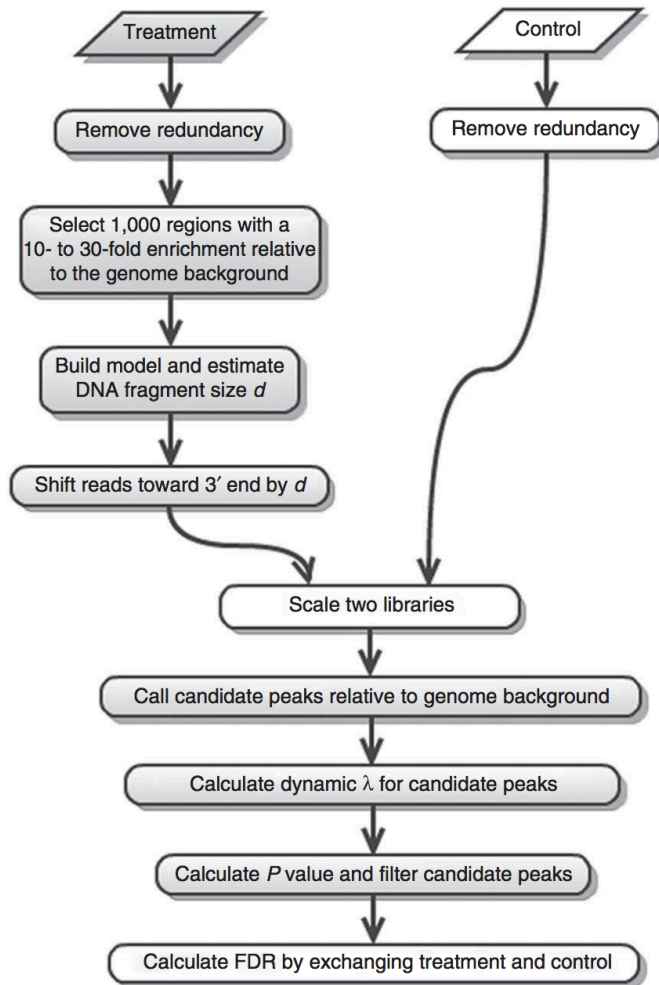
#### Modeling the Shift Size of ChIP-seq Tags

Theoretically, the ChIP DNA fragments have the same chances to get sequenced from both ends, so the Watson strand tags and the Crick strand tags around a true binding site should form a bimodal enrichment pattern (Figure 3.9 a). MACS uses this feature to model the shift size to locate the real binding sites.

MACS defines the sonication size as *bandwidth* and slides  $2*bandwidth$  windows across the reference genome to find regions which have tags more than *m fold* (a high-confidence fold-enrichment) enriched compared to a random tag distribution across the genome. 1000 of the resulting high-quality regions are randomly sampled by MACS. Their Watson and Crick tags

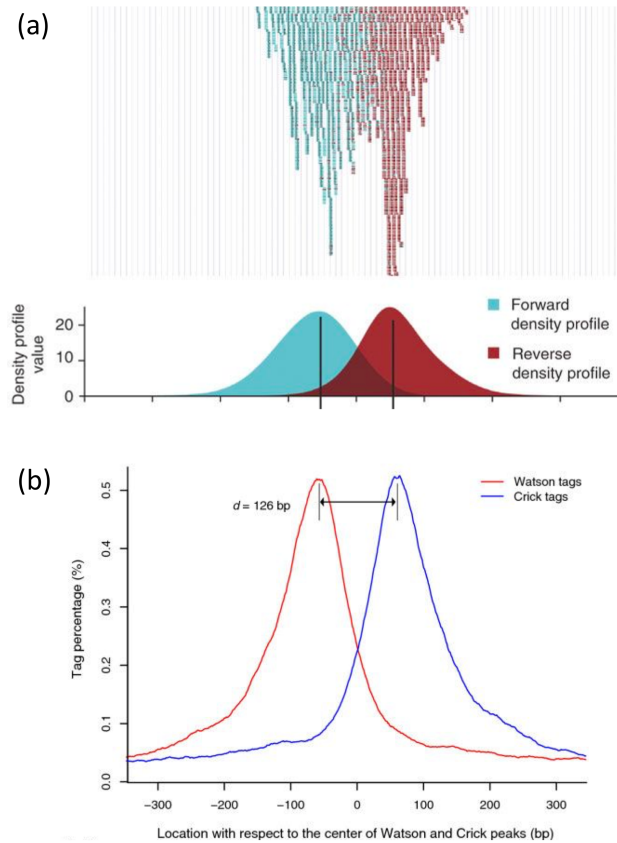


are then separated and aligned by the center of their Watson and Crick peaks (Figure 3.9 b). The distance between the modes of the Watson and Crick peaks is defined as “ $d$ ”. MACS then shifts all the reads by  $d/2$  in the 5’ to 3’ direction to locate the most probable true binding sites.



**Figure 3.8 Workflow of MACS**

The steps shown in white boxes are skipped when there is no control sample. Figure taken from Feng *et al.* (2012)<sup>110</sup>.



**Figure 3.9 Modeling the shift size of ChIP-seq tags**

(a). Forward/Watson and reverse/Crick read density profiles. (b). The 5' ends of strand-separated tags from 1000 high-quality peak are aligned by the center of their Watson and Crick peaks. Figure taken from Valouev *et al.* (2008)<sup>185</sup> and Zhang *et al.* (2008)<sup>109</sup>.

### Peak Detection: Poisson Distribution

ChIP-seq experiments usually generate millions of reads. With this high genome coverage, read distribution across the genome can be modeled by Poisson distribution (Eq 34).  $\lambda$  is the only parameter for Poisson distribution, which is here determined by total read number ( $n$ ), read length ( $l$ ) and effective genome size ( $s$ ):

$$P(k \text{ events}) = e^{-\lambda} \frac{\lambda^k}{k!} \quad (34)$$

$$\lambda = \frac{nl}{s}$$

( $n$ : Total read number;  $l$ : Read length;  $s$ : Effective genome length)

After  $d$  is defined, MACS shifts every tag by  $d/2$  distance and slides  $2d$  windows across the genome to find the regions which are significantly enriched in the ChIP sample (default  $p < 10e-5$ ). Instead of a global  $\lambda_{BG}$  determined by the whole genome, a dynamic parameter  $\lambda_{local}$ , defined for each candidate peak, is utilized by MACS:

$$\lambda_{local} = \max(\lambda_{BG}, [\lambda_{1k}], \lambda_{5k}, \lambda_{10k}) \quad (35)$$

$\lambda_{local}$  takes the maximum value of  $\lambda_{BG}$ ,  $[\lambda_{1k}]$ ,  $\lambda_{5k}$  and  $\lambda_{10k}$ , where  $\lambda_{1k}$ ,  $\lambda_{5k}$  and  $\lambda_{10k}$  are estimated from window sizes of 1 kb, 5 kb or 10 kb around the center of the peak location in the control sample. If there is no control sample, then the ChIP-seq sample itself is used for estimation (in this case  $\lambda_{1k}$  is not used). The advantage of using  $\lambda_{local}$  is that it captures the effect of local biases at least from DNA amplification and local chromatin structure. In addition, it is capable to deal with the situation of low read counts at small local regions<sup>109</sup>. By defining a threshold p-value, candidate peaks with a p-value lower than this threshold are called and the overlapping peaks are merged.

### Estimate False Discovery Rate (FDR)

When a control sample is available, MACS can estimate an empirical FDR for each peak by swapping the ChIP and control samples. At each p-value, the same parameters are used to detect ChIP peaks over control and control peaks over ChIP. The FDR is defined as “Number of control peaks / Number of ChIP peaks”<sup>109</sup>.

### 3.9.2 Association of Peaks to Genes

The obtained ChIP-seq peaks were then assigned to genomic regions (promoter, genic and intergenic) and potential target genes which are annotated in the RefSeq database. The peaks that overlap with -5kb/+2kb of the TSS (transcription start site) regions were categorized as

promoter-associated and assigned to the corresponding genes. The peaks that overlap with +2kb from the TSS to +5kb after the TES (transcription end site) regions were categorized as genic-associated and assigned to the corresponding genes. The remaining peaks were defined to be intergenic and were associated with the closest up- and down-stream genes.

### 3.10 RNA-seq Analysis

For each RNA-seq sample, all of the sequencing reads were mapped to the mouse genome (mm10) using TopHat (version 2.0.11), a program built on Bowtie to align RNA-seq reads to a genome, with settings `-M -g 1` to exclude the multi-mapped reads, thus keeping only the uniquely mapped reads<sup>111</sup>. Cuffdiff was then used to calculate the normalized FPKM (fragments per kilobase of transcript per million fragments mapped) which is a way to estimate gene expression levels. In this study, the default normalization method “geometric” was used for Cuffdiff<sup>112</sup>. For the detection of differential genes from RNA-seq data, the cutoff of log2 fold change 1 or 0.8 was used for distinct datasets.

The programming language perl (<https://www.perl.org/>), R statistical program (R 3.1.2; <https://www.r-project.org/>), R packages “venneuler” (<http://www.rforge.net/venneuler/>) and “gplots” (<http://CRAN.R-project.org/package=gplots>) were used for hierarchical clustering, generation of venn diagrams and heatmaps related to RNA-seq data analysis in this project.

### 3.11 ATAC-seq Analysis

#### ATAC-seq Read Mapping

Based on the procedure published by Koch *et al.*<sup>48</sup>, ATAC-seq data was firstly treated by using fastq-mcf of ea-utils (<http://expressionanalysis.github.io/ea-utils/>, version 1.04.807) to remove sequencing adapters (Supplementary Note 1) with the parameters `-0 -S -K -C 1000000`.

Because of a bug in earlier versions of bowtie, due to which a pair of identical forward and reverse reads fail to map, the forward and reverse reads in which the adapters were found and removed were mapped separately using bowtie (version 1.0.0)<sup>113</sup> with the options `-y -m 1 -S`. The fixmate function from samtools (v 0.1.19) was used to mate the related reads after merging the results from bowtie. The mapped mates with a maximum fragment size of 2000

bp were kept and the possible PCR duplicates were further removed using Picard (<https://broadinstitute.github.io/picard>, version 1.103).

The reads in which no adapters could be found were trimmed by 5 bp to remove potential short adapter sequences which were not detected. The trimming was performed by using `fastx_trimmer` which is part of FASTX-Toolkit ([http://hannonlab.cshl.edu/fastx\\_toolkit/](http://hannonlab.cshl.edu/fastx_toolkit/), version 0.0.13). Bowtie was then used for paired-end mapping with the options `-y -m 1 -S -X 2000`, followed by PCR duplicates removal with Picard.

All of the resulting paired-end reads were combined in the final alignment file. Reads located in artefact regions (Supplementary Note 2)<sup>48</sup> were removed by samtools and reads mapped to the Y and M chromosomes were filtered out.

For each sample, the `macs2 callpeak` function was used for peak calling to get broad peaks. The broad peaks from all samples were combined to generate a consensus peak set using the `bedtools merge` function. This `bedtools merge` function was further used for each sample to count the number of reads that overlap with each peak in the consensus peak set. The resulting read numbers were used to calculate pairwise spearman's rank correlation coefficients for all samples and assess the reproducibility of the replicates (results in section 4.3.1).

### **Identification of Differential Dips**

To detect regions potentially occupied by TFs, the replicates for each condition were merged and pair-end fragments longer than 120 bp were filtered out. The remaining fragments were identified as nucleosome-free regions, which are potentially bound by TFs. In order to specify the TF binding sites, these remaining fragments were trimmed to keep only the first 10 bp from 5' to 3' direction of each read (after shifting 3 bp right for the positive strand and 1 bp left for the negative strand). These modified reads, namely the Tn5 binding sites, were used to call peak summits by `macs2` with the option `call-summits`. And then pairs of adjacent summits separated by less than 150 bp were retained. The insert regions between these summit pairs were defined as “dips” and are considered to be the TF binding sites. The R packages `GenomicRanges` and `rtracklayer` were used at this step for the calculation.

To identify the regions with differential chromatin accessibility between different time points, the pair-end fragments were analyzed by diffreps<sup>114</sup>. Differential dips between corresponding time points were then detected by overlapping the dips with the differential chromatin accessibility regions. To validate the discovered dips, the profile plot was made to show the distributions of merged ATAC-seq signals in d1 to d3 and d2 to d4 samples around Eomes and T motifs in differential dips using deepTools<sup>115</sup> (section 4.3.1, Figure 4.17).

### 3.12 Motif Analysis

Discovering TF binding sites from a set of DNA sequences during ChIP-seq and ATAC-seq analysis is important for decoding gene regulatory networks. A TF binding site can be represented by a motif, which is a sequence pattern repeatedly occurring in a group of sequences. The computational methods for motif discovery include AlignACE<sup>116</sup>, MotifSampler<sup>117</sup>, Homer<sup>118</sup>, MEME<sup>119</sup>, etc. AlignACE and MotifSampler are based on Gibbs sampling<sup>120</sup>, which was shown to be ineffective with long sequences<sup>121</sup>. MEME (Multiple Expectation Maximization Estimation) is based on EM algorithm described in section 3.5. It takes a group of sequences as input and discovers as many motifs as requested (<http://meme-suite.org/doc/meme.html>). Homer is a differential motif discovery algorithm, which aims to find motifs enriched in one group of sequences relative to the other group. The motif enrichment with Homer is determined using hypergeometric enrichment or binomial calculations<sup>118</sup>.

### 3.13 Gene Ontology (GO) Term Enrichment Analysis

After obtaining the target genes of a TF of interest from ChIP-seq, differentially expressed (DE) genes from RNA-seq or genes associated with regions of differential chromatin accessibility from ATAC-seq, we would like to functionally profile those genes, such as to determine the enriched biological processes of those genes. The Gene Ontology Consortium was formed with the purpose of constructing Gene Ontology (GO) terms to annotate genes. The GO terms describe gene function from three aspects: cellular component (the locations in the cell where a gene works, e.g., “nuclear membrane”), biological process (the biological processes a gene is involved, e.g., “mesoderm formation”) and molecular function (the function of a gene at the molecular level, e.g., “enzyme” and “transporter”)<sup>122</sup>. GO not only defines terms with respect to gene functions, but also comprises well-defined relationships between the terms, which helps to get better knowledge about the genes of interest.

GO enrichment analysis aims to find the overrepresented GO terms in a gene set of interest. There are several tools available, such as DAVID<sup>123</sup> and BiNGO<sup>124</sup>, which differ in the underlying algorithms<sup>125</sup>. The most common algorithm for GO enrichment analysis is hypergeometric distribution (Eq. 36).

$$P(X = x) = \frac{\binom{K}{x} \binom{N-K}{n-x}}{\binom{N}{n}} \quad (36)$$

In this equation,  $N$  is the total number of genes (background genes/population genes), of which  $n$  genes are of interest (e.g., differentially expressed genes from RNA-seq).  $K$ , a subset of  $N$ , is the total number of genes with the tested GO term, of which  $x$  genes belong to those  $n$  genes. The hypergeometric test uses the hypergeometric distribution to determine how significant the tested GO term is. To measure whether a GO term is overrepresented in these  $n$  genes, the p-value of the hypergeometric test is calculated as the probability or chance of seeing at least  $x$  genes out of all  $n$  genes in the list of the tested GO term. The smaller the p-value is, the more significant this tested GO term is. The hypergeometric test is the same as the one-tailed Fisher's exact test.

DAVID, the tool used in this study, is built with a modified hypergeometric test (one-tailed Fisher's exact test). To make the result more conservative, it modified the number of  $x$  in Equation 36 as  $x - 1$  and then calculated the p-value for each term. This p-value indicates whether  $(x - 1)/n$  is more likely than by random chance as compared to the background of  $K/N$ <sup>126</sup>. The enrichment p-values can be corrected to control false discovery rate by multiple testing correction methods Bonferroni, Benjamini or FDR provided by DAVID<sup>127</sup>.

### 3.14 Software

**Table 3-3 Tools used in this study**

| Software                            | Publication  | Weblink   |
|-------------------------------------|--|---|
| <b>BEDTools</b><br>(v 2.27.1)       | Quinlan and Hall, 2010                                       | <a href="http://bedtools.readthedocs.io/">http://bedtools.readthedocs.io/</a>   |
| <b>Bowtie</b><br>(v 1.0.0)          | Langmead <i>et al.</i> , 2009                                | <a href="http://bowtie-bio.sourceforge.net/index.shtml">http://bowtie-bio.sourceforge.net/index.shtml</a>   |
| <b>Cluster 3.0</b>                  | de Hoon <i>et al.</i> , 2004                                 | <a href="http://bonsai.hgc.jp/~mdehoon/software/cluster/">http://bonsai.hgc.jp/~mdehoon/software/cluster/</a>   |
| <b>Cufflinks</b><br>(v 2.1.1)       | Trapnell <i>et al.</i> , 2010; Trapnell <i>et al.</i> , 2012 | <a href="http://cole-trapnell-lab.github.io/cufflinks/">http://cole-trapnell-lab.github.io/cufflinks/</a>   |
| <b>CummeRbund</b><br>(2.24.0)       | Goff <i>et al.</i> , 2018                                    | <a href="https://bioconductor.org/packages/release/bioc/html/cummeRbund.html">https://bioconductor.org/packages/release/bioc/html/cummeRbund.html</a>   |
| <b>DAVID</b><br>(v 6.8)             | Dennis <i>et al.</i> , 2003                                  | <a href="https://david.ncifcrf.gov/">https://david.ncifcrf.gov/</a>   |
| <b>deepTools</b><br>(v 2.5.3)       | Ramírez <i>et al.</i> , 2016                                 | <a href="https://deeptools.readthedocs.io/en/develop/">https://deeptools.readthedocs.io/en/develop/</a>   |
| <b>diffreps</b><br>(v 1.55.6)       | Shen <i>et al.</i> , 2013                                    | <a href="https://github.com/shenlab-sinai/diffreps">https://github.com/shenlab-sinai/diffreps</a>   |
| <b>DREM 2.0</b>                     | Schulz <i>et al.</i> , 2012                                  | <a href="http://www.sb.cs.cmu.edu/drem/">http://www.sb.cs.cmu.edu/drem/</a>   |
| <b>Ea-utils</b>                     | /  | <a href="https://expressionanalysis.github.io/ea-utils/">https://expressionanalysis.github.io/ea-utils/</a>   |
| <b>Homer</b><br>(v 4.8.2)           | Heinz <i>et al.</i> , 2010                                   | <a href="http://homer.ucsd.edu/homer/motif/">http://homer.ucsd.edu/homer/motif/</a>   |
| <b>Java TreeView</b><br>(v 1.1.6r4) | Saldanha, 2004   | <a href="https://sourceforge.net/projects/jtreeview">https://sourceforge.net/projects/jtreeview</a>   |
| <b>MACS</b><br>(v 2.1.0)            |  |   |
| <b>MEME</b><br>(v 2.1.0)            | Bailey and Elkan, 1994                                       | <a href="http://meme-suite.org/">http://meme-suite.org/</a>   |
| <b>Meme-CHIP</b><br>(v 4.11.1)      | Machanic and Bailey, 2011                                    | <a href="http://meme-suite.org/">http://meme-suite.org/</a>   |
| <b>mergeShuffledFastqSeqs.pl</b>    | /  | <a href="https://github.com/broadinstitute/viral-ngs/blob/master/tools/scripts/mergeShuffledFastqSeqs.pl">https://github.com/broadinstitute/viral-ngs/blob/master/tools/scripts/mergeShuffledFastqSeqs.pl</a> |
| <b>Picard</b><br>(v 2.18.27)        | /  | <a href="https://broadinstitute.github.io/picard/">https://broadinstitute.github.io/picard/</a>   |
| <b>R</b><br>(v 3.5.1)               | /  | <a href="https://www.r-project.org/">https://www.r-project.org/</a>   |
| <b>SAMtools</b><br>(v 0.1.19)       | Li <i>et al.</i> , 2009                                      | <a href="http://www.htslib.org/">http://www.htslib.org/</a>   |
| <b>TopHat</b><br>(v 2.0.11)         | Kim <i>et al.</i> , 2013                                     | <a href="https://ccb.jhu.edu/software/tophat/index.shtml">https://ccb.jhu.edu/software/tophat/index.shtml</a>   |



## 4 Results

### 4.1 Time-series Transcriptome Analysis of Mesoderm Formation *in Vitro*

To determine the genome-wide transcriptome changes during the transition of stem cells to mesoderm and to assess the success of *in vitro* differentiation procedure, we performed time-series RNA-seq using the *in vitro* system of the mouse embryonic stem cells (mESCs) differentiated to mesoderm. We decided to monitor the transcriptome at 10 stages of differentiation, including undifferentiated (ES), early hourly stages (1h, 6h, 12h) and later daily stages (d1-d6). Our experimental results showed that the differentiated cells at day 6 were fully committed to cardiac mesodermal fate and at day 8 they became functional contracting cardiomyocytes (data not shown).

The analysis workflow for the time-series RNA-seq included: (1) read mapping, (2) read counting, (3) differential gene expression analysis, (4) clustering and (5) functional enrichment analysis<sup>128</sup>. From the analysis, the *in vitro* mesoderm formation system was assessed by observing whether the temporal aspect of gene expression mimics that of mesoderm formation *in vivo* (i.e., down-regulation of pluripotency genes, up-regulation of mesodermal genes, expression of the EMT genes, and high expression of the cardiovascular system associated genes at day 6). After the *in vitro* differentiation approach was validated, it was used as the foundation in order to characterize the regulators involved in this process and develop a method which allows us to use an unbiased approach for a global analysis of the molecular mechanisms underlying EMT and the formation of mesoderm.

#### 4.1.1 Differential Gene Expression Analysis and Clustering

All of the RNA-seq reads for 10 samples with 2 replicates each were treated as described in section 3.10 and the percentages of aligned reads for each sample were ranging from 53% to 62% (Supplementary Table 1). Cuffdiff was used to calculate the FPKM values, which were scaled via the median of the geometric means of fragment counts across all sample libraries. The Pearson's correlations between replicates show high reproducibility of the data (Supplementary Table 2).

To determine the differentially expressed genes across the time course, the normalized FPKM values of 10 samples were used. The genes with FPKM values lower than 1 in all

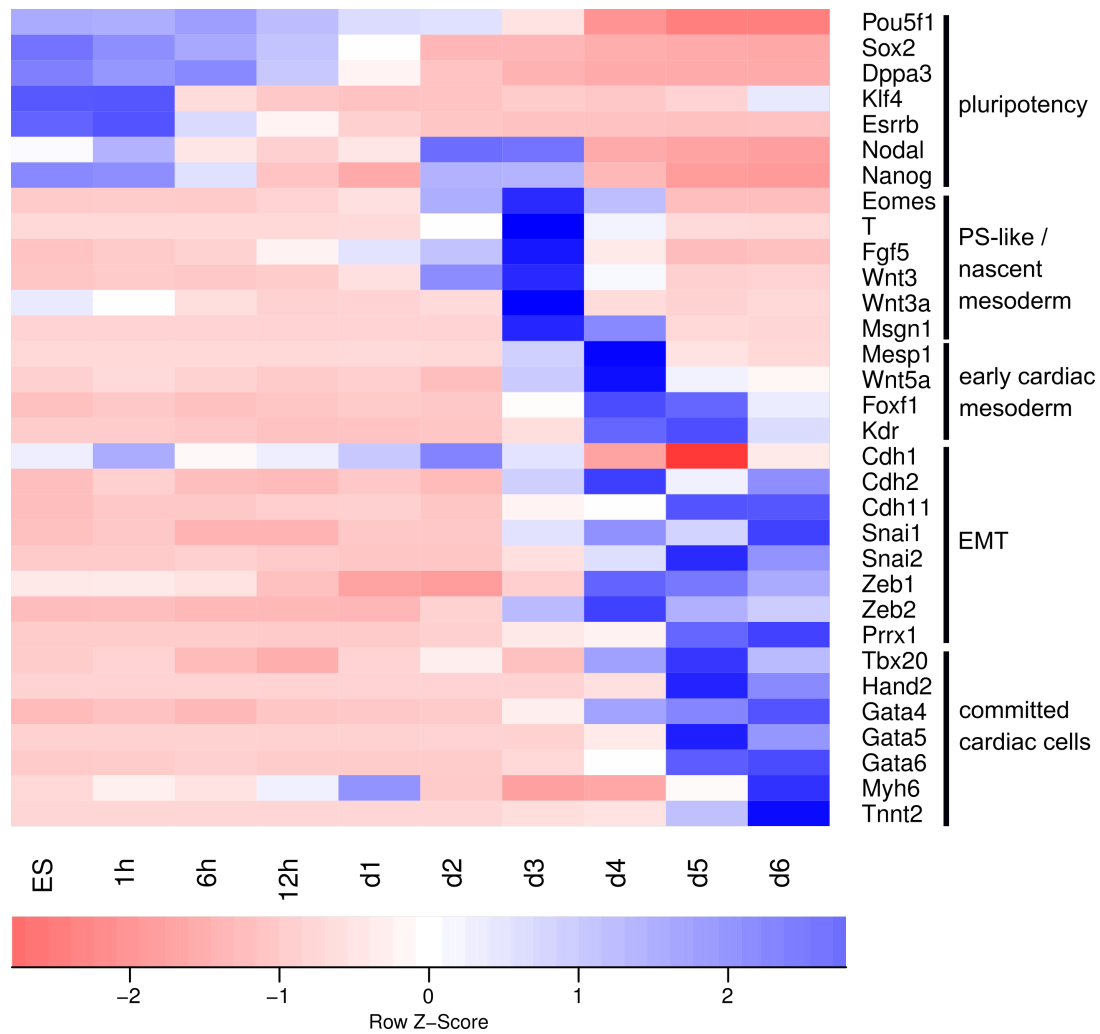
samples (defined as “not expressed”) and the genes which are not annotated in RefSeq were firstly removed. Then, log<sub>2</sub> fold change (log<sub>2</sub>FC) of the maximal FPKM against the minimal FPKM among the 10 samples was calculated for each of the genes. Application of the cutoff value of log<sub>2</sub>FC  $\geq 1$  resulted in a total of 9888 differentially expressed genes.

To define the stages of *in vitro* differentiation equivalent to *in vivo* development, the expression patterns of marker genes were observed, including markers of pluripotency, PS-like/nascent mesoderm, early cardiac mesoderm, committed cardiac cells and EMT (Figure 4.1; Supplementary Table 3).

Pou5f1, Sox2 and Nanog are pluripotency markers that define ES cells identity<sup>129</sup>, while Nodal is also involved in the development of all three germ layers<sup>130</sup>. Our data showed that the expression of pluripotency genes was relatively high at very early stages and dramatically decreased at later stages (Figure 4.1), roughly defining the time interval from ES to d1 as pluripotency/exit from pluripotency stage. d2 to d3 was defined as PS-like/nascent mesoderm stage, because PS and mesodermal marker genes, including *Bmp4*, *Eomes*, *T*, *Fgf5*, *Wnt3/3a* and *Mso1* became highly up-regulated at d2 or d3 (Figure 4.1). It has been shown that *Bmp4* homozygous mutant mouse embryos die between E6.5 to E9.5 and show little or no sign of mesoderm development<sup>131</sup>. *Wnt3a* directly targets *T* and the mutation of *Wnt3a* leads to failure of the formation of paraxial mesoderm progenitors<sup>55</sup>. Then, d4 was defined as the stage of early cardiac mesoderm, based on the high expression of *Mesp1*, *Foxf1* and *Kdr* (Figure 4.1). *Mesp1* is essential for early cardiac mesoderm formation, because it initiates its generation by controlling the expression of downstream TFs, including *Gata4* and *Hand2*<sup>39,132</sup>. *Foxf1*-deficient mice embryos show cardiac ventricular hypoplasia<sup>133</sup>. Studies in mice have shown that *Kdr* is expressed in cardiovascular progenitors which are multipotent and give rise to the three lineages of a functional heart<sup>134,135</sup>. Genes of committed cardiac cells, including *Tbx20*, *Hand2*, *Gata4*, *Gata5*, *Gata6*, *Myh6* and *Tnnt2*, were not expressed at the early stages but became highly expressed at d5 or d6 (Figure 4.1). Absence of *Gata4* and *Gata6* prevents cardiac myocyte differentiation and leads to acardia in mice<sup>136</sup>. *Myh6* is a subunit of type II myosin which is needed for cardiac muscle contraction<sup>137</sup>. *Tnnt2* is also related to cardiac muscle contraction, because homozygous knockout of *Tnnt2* produces disorganized sarcomeres and noncontractile hearts<sup>138</sup>. EMT marker genes, including *Cdh2*, *Cdh11*, *Snai1*, *Snai2*, *Zeb1*, *Zeb2* and *Prrx1*, get highly expressed during the differentiation starting at around d3 (Figure 4.1). *Cdh1* and *Cdh2* are related to cell-cell adhesion. The

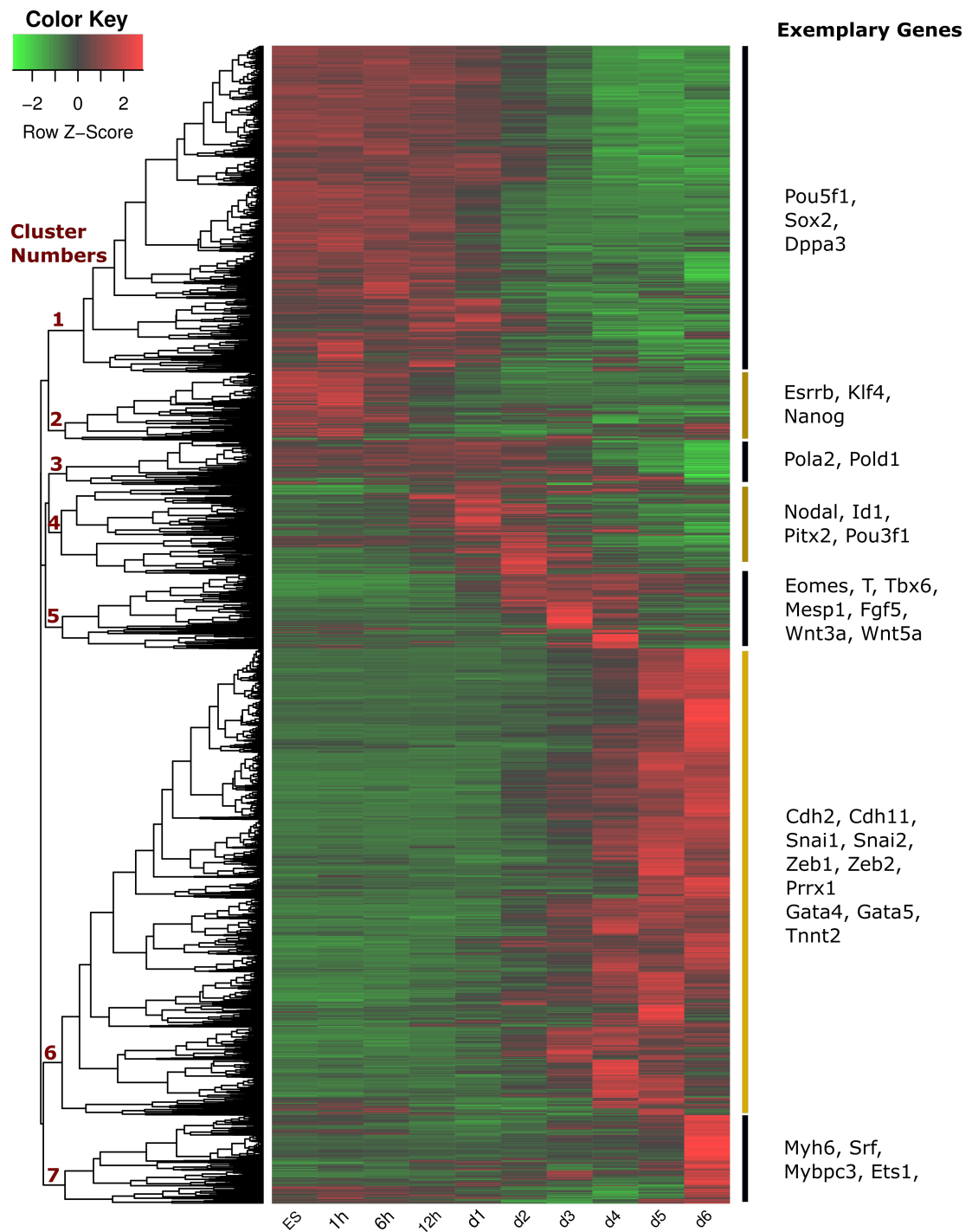
epithelial marker *Cdh1* (E-Cadherin) gets downregulated while the mesenchymal marker *Cdh2* (N-Cadherin) gets up-regulated. *Prrx1* is involved in EMT induction and its expression is restricted in mesodermal cell types during embryonic development<sup>139,140</sup>. *Snai1*, *Snai2*, *Zeb1* and *Zeb2* are transcriptional repressors triggering EMT by down-regulating *Cdh1*<sup>141,142</sup>. In total 711 differentially expressed TFs were detected during the process of mesoderm formation. A number of those TFs have been studied previously. The mechanisms of transcriptional regulation by the essential mesodermal TFs Smads, Eomes and T were characterized in detail (section 4.2). This data not only sheds new light on the function of these TFs, but also allows us to evaluate the success of our global unbiased approach to building mesodermal GRNs (section 4.3).

Genes sharing the same expression pattern are often involved in the related functions by forming gene regulatory networks. After the differentially expressed genes were obtained, they were clustered using the normalized FPKM values produced by Cuffdiff. The distance matrix was calculated by Pearson's correlation between genes. Figure 4.2 shows the clustering result. During clustering, the bigger the correlation coefficient is, the closer the genes are (Figure 4.2).



**Figure 4.1 Heatmap showing differential expression of marker genes during the time course of mesoderm formation**

The expression of marker genes serves as basis for defining the stages of *in vitro* differentiation. Pluripotency genes are highly expressed at early stages and drop significantly at later stages, defining ES to d1 as pluripotency/exit from pluripotency stage. The expression of PS-like/nascent mesoderm markers peaks at d3 and roughly defined d2 to d3 as PS-like/nascent mesoderm stage. Early cardiac mesoderm markers are highly expressed at d4. Starting at d5, cardiac mesodermal precursors and committed cardiac cells are formed. EMT starts from around d3 with the reduction of epithelial marker Cdh1 expression levels and increased expression of mesenchymal maker genes.



**Figure 4.2 Hierarchical clustering of 9888 differentially expressed genes**

Pearson's correlation was used as the distance for clustering. The bigger the correlation coefficient is, the closer the genes are. The numbers (in red) marked on the left show the seven sub-clusters (section 4.1.2). On the right are selected marker genes for each sub-cluster (section 4.1.2).

### 4.1.2 Sub-Cluster Analysis

To gain more insight about what types of genes are expressed in a similar fashion during the *in vitro* mesoderm formation process, the hierarchical clustering tree was divided into seven sub-clusters as marked in Figure 4.2 and functional enrichment analysis was then performed for genes of each sub-cluster.

Functional enrichment analysis is a common way to study big datasets, by which we can identify candidate genes or proteins sharing biological functions that are over-represented in a specific dataset. DAVID (<https://david.ncifcrf.gov/>) was used in this study to obtain enriched Gene Ontology (GO) terms<sup>122</sup> for each cluster.

Table 4-1 shows the gene expression patterns and the enriched GO terms for each of the seven clusters. In order to get the more specific GO terms for each cluster, all of the 9888 DE genes rather than all mouse genes were used as the background.

Cluster 1 included genes highly expressed at early stages which decreased quickly after d1. Those genes were related to GO terms “inner cell mass cell proliferation” (including genes *Setdb1*, *Brca2* and *Sall4*) and “stem cell population maintenance” (including genes *Pou5f1*, *Sox2*, *Dppa2*, *Fgf4* and *Sall4*). *Setdb1* is a histone lysine methyltransferase. *Setdb1*-null blastocysts exhibit defective ICM, from which mouse ES cells cannot develop<sup>143</sup>. *Sall4* is a TF expressed in early embryo and germ cells, with the expression pattern similar to *Pou5f1* and *Sox2*<sup>144</sup>. *Sall4* is required for early embryonic development and ES cells pluripotency by forming an interconnected autoregulatory network with *Pou5f1*, *Sox2* and *Nanog* in ES cells<sup>145</sup>.

Cluster 2 included genes peaking at 1h and dropping sharply afterwards, which associated with the GO term “cell adhesion” and “stem cell population maintenance”. Genes annotated with “cell adhesion” included *Jup*, *Itga1* and *Itga6*. The well-known pluripotency marker genes *Nanog*, *Tbx3*, *Esrrb*, *Klf4* and *Stat3* were also in this cluster associated with the GO term “stem cell population maintenance”.

Genes in cluster 3 on average were highly expressed at d1 and then dropped continuously. *Pola2*, *Pold1*, *Dscc1*, *Mcm7* and *Mcm8* were related to the enriched GO term “DNA replication”, while many histone-related genes, such as *Hist1b4b*, *Hist1b4f*, *Hist1b4b*, *Hist1b4i* and *Hist2b4*,

were annotated with “DNA methylation on cytosine”, “nucleosome assembly” and “protein heterotetramerization”, indicating reduced cell proliferation.

Genes in cluster 4 were highly expressed at d1 and d2. They were mainly related to “brain development” which included genes *Id1*, *Pitx2* and *Pou3f1*, indicating transient expression of neural marker genes, some of which are also important for mesoderm development.

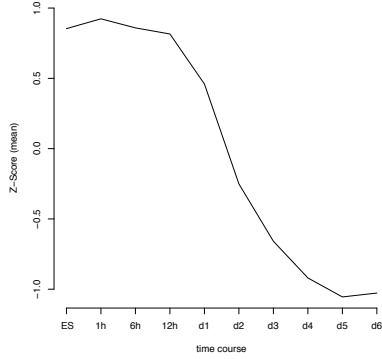
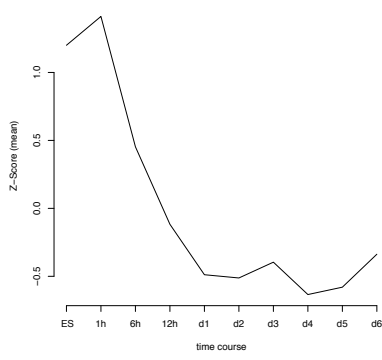
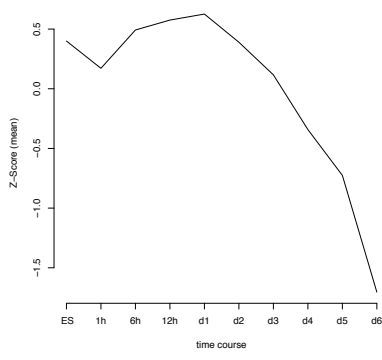
Genes in cluster 5 include those that reach maximum expression levels at d3 and d4. Manual inspection revealed that many mesodermal marker genes, including *Eomes*, *T*, *Mesp1*, *Tbx6*, *Fgf5*, *Wnt3a*, *Wnt5a*, belonged to this cluster. DAVID produced mesoderm-related GO terms such as "somitogenesis" or “lung development” as well as "Wnt signaling pathway", which plays important roles in mesoderm formation.

Genes in cluster 6 were highly expressed starting at d3 and d4, reaching maximum expression levels at d5 and d6. They were related to “heart development”, which fit the development process correctly. Cluster 6 also contained EMT marker genes, including *Cdh2*, *Chd11*, *Snai1*, *Snai2*, *Zeb1*, *Zeb2* and *Prrx1*, and well-known markers associated with cardiac cells, including *Gata4*, *Gata5* and *Tnnt2*.

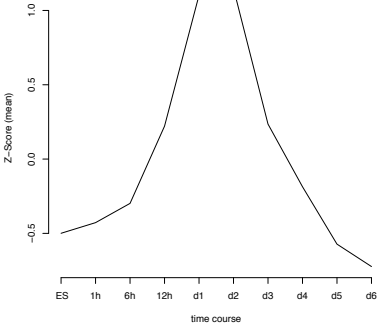
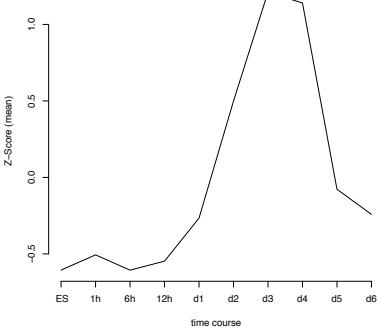
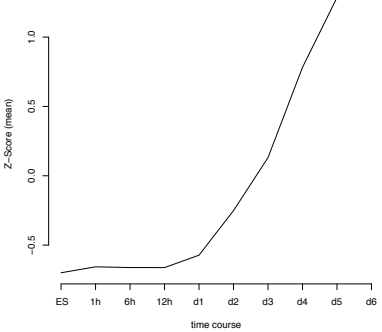
Genes in cluster 7 were highly expressed starting at d5, and included genes associated with “angiogenesis” (e.g., *Ephb2*, *Smad5*, *Xbp1*, *Ccbe1* and *Sbb*) and genes associated with “sarcomere organization” (e.g., *Myb6*, *Srf* and *Mybpc3*).

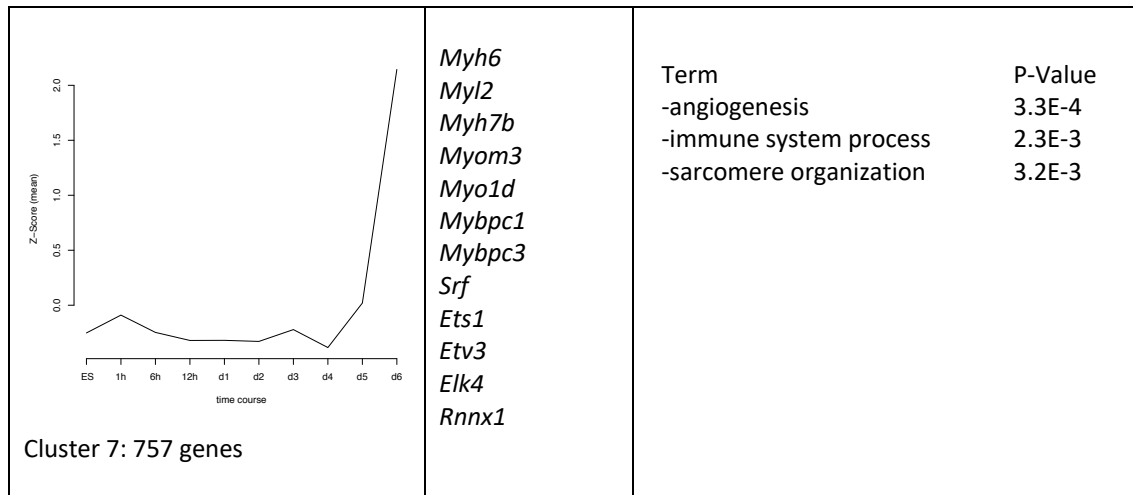
The stages of *in vitro* differentiation was previously roughly defined by manually looking at the expression patterns of marker genes: pluripotency (ES ~ d1), PS-like/nascent mesoderm (d2 ~ d3), early cardiac mesoderm (d4), committed cardiac cells (d5 ~ d6). Unbiased clustering based on expression patterns grouped functionally related marker genes. Thus, the *in vitro* differentiation system was shown to recapitulate the *in vivo* process. Moreover, the genes in the same cluster are likely to be regulated by the same mechanism and are potentially involved in the same biological processes, collaborating by forming gene regulatory networks.

**Table 4-1. GO term analysis for seven sub-clusters**

| Clusters  | Associated Genes (examples)  | GO terms   | P-Value   |
|---|--|--|---|
|  <p>Cluster 1: 2788 genes</p>  | <i>Pou5f1</i><br><i>Sox2</i><br><i>Mtf2</i><br><i>Dppa2</i><br><i>Fgf4</i>   | Term<br>-inner cell mass cell proliferation<br>-stem cell population maintenance   | 4.2E-4<br>6.7E-3                                |
|  <p>Cluster 2: 579 genes</p>  | <i>Nanog</i><br><i>Tbx3</i><br><i>Esrrb</i><br><i>Stat3</i><br><i>Tet1</i><br><i>Klf4</i><br><i>Prdm14</i><br><i>Tcl1</i><br><i>Jup</i><br><i>Gbx2</i><br><i>Gli2</i><br><i>Pou4f2</i> | Term<br>-Cell adhesion<br>-stem cell population maintenance<br>-germ cell development<br>-axon guidance                              | 2.6E-3<br>1.6E-2<br>2.0E-2<br>3.0E-2            |
|  <p>Cluster 3: 383 genes</p> | <i>Pola2</i><br><i>Pold1</i><br><i>Dscc1</i><br><i>Pten</i><br><i>Tert</i><br><i>Hdac11</i><br><i>Prmt7</i><br><i>Mcm7</i><br><i>Mcm8</i>  | Term<br>-nucleosome assembly<br>-DNA methylation on cytosine<br>-protein heterotetramerization<br>-cell division<br>-DNA replication | 4.7E-10<br>5.7E-8<br>6.9E-6<br>5.3E-4<br>2.9E-3 |



|  <p>Cluster 4: 760 genes</p>    | <p><i>Pou3f1</i><br/><i>Gabra5</i><br/><i>Id1</i><br/><i>Pitx2</i><br/><i>Cxcl12</i><br/><i>Arnt2</i><br/><i>Oxct1</i></p>   | <table border="0"> <thead> <tr> <th>Term</th> <th>P-Value</th> </tr> </thead> <tbody> <tr> <td>-brain development</td> <td>5.5E-3</td> </tr> <tr> <td>-neuromuscular process</td> <td></td> </tr> <tr> <td>controlling balance</td> <td>6.6E-3</td> </tr> <tr> <td>-synapse organization</td> <td>1.7E-2</td> </tr> </tbody> </table>  | Term | P-Value | -brain development     | 5.5E-3 | -neuromuscular process |        | controlling balance                       | 6.6E-3 | -synapse organization                 | 1.7E-2 |                                    |        |                        |        |                           |        |
|--|--|--|------|---------|------------------------|--------|------------------------|--------|---|--------|---------------------------------------|--------|------------------------------------|--------|------------------------|--------|---------------------------|--------|
| Term   | P-Value  |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -brain development   | 5.5E-3   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -neuromuscular process   |  |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| controlling balance  | 6.6E-3   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -synapse organization  | 1.7E-2   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
|  <p>Cluster 5: 638 genes</p>   | <p><i>Eomes</i><br/><i>T</i><br/><i>Mesp1</i><br/><i>Tbx6</i><br/><i>Fgf5</i><br/><i>Wnt3a</i><br/><i>Wnt5a</i></p>  | <table border="0"> <thead> <tr> <th>Term</th> <th>P-Value</th> </tr> </thead> <tbody> <tr> <td>-lung development</td> <td>1.7E-5</td> </tr> <tr> <td>-somitogenesis</td> <td>7.2E-5</td> </tr> <tr> <td>-anterior/posterior pattern specification</td> <td>1.3E-4</td> </tr> <tr> <td>-canonical Wnt signaling pathway</td> <td>5.8E-4</td> </tr> <tr> <td>-compartment pattern specification</td> <td>1.1E-3</td> </tr> </tbody> </table>   | Term | P-Value | -lung development      | 1.7E-5 | -somitogenesis         | 7.2E-5 | -anterior/posterior pattern specification | 1.3E-4 | -canonical Wnt signaling pathway      | 5.8E-4 | -compartment pattern specification | 1.1E-3 |                        |        |                           |        |
| Term   | P-Value  |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -lung development  | 1.7E-5   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -somitogenesis   | 7.2E-5   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -anterior/posterior pattern specification  | 1.3E-4   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -canonical Wnt signaling pathway   | 5.8E-4   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -compartment pattern specification   | 1.1E-3   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
|  <p>Cluster 6: 3983 genes</p> | <p><i>Cdh2</i><br/><i>Cdh11</i><br/><i>Snai1</i><br/><i>Snai2</i><br/><i>Zeb1</i><br/><i>Zeb2</i><br/><i>Prrx1</i></p> <p><i>Gata4</i><br/><i>Gata6</i><br/><i>Tnnt2</i></p> | <table border="0"> <thead> <tr> <th>Term</th> <th>P-Value</th> </tr> </thead> <tbody> <tr> <td>-cartilage development</td> <td>6.5E-6</td> </tr> <tr> <td>-heart development</td> <td>1.9E-4</td> </tr> <tr> <td>-smooth muscle cell differentiation</td> <td>1.4E-4</td> </tr> <tr> <td>-epithelial to mesenchymal transition</td> <td>2.9E-4</td> </tr> <tr> <td>-cell migration</td> <td>9.0E-4</td> </tr> <tr> <td>-BMP signaling pathway</td> <td>2.3E-3</td> </tr> <tr> <td>-blood vessel development</td> <td>2.3E-3</td> </tr> </tbody> </table> | Term | P-Value | -cartilage development | 6.5E-6 | -heart development     | 1.9E-4 | -smooth muscle cell differentiation       | 1.4E-4 | -epithelial to mesenchymal transition | 2.9E-4 | -cell migration                    | 9.0E-4 | -BMP signaling pathway | 2.3E-3 | -blood vessel development | 2.3E-3 |
| Term   | P-Value  |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -cartilage development   | 6.5E-6   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -heart development   | 1.9E-4   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -smooth muscle cell differentiation  | 1.4E-4   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -epithelial to mesenchymal transition  | 2.9E-4   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -cell migration  | 9.0E-4   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -BMP signaling pathway   | 2.3E-3   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |
| -blood vessel development  | 2.3E-3   |  |      |         |                        |        |                        |        |   |        |                                       |        |                                    |        |                        |        |                           |        |



## 4.2 Gene Regulation by Transcription Factors Smads, Eomes and T during Mesoderm Formation

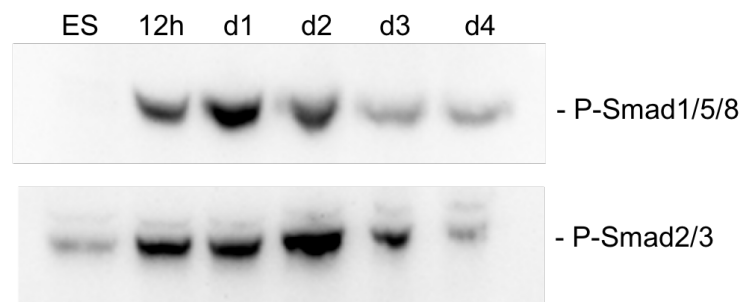
Our time course transcriptome analysis showed that the marker genes of different developmental stages followed the expected order of *in vivo* mesoderm development, suggesting that this transcriptome dataset can be used as the foundation to study the regulators involved in mesoderm formation and EMT.

Mesodermal differentiation time course transcriptome (RNA-seq) and chromatin accessibility (ATAC-seq) data were used in combination to build global gene regulatory networks in an unbiased way (section 4.3). For validation of the constructed global gene regulatory networks, we decided to build detailed GRNs centralized on the established essential mesodermal TFs Smads, Eomes and T. In addition to validating the global GRN, identification of downstream targets of Smads, Eomes and T in the same model system allowed to directly assess the extent of their cooperation and characterize the mode of regulation of their common and unique target genes. To achieve these purposes, we performed ChIP-seq and RNA-seq using wild-type and mutant cells for each TF. While the ChIP-seq data uncovers the DNA binding sites of a TF, RNA-seq from wild-type and mutant samples would show the genes up- or down-regulated by a TF. For a specific TF, by combining the results of ChIP-seq and RNA-seq, the regulated genes with TF binding sites, i.e. direct target genes, can be discovered.

### 4.2.1 Smads

R-Smads transduce the signals initiated by ligands of the TGF- $\beta$  family. In particular, Smad2 and Smad3 are activated in response to TGF $\beta$ , Activin and Nodal, while Smad1, Smad5 and Smad8 mediate BMP signaling. Smad4, which binds to phosphorylated Smad2/3 and to phosphorylated Smad1/5/8 is a common partner involved in both pathways (Figure 1.6).

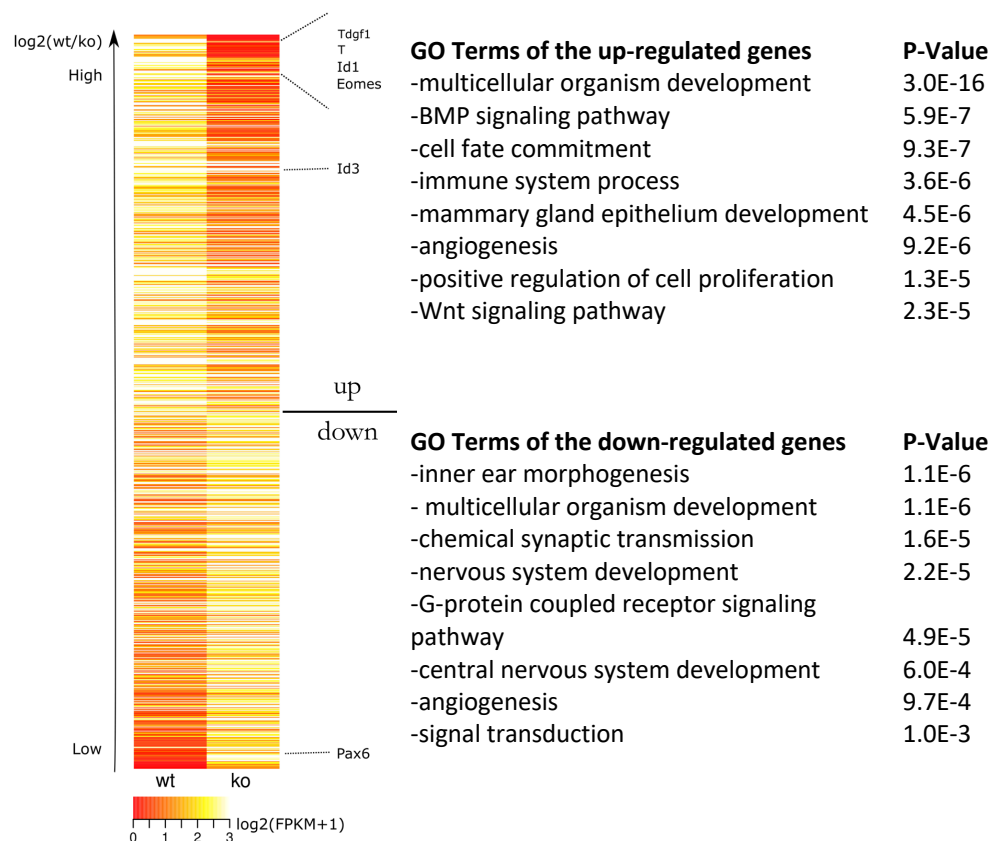
We set out to monitor the time-dependent strength of Smad-signaling by detecting the levels of phosphorylated Smad1/5/8 and Smad2/3 during the course of mesoderm differentiation. Both pathways were activated as early as 12h after Bmp4 treatment, and the levels of phosphorylated Smads reached the maximum on d1 and d2 (Figure 4.3). Therefore, we decided to perform ChIP-seq and RNA-seq using samples collected on d2 of differentiation. Using this time point allows us to study Smad signaling at the peak of its activity and obtain enough material for ChIP experiment.



**Figure 4.3 Phosphorylation levels of Smad1/5/8 and Smad2/3 proteins detected by Western blotting** (Figure provided by Dr. Pavel Tsaytler)

To identify downstream genes of Smad signaling pathways, we performed RNA-Seq for Smad4 wild-type and knockout cells. Using the cutoff  $\log_2FC \geq 0.8$ , 1062 differentially expressed genes, including 550 up- and 512 down-regulated genes, were obtained (Figure 4.4). The genes up-regulated by Smad4 included *Id1*, *Id2*, *Id3*, *TdGF1*, Wnt (*Wnt3*, *Wnt4*, *Wnt5b*, *Wnt6*, *Wnt7b*, *Wnt8a*), FGF (*Fgf8*, *Fgf17*), *Notch3*, *Nodal*, *Nanog*, *Axin2*, *Mixl1*, *Eomes* and *T*. *Id1* and *Id3* have been shown to play major roles during cardiac development<sup>146</sup>. In the mouse, loss of *Axin1* leads to embryonic lethality, accompanied by many malformations<sup>147</sup>. Strikingly, expression of mesodermal genes *Wnt3*, *Eomes*, *T*, *Fgf8* and *Pitx2* shows very strong dependency on Smad signaling. The genes down-regulated by Smad4 included *Pax6*, a key

TF regulating eye and brain development. Mutation of *Pax6* is associated with aniridia<sup>148</sup>. Functional enrichment analysis, which used all the expressed genes in wild-type or knockout samples as the background, showed that genes up-regulated by Smad4 were related to BMP and WNT signaling pathways, while genes down-regulated by Smad4 were highly related to nervous system development (Figure 4.4).

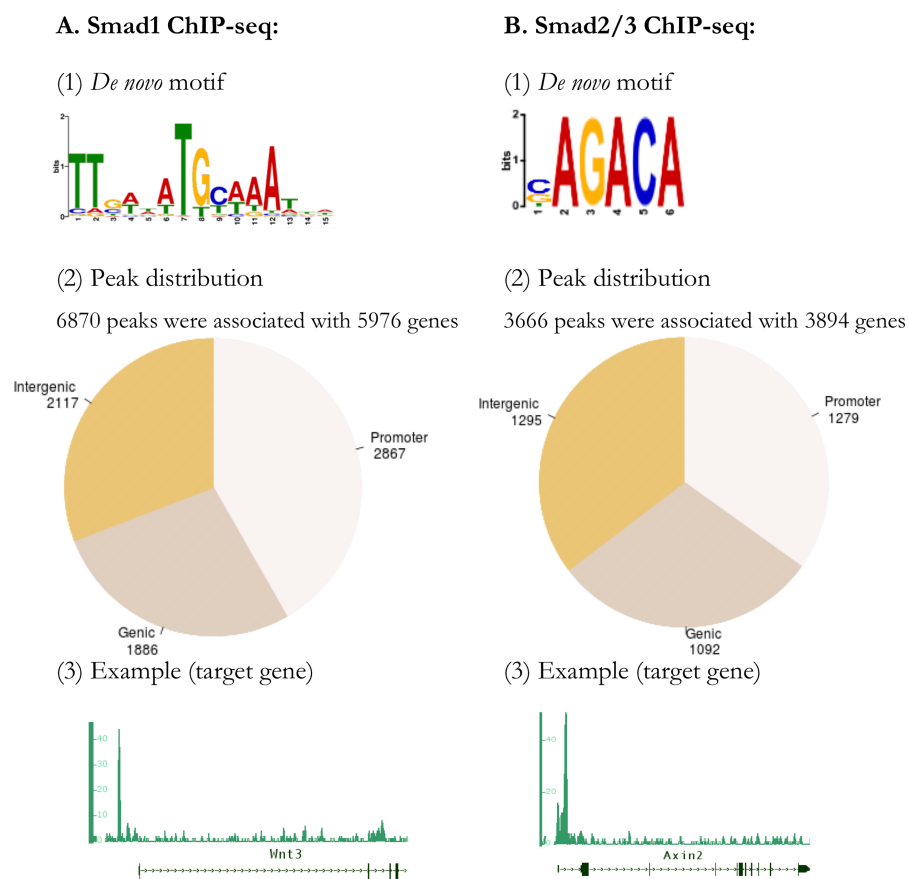


**Figure 4.4 Genes up- or down-regulated by Smad4 KO**

Top 8 GO terms for genes up- or down-regulated by Smad4 KO are shown.

To identify direct targets of Smad1/5/8 and Smad2/3 pathways, we performed ChIP-seq assays on Smad1 and Smad2/3 separately. Smad1/5/8 were shown to bind to overlapping regions, and so Smad1 binding sites represent those of Smad1/5/8<sup>24</sup>. The ChIP-seq data analysis of Smad1 by macs2 (q value 0.05) yielded 6870 high confidence peaks. The same method yielded 3666 peaks for Smad2/3 (p value 0.0002). The *de novo* motifs found using MEME are shown in Figure 4.5. The Smad1 motif resembles that of Pou5f1/Sox2 and was also shown to be the most significant Smad1 motif by Chen *et al.*<sup>34</sup>. The Smad2/3 motif is the same as Smad3 motif identified by Badis *et al.*<sup>149</sup>. The detected motifs of Smad1 and Smad2/3 validated our experiments and indicated functional binding sites.

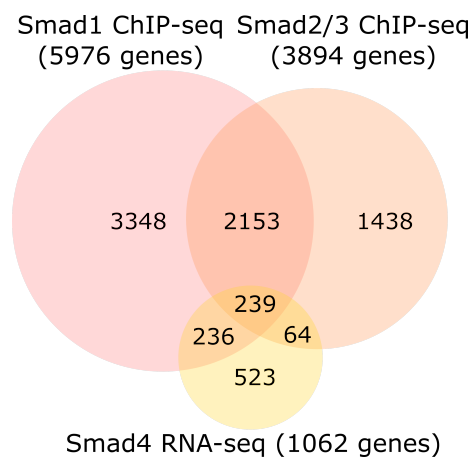
42, 27 and 31% of Smad1 peaks were located at the promoter, genic and intergenic regions respectively. 5976 genes were associated with the 6870 Smad1 peaks (Figure 4.5 A). The results of the similar analysis of Smad2/3 ChIP-seq data are shown in Figure 4.5 B. In total, 3894 target genes associated with 3666 Smad2/3 peaks were obtained. 35, 30 and 35% of those peaks were located at the promoter, genic and intergenic regions respectively. Notably, the actual number of true target genes associated with intergenic peaks is likely to be less than predicted by our approach, since each intergenic peak was associated with both up- and down-stream closest genes.



**Figure 4.5 ChIP-seq analysis of Smads**

(A) and (B) show ChIP-seq results of Smad1 and Smad2/3 respectively: (1) top binding motif; (2) peak distribution; (3) illustration of ChIP-seq binding sites on a selected target gene (*Wnt3*/*Axin2*).

By overlapping the genes associated with Smad1 and Smad2/3 DNA binding sites with the genes differentially expressed in Smad4 WT/KO cells, 539 direct targets of Smad4 were obtained (Figure 4.6). Among those 539 genes, 236 genes bound by Smad1 are unique for BMP pathway while 64 genes bound by Smad2/3 are unique for TGF $\beta$ /Nodal pathway. In addition, 239 genes are shared by both pathways. The direct target genes of both pathways include *Id2*, *Tdgf1*, *Wnt3*, *Wnt8a*, *Fgf8*, *Nodal*, *Nanog*, *Notch3*, *Axin2*, *Mixl1*, *Eomes* and *T*. In mice, the gene *Mixl1*, expressed in the PS and nascent mesoderm at the beginning of gastrulation and involved in the formation of heart and gut, is reported to be bound by Foxh1 in complex with Smad2/4 or Smad3/4<sup>150,151</sup>. Gain-of-function transgenic reporter assays showed that *Eomes* is regulated by Smad2/3 in the early mouse embryo<sup>152</sup>. It was also shown to be activated by Smad2/3, which recruit Jmjd3 to chromatin in response to Nodal signaling<sup>153</sup>.



**Figure 4.6 Venn diagram showing direct target genes of Smad1 and Smad2/3**

Direct targets of Smad1 (475 genes) and Smad2/3 (303 genes) were identified by overlapping genes related to Smads ChIP-seq peaks with DE genes from Smad4 WT/KO RNA-seq.

To determine the position of Smad targets in the context of global transcriptome changes during the differentiation, I checked the distribution of 475 Smad1 target genes and 303 Smad2/3 target genes in the seven sub-clusters from the transcriptome analysis (Table 4-1; Figure 4.2). It showed that Smad1 and Smad2/3 targets were significantly enriched in clusters 4 and 5 (section 3.2) (Table 4-2).

**Table 4-2. Enrichment analysis for target genes of Smad1 and Smad2/3**

"Cluster 1 (2788)" indicates that there are 2788 genes in cluster 1. "75 (2.7%)" indicates that 75 (or 2.7%) of the 2788 genes from cluster 1 are among the 475 Smad1 target genes. Significant p values (< 0.0001) from Fisher's exact test are marked in bold.

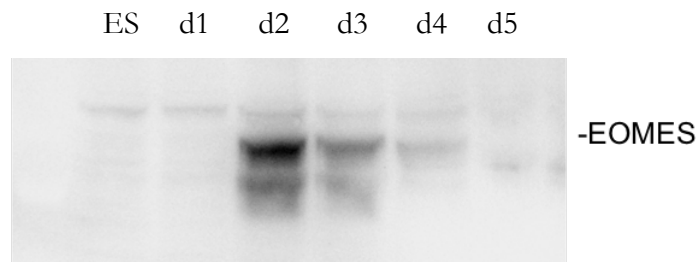
|         |                            | Cluster 1<br>(2788) | Cluster 2<br>(579) | Cluster 3<br>(383) | Cluster 4<br>(760) | Cluster 5<br>(638) | Cluster 6<br>(3983) | Cluster 7<br>(757) |
|---------|----------------------------|---------------------|--------------------|--------------------|--------------------|--------------------|---------------------|--------------------|
| Smad1   | 475<br>(direct targets)    | 75 (2.7%)           | 41 (7.1%)          | 7 (1.8%)           | 68 (8.9%)          | 67 (9.8%)          | 148 (3.7%)          | 34 (4.5%)          |
|         | Fisher's test<br>(p value) | 1                   | 0.002              | 0.999              | <b>1.469e-08</b>   | <b>2.112e-11</b>   | 0.999               | 0.504              |
| Smad2/3 | 303<br>(direct targets)    | 42 (1.5%)           | 28 (4.8%)          | 3 (0.8%)           | 43 (5.7%)          | 38 (6.0%)          | 101 (2.5%)          | 23 (3.0%)          |
|         | Fisher's test<br>(p value) | 1                   | 0.004              | 0.999              | <b>7.721e-06</b>   | <b>8.741e-06</b>   | 0.930               | 0.386              |

Genes in clusters 4 and 5 become up-regulated very early during differentiation – 12h and d1, respectively. As SMAD signaling is the earliest response to differentiation cues, it is unsurprising to detect enrichment of Smad targets in these clusters. Notably, the majority of genes in clusters 4 and 5 are not direct Smad targets, so other regulators controlling expression of those genes must exist.

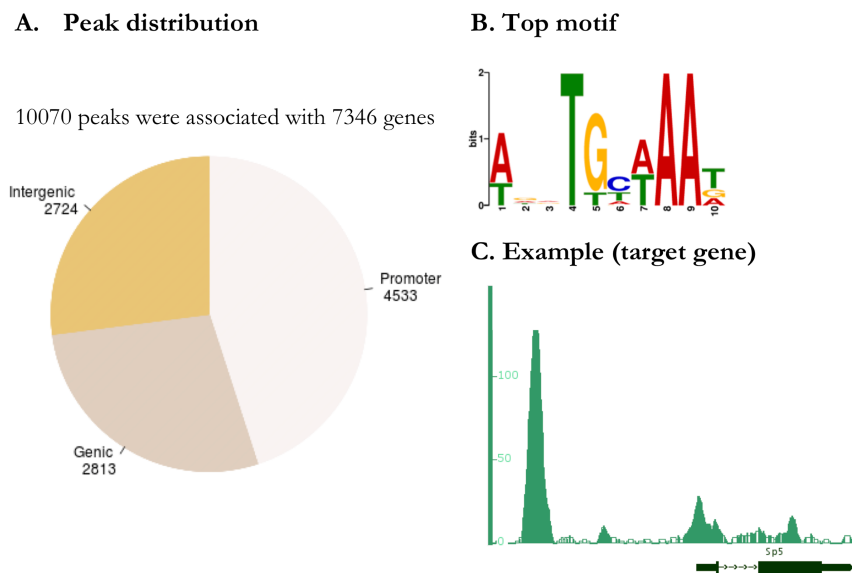
#### 4.2.2 Eomes

Eomes is required for mouse mesoderm formation by influencing the movement of cells from the epiblast to the future mesoderm<sup>36</sup>. In the time course of the mouse ES cells differentiated to mesoderm, on the protein level, Eomes was highly expressed at d2 (Figure 4.7). We therefore performed ChIP-seq and RNA-seq using samples collected on day 2 of differentiation. To build Eomes-mediated GRN, the same approach as for Smad-mediated GRN was used. Namely, ChIP-seq was performed to identify the binding sites of Eomes and then the RNA-seq in Eomes WT vs. KO cells was used to uncover the downstream target genes of Eomes. By combining ChIP-seq and RNA-seq results, direct targets of Eomes were identified.

**Figure 4.7 Expression levels of Eomes during the differentiation time course detected by Western blotting** (Figure provided by Dr. Pavel Tsaytler)



10070 Eomes peaks were obtained from the ChIP-seq analysis, among which 45, 28 and 27% of peaks were located at the promoter, genic and intergenic regions respectively. Compared to Smads and T (next section), a greater ratio of Eomes binding sites was found at promoters rather than in genic or intergenic regions (Figure 4.5, Figure 4.8, Figure 4.11). The *de novo* motif analysis with MEME for all of the peaks detected Eomes motif as the most significant<sup>149</sup>. These Eomes peaks were associated with a total of 7346 genes (Figure 4.8).

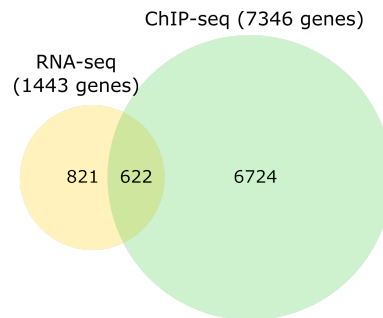


**Figure 4.8 ChIP-seq analysis of Eomes**

(A) Distribution of 10070 Eomes peaks. (B) The most significant motif from *de novo* motif analysis. (C) Illustration of ChIP-seq binding sites on a selected target gene of Eomes (*Sp5*).



By comparing the transcriptomes of Eomes WT and KO cells, 1443 differentially expressed genes were obtained. Overlapping these 1443 genes with the 7346 genes from ChIP-seq led to the identification of 622 direct target genes of Eomes (Figure 4.9). The up-regulated targets, in total 371 genes, included mesodermal markers *T* and *Fgf5*, while down-regulated targets, in total 251 genes, included *Sox2* (a marker gene of neuroectoderm), *Lef1* (a mediator of Wnt signaling), *Id1/3* (inhibitors of bHLH TFs), *Cdx1/2*, *Nkx1-2* and *Stat4*.



**Figure 4.9 Venn diagram showing direct target genes of Eomes**

622 direct target genes of Eomes were identified by overlapping genes related to Eomes ChIP-seq peaks with DE genes from Eomes WT/KO RNA-seq.

Target genes of Eomes were combined with the time-series transcriptome analysis in order to test whether Eomes targets are enriched in any specific sub-clusters (Table 4-1; Figure 4.2). The resulting overlaps are shown in Table 4-3. The target genes of Eomes are significantly enriched in cluster 4 and 5.

**Table 4-3. Enrichment analysis for target genes of Eomes**

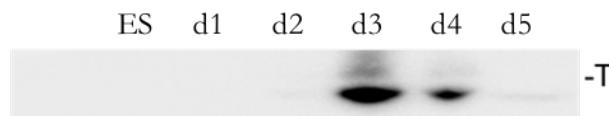
"Cluster 1 (2788)" indicates that there are 2788 genes in cluster 1. "93 (3.3%)" indicates that 93 (or 3.3%) of the 2788 genes from cluster 1 are among the 622 Eomes target genes. Significant p values (< 0.0001) are marked in bold.

|                         | Cluster 1<br>(2788) | Cluster 2<br>(579) | Cluster 3<br>(383) | Cluster 4<br>(760) | Cluster 5<br>(638) | Cluster 6<br>(3983) | Cluster 7<br>(757) |
|-------------------------|---------------------|--------------------|--------------------|--------------------|--------------------|---------------------|--------------------|
| 622 Eomes Targets       | 93(3.3%)            | 36 (6.2%)          | 21 (5.5%)          | <b>76 (10%)</b>    | <b>72 (11.3%)</b>  | 242 (6.1%)          | 28 (3.7%)          |
| Fisher's Test (p value) | 1                   | 0.333              | 0.621              | <b>9.346e-07</b>   | <b>1.572e-08</b>   | 0.132               | 0.997              |

*Eomes* itself belongs to cluster 5, where the expression levels of genes peak at d3 and d4. Cluster 5 contains many mesoderm-related genes, which apart from *Eomes* include *T*, *Mesp1*, *Tbx6* and *Wnt3a*. The enrichment of direct *Eomes* targets in cluster 5 points at its role as one of the master regulators for mesoderm formation. Among the 72 *Eomes* target genes in cluster 5, 60 and 12 genes are up- and down-regulated by *Eomes* respectively. The genes in cluster 4, with expression levels peaking at d1 and d2 and then declining dramatically, are also mostly up-regulated by *Eomes*. Among the 76 *Eomes* targets in cluster 4, 51 genes are up-regulated by *Eomes* while 25 genes are down-regulated by *Eomes*.

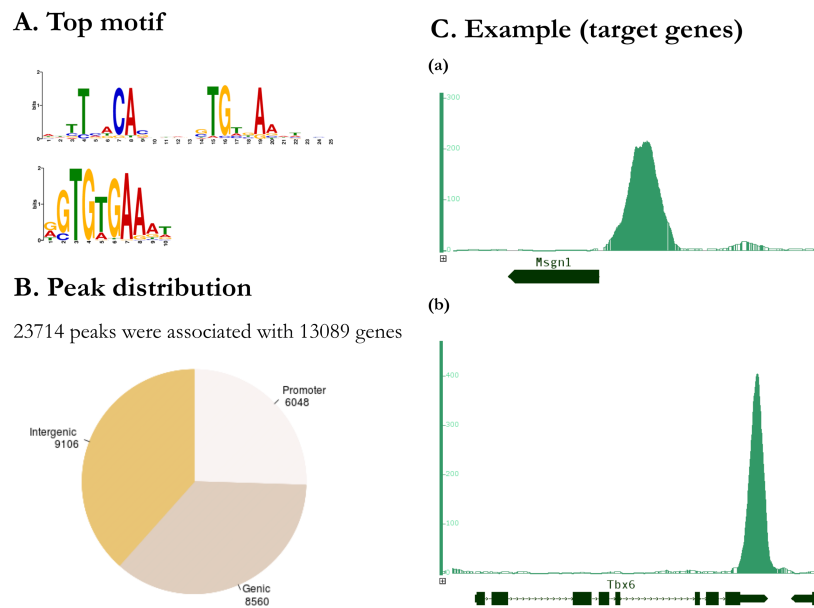
### 4.2.3 T

Previous studies have shown that T plays an important role in mesoderm formation. Homozygous mutations of T in mice result in incomplete mesoderm and the dysfunction of mesoderm-derived tissues<sup>154,155</sup>. To identify targets of T in our *in vitro* differentiation system, ChIP-seq for T and RNA-seq for T WT/KO were performed on differentiated ES cells at d3 when the expression of T reached its peak (Figure 4.10).



**Figure 4.10 Expression levels of T during the differentiation time course detected by Western blotting** (Figure provided by Dr. Pavel Tsaytler)

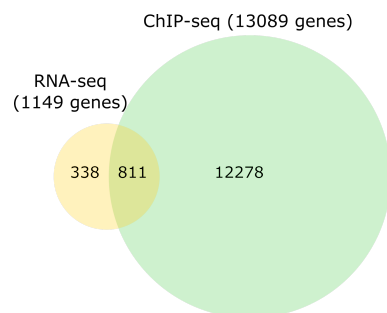
For T ChIP-seq, 23714 peaks (q value: 0.00001) called by macs2 were obtained and associated with 13089 genes (Figure 4.11). T is known to bind as a homodimer to an 18 bp palindromic motif or to a T-box motif, which comprises a half of the long palindromic motif<sup>52,53</sup>. The *de novo* motif analysis with MEME of T peaks revealed binding motifs of T on top of the list: the palindromic motif (depicted in the motif database as T\_full motif)<sup>156</sup>, and T-box motif, which is the DNA consensus sequence that can be bound by all members of T-box family. Figure 4.11 B shows the distribution of T peaks (26, 36 and 38% peaks were located at the promoter, genic and intergenic regions respectively), suggesting that T peaks are mainly enhancer associated. Figure 4.11 C shows two T targets *Msgn1* and *Tbx6* which were already shown to be regulated by T<sup>48,157</sup>.



**Figure 4.11 ChIP-seq analysis of T**

(A) Both binding motifs of T are on top of the *de novo* motif analysis result: T\_full motif and T-box motif. (B) Most of peaks are located at the intergenic regions. (C) Illustration of ChIP-seq binding sites on two selected target genes of T (*Msgn1* and *Tbx6*).

By comparing the transcriptomes of T WT and KO cells, 1149 differentially expressed genes were found. By overlapping those 1149 genes with the 13089 genes from ChIP-seq, 811 direct target genes of T were obtained, including 536 up- and 275 down-regulated genes (Figure 4.12). Mesodermal markers *Fgf8*, *Eomes*, *Mesp1* and *Lef1* were directly up-regulated by T, while *Fos*, *Id3*, *Igf2*, *Ascl2*, *Acer2*, *Heg1*, *Gata3*, *Fgf4* and *Sox2* were inhibited by T.



**Figure 4.12 Venn diagram showing direct target genes of T**

811 direct target genes of T were identified by overlapping genes related to T ChIP-seq peaks with DE genes from T WT/KO RNA-seq.

Similarly to Smads and Eomes, the distribution of T direct targets in the 7 clusters of the global time course transcriptome was determined (Table 4-1; Figure 4.2). Table 4-4 shows that they are highly enriched in cluster 5. It was shown in previous section that cluster 5 contains a significant number of Eomes target genes (11.3%) and most of the known mesodermal marker genes, including *Wnt3a*, *Eomes*, *T*, *Mesp1* and *Tbx6*. The results here show that over 25% of the genes in cluster 5 are directly controlled by T.

**Table 4-4. Enrichment analysis for target genes of T**

"Cluster 1 (2788)" indicates that there are 2788 genes in cluster 1. "102 (3.7%)" indicates that 102 (or 3.7%) of the 2788 genes from cluster 1 are among the 811 T target genes. Significant p values (< 0.0001) are marked in bold.

|                         | Cluster 1<br>(2788) | Cluster 2<br>(579) | Cluster 3<br>(383) | Cluster 4<br>(760) | Cluster 5<br>(638)           | Cluster 6<br>(3983) | Cluster 7<br>(757) |
|-------------------------|---------------------|--------------------|--------------------|--------------------|------------------------------|---------------------|--------------------|
| 811 T Targets           | 102<br>(3.7%)       | 49<br>(8.5%)       | 16<br>(4.2%)       | 76<br>(10%)        | <b>160</b><br><b>(25.1%)</b> | 304<br>(7.6%)       | 60<br>(7.9%)       |
| Fisher's Test (p value) | 1                   | 0.278              | 0.999              | 0.012              | <b>&lt;2.2e-16</b>           | 0.662               | 0.450              |

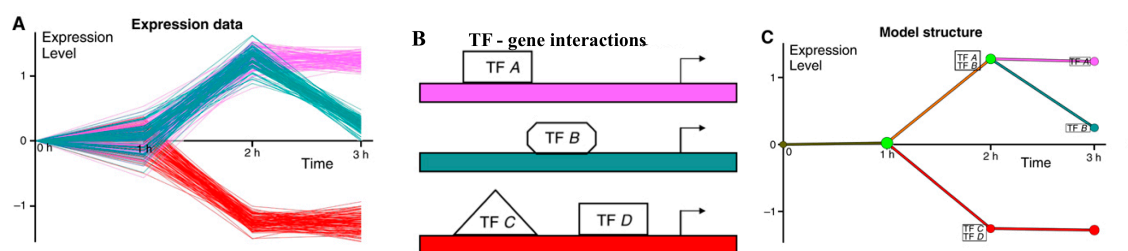
T and Eomes are both T-box TFs. Our results showed that the direct targets of Eomes and T are both enriched in cluster 5 of the time-series transcriptome data (Table 4-3 and 4-4), including 189 genes, of which 29 are targets of Eomes, 117 are targets of T, and 43 are targets of both. Furthermore, 44% (4480 out of 10070) Eomes ChIP-seq peak summits were within 500 bp distance from T summits (Supplementary Figure 2 A), and Eomes and T share many downstream genes (Supplementary Figure 2 B). Previous studies have shown that Eomes and T bind to the same genome regions during gastrulation in *Xenopus*<sup>60</sup> and that the genomic binding sites of T are in close proximity to those of Eomes in differentiating human ES cells<sup>35</sup>. These observations suggested that Eomes and T might have overlapping activities. To study the combinatorial functions of Eomes and T, we performed RNA-seq for Eomes knockout, T knockout and Eomes/T double knockout cells at d3 of differentiation. The k-means clustering of DE genes produced different categories with distinct gene expression patterns (Supplementary Figure 3)<sup>158,159</sup>. It clearly showed that a great portion of DE genes were regulated by the combination of Eomes and T (such as genes in cluster 1 to 4), while others depended on only one of these TFs (such as genes in cluster 5 to 7) (Supplementary Figure 3). Our preliminary basic analysis of Eomes and T peaks associated with genes in the k-means clusters did not show any significantly enriched peak patterns for any of the clusters

(data not shown). In the future, more tests should be conducted to decipher the mechanisms of combinatorial Eomes and T interactions.

### 4.3 Reconstruction of the Dynamic Regulatory Network Underlying Mesoderm Formation

Combining ChIP-seq and RNA-seq for specific TFs to study the GRNs is limited to TFs for which KO cell lines and good quality antibodies are available. Otherwise, they have to be produced first, which is costly and time consuming. To overcome this limitation, we decided to use the following approach for global characterization of the molecular mechanisms of mesoderm formation process.

The dynamic regulatory network of the whole mesoderm differentiation process was built by combining the time-series RNA-seq transcriptome dataset with the time-series ATAC-seq dataset. In general, this approach consists of three major steps. In the first step, the time-series RNA-seq data was used to cluster genes into paths that exhibit defined unique expression patterns based on Hidden Markov Model (Figure 4.13 A). In the second step, the ATAC-seq data was used to generate a TF-target relationship table (Figure 4.13 B). In the third step, the TFs were assigned to the paths based on the enrichment analysis of TF targets among the genes in the paths. This approach results in building the global gene regulatory network (Figure 4.13 C) underlying the process of mesoderm formation. This network allows us to predict TFs responsible for regulation of a subset of genes at every point of differentiation. Combining this method with GO term analysis reveals TFs associated with the determination of various cell fates (Figure 4.13 C).



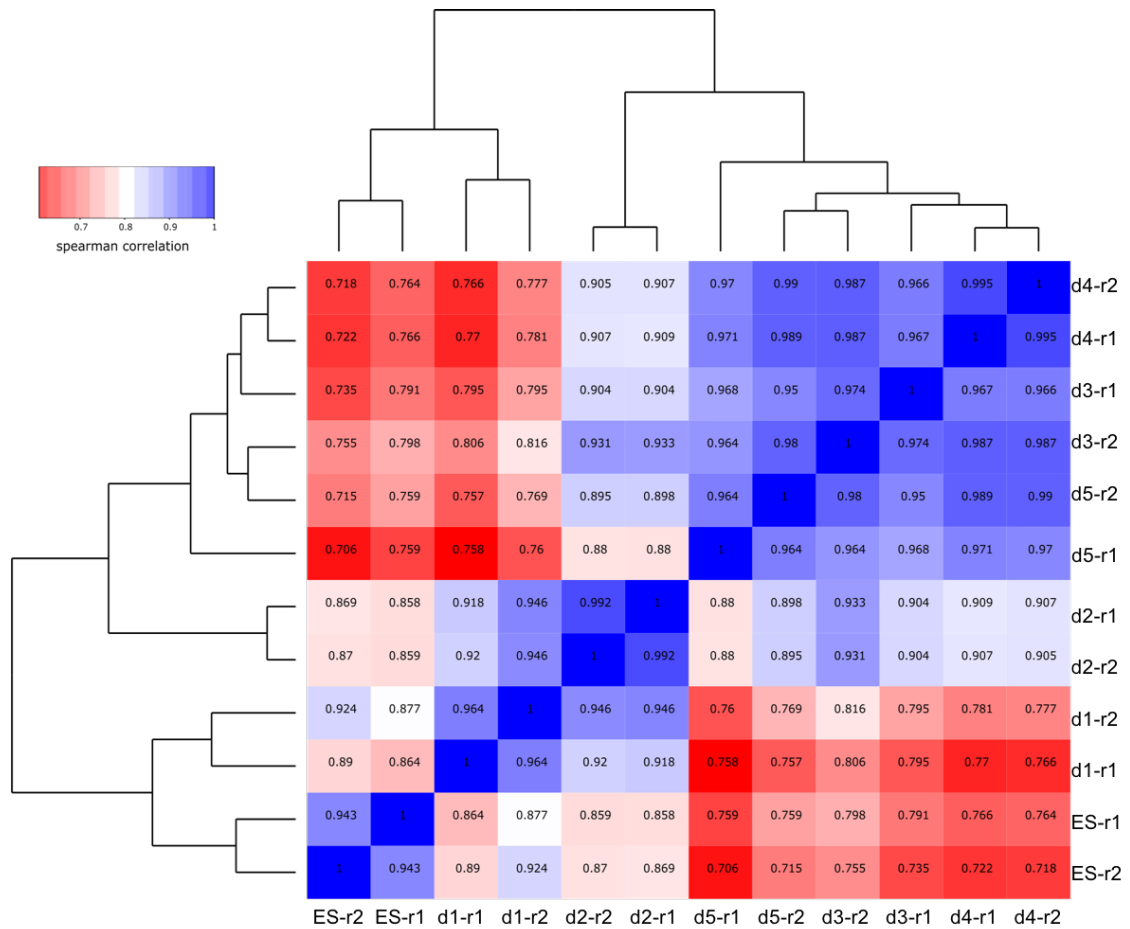
**Figure 4.13 Three major steps to build the dynamic gene regulatory network/tree** (A) Clustered time-series expression data. (B) TF-gene interactions. In this example, genes are clustered into 3 groups. Most of the genes in pink, blue and red paths are regulated by TF A, TF B, TF C/D, respectively. (C) The model structure generated from data (A) and (B). Figure taken and edited from Ernst *et al.* (2007)<sup>66</sup>.

The first and third steps of this approach were performed with DREM as described in section 3.7<sup>102,160</sup>. The advantages of this approach rely on integrating ATAC-seq and motif discovery analyses, which allows us to locate the potential binding sites of any TF with known motif. The chromatin regions that open or close differentially over time usually reflect that certain TFs exert their function there. I detected differentially open regions, discovered TFs that could potentially occupy those regions and associated them to the neighboring genes thus building the TF-target interaction table. This table was then used to calculate the enriched TFs for each path in the global gene regulatory network.

In the end, the final global network was validated by comparing it to the detailed GRNs of Smads, Eomes and T.

#### **4.3.1 Inferring TF targets from ATAC-seq Data**

ATAC-seq was performed for 6 time points (ES, d1 to d5) with 2 replicates for each condition. All of the reads were treated as described in section 3.11 and the percentages of aligned reads for each sample are shown in Supplementary Table 4. The Spearman's correlations between different samples show that the replicates are reproducible (Figure 4.14).



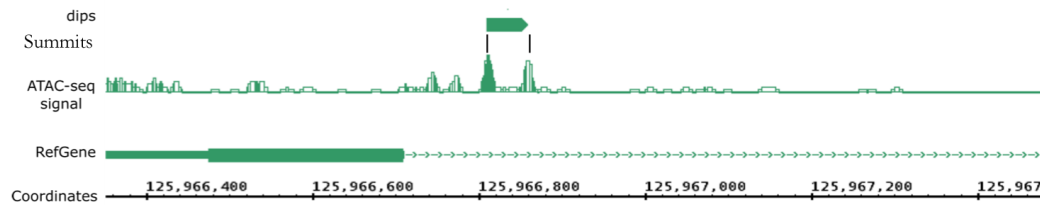
**Figure 4.14 Spearman's rank correlation coefficients between samples**

The heatmap of clustering based on the Spearman's correlations (listed in the cells) between different samples shows that the replicates are reproducible. r1 and r2 indicate replicates 1 and 2 respectively.

To find TFs binding to differentially accessible chromatin regions between two samples, for example d2 and d3, the ATAC-seq “dips” of sample 1 and sample 2 were defined separately, as depicted in the scheme of ATAC-seq analysis procedure (Figure 4.17). The “dips” are the chromosome locations potentially bound by TFs (section 3.11).

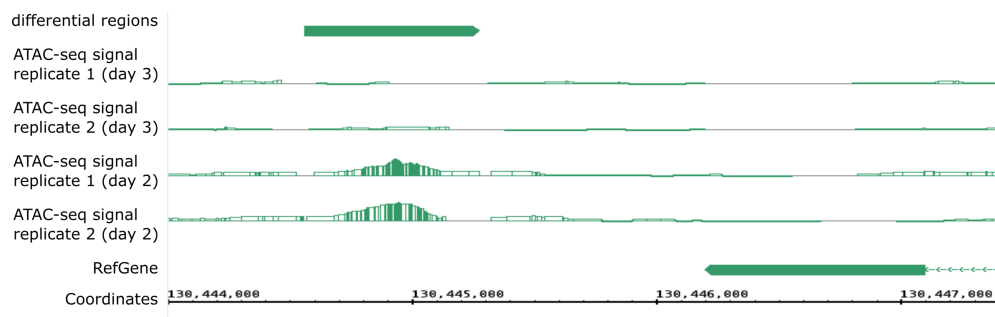
The size distribution of mapped ATAC-seq fragments shows that fragments longer than 120 bp represent DNA regions occupied by nucleosomes<sup>13</sup>. Since we were interested in nucleosome-free TFs binding regions, fragments with a length of less than 120 bp, which correspond to regions devoid of nucleosomes, were only kept (Supplementary Figure 1). Then, to detect transposase insertion sites, the remaining reads were modified as described in section 3.11. Peak calling was performed after combining the modified reads of both

replicates. I then only kept the peak pairs with a distance of less than 150 bp between two peak summits and defined the insert regions between those peak pairs as dips (Figure 4.15). The dips represent genomic regions protected from Tn5 binding and transposition, and since these regions are nucleosome-free, they likely represent binding sites of TFs.



**Figure 4.15 An example of the identified dips**

The next step was to detect dips located at regions that undergo changes in chromatin accessibility during the time course. For that differential regions were firstly detected with "diffreps"<sup>114</sup> using mapped ATAC reads (Figure 4.16). Then, by overlapping all of the dips of sample 1 and 2 with the differential regions between those two samples, the differential dips were defined. The same procedure was applied to find the differential dips between all adjacent time points and the results are summarized in Table 4-5.

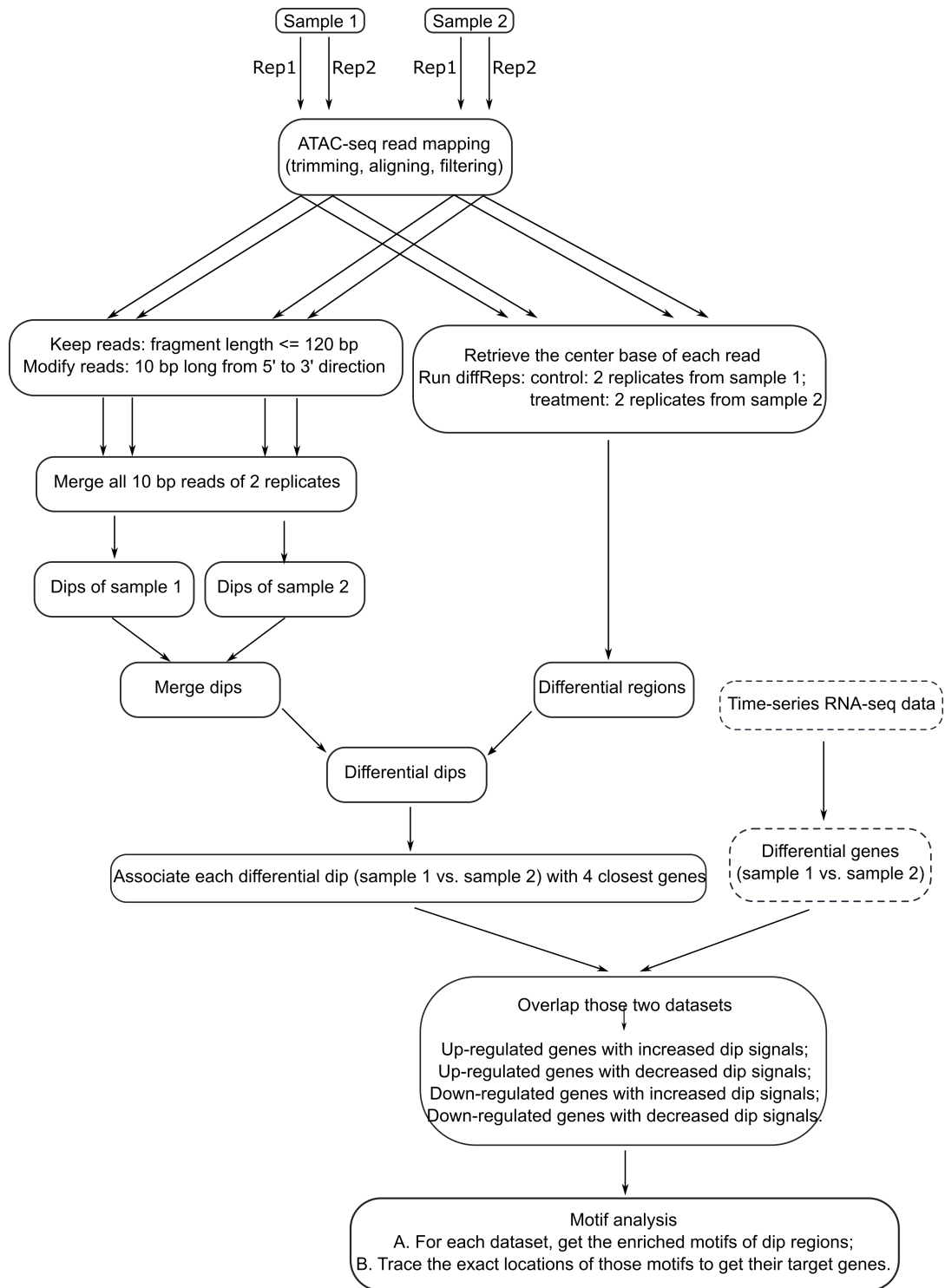


**Figure 4.16 An example of the identified differential regions**

**Table 4-5. Differential regions and differential dips between samples**

|                             | <i>ES vs. D1</i> | <i>D1 vs. D2</i> | <i>D2 vs. D3</i> | <i>D3 vs. D4</i> | <i>D4 vs. D5</i> |
|-----------------------------|------------------|------------------|------------------|------------------|------------------|
| <i>Differential regions</i> | 16090            | 19854            | 16558            | 4377             | 3367             |
| <i>Differential dips</i>    | 34321            | 32166            | 42013            | 10272            | 8008             |

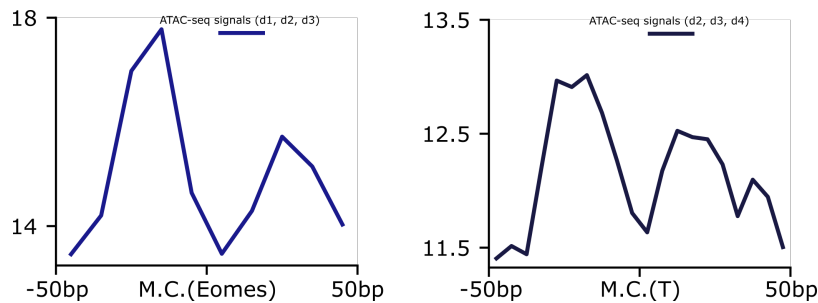




**Figure 4.17 Schematic procedure of ATAC-seq analysis to build TF-target gene interactions**

The steps in the dotted circles are from a parallel RNA-seq data analysis.

To assess the validity of the (differential) dips, I made use of the Eomes and T ChIP-seq datasets. In particular, I detected the precise locations of Eomes motif in 1799 sequences and T motif in 4787 sequences and overlapped them with the differential dips. Then, the profile plot was created to show the distributions of merged d1, d2, d3 ATAC-seq signals around Eomes motifs and d2, d3, d4 ATAC-seq signals around T motifs. The results show that the motifs of Eomes and T are on average located within the dips (Figure 4.18).



**Figure 4.18** The distributions of merged d1, d2, d3 ATAC-seq signals around Eomes motifs and d2, d3, d4 ATAC-seq signals around T motifs

\*M.C. indicates “Motif Center” (*e.g.*, The motif center for Eomes is calculated as  $10/2$ , since Eomes motif is 10 bp long).

The differential dips were then associated with four closest genes, including two upstream and two downstream genes. Those genes can potentially be regulated by the TFs located at the differential dips. Then only genes that are differentially expressed between two consecutive time points were kept. For example, for d2 vs. d3, the genes associated with the differential dips of d2 vs. d3 were overlapped with the differentially expressed genes of the same two days from the transcriptome RNA-seq analysis (Figure 4.17). To determine which TF binding motifs were enriched at the differential dips associated with differentially expressed genes, motif analysis tool Homer was used. At this step, the enriched motifs were obtained, but Homer did not output their genomic locations. Thus, as a separate step, their exact locations within differential dips were then identified by a trace-back analysis with Homer to build their connections with their target genes. Using the resulting datasets, the list of TF-target gene interactions were built for every time point in the form represented in Table 4-6. The final list of TF-target gene interactions is the combination of all time points. For each specific TF-target interaction, it is kept in the final list only if this TF is expressed (“ $\text{FPKM} \geq 1$ ” and “ $\text{FPKM} \geq 0.2 \times \max(\text{FPKM of any time point})$ ”) at either the current

or the next time point (Table 4-6). Notably, during the motif analysis, for a TF which belongs to a TF group (where all TFs have the similar motif) (Supplementary Table 5), all members in this group share all of the target genes associated with this group.

**Table 4-6. Final list of TF-gene interactions**

The columns are TFs, target genes, input value (1: the TF-target gene pair is present; 0: the TF-target gene pair is not present) and time points. Part of the results are shown in this table.

| TF    | Gene    | Input | Timepoint |
|-------|---------|-------|-----------|
| ATF1  | AACS    | 1     | ES        |
| ATF1  | ABCA1   | 1     | ES        |
| ...   | ...     | ...   | ...       |
| ELF1  | MEP1B   | 1     | d1        |
| ELF1  | METRNL  | 1     | d1        |
| ...   | ...     | ...   | ...       |
| EOMES | IGFBP3  | 1     | d2        |
| EOMES | IL33    | 1     | d2        |
| ...   | ...     | ...   | ...       |
| T     | SAMD3   | 1     | d3        |
| T     | SCARA3  | 1     | d3        |
| ...   | ...     | ...   | ...       |
| FOXA2 | CCDC162 | 1     | d4        |
| FOXA2 | CCDC40  | 1     | d4        |
| ...   | ...     | ...   | ...       |
| NANOG | BLNK    | 1     | d5        |
| NANOG | BMF     | 1     | d5        |
| ...   | ...     | ...   | ...       |

**4.3.2 Reconstructing the Dynamic Regulatory Network Controlling Mesoderm Formation**

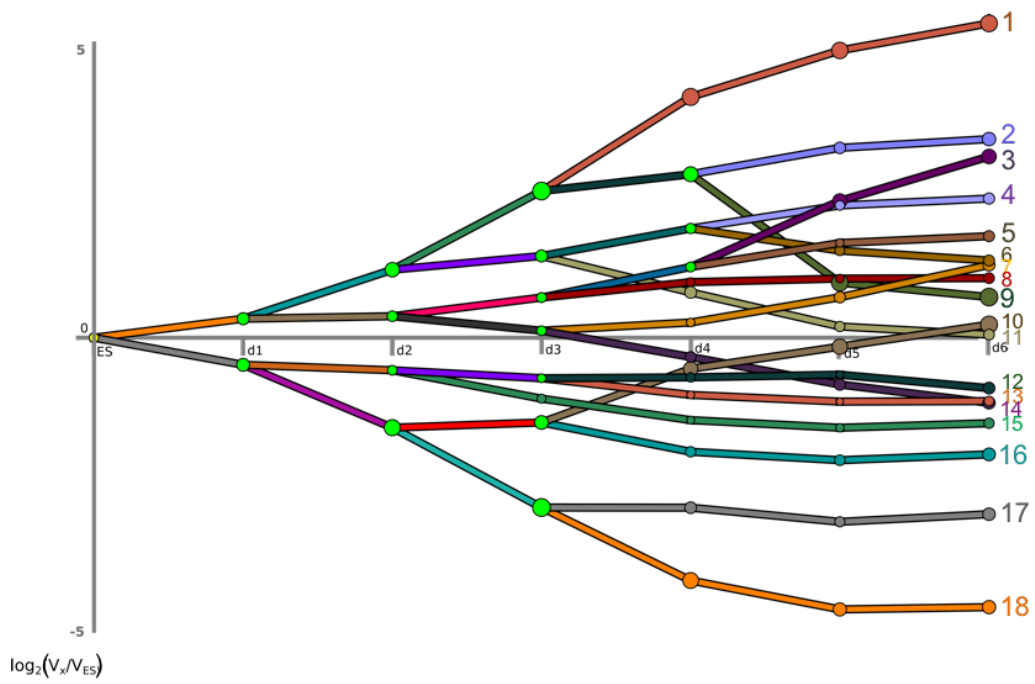
The dynamic TF-target gene interaction table created as described in section 4.3.1 contains TFs that can either induce opening of the chromatin or bind to open chromatin regions. This TF-targets dataset was integrated with our gene expression data from the time-series RNA-seq assay (ES and samples from d1 to d6) and used as the input for DREM to identify major regulators underlying the process of mesoderm formation.

The gene pattern clustering was calculated using the time-series RNA-seq gene expression values. The FPKM values were firstly adjusted by adding 1 to avoid  $V_{ES} = 0$ , where  $V_{ES}$  indicates the gene expression value at the time point ES, and then normalized to  $\log_2(V_x/V_{ES})$ , where  $V_x$  indicates the gene expression value for the time point x, i.e., d1 to

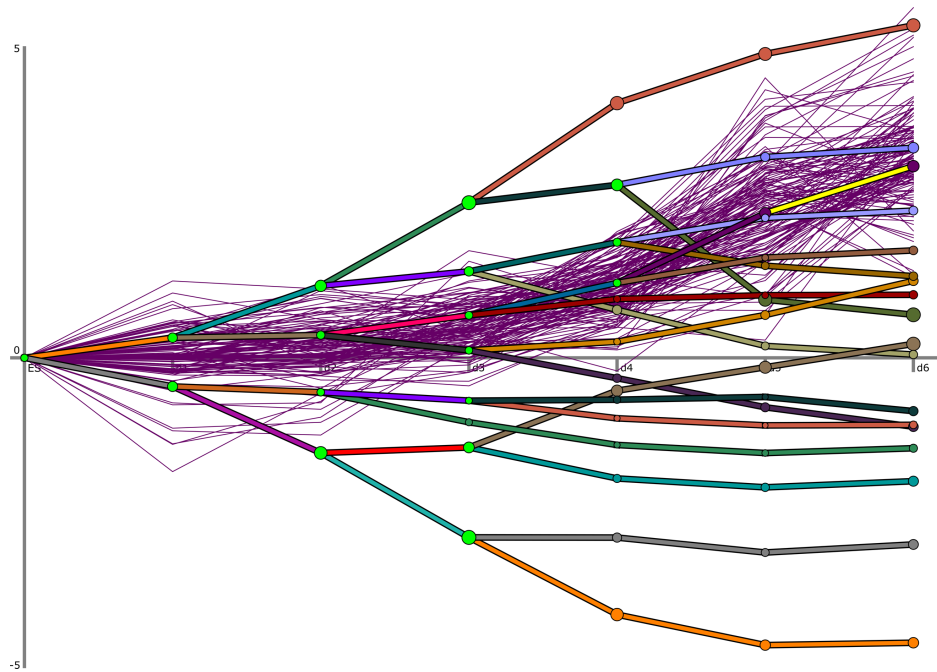
d6. The genes with normalized FPKM values lower than 1 for all time points, considered not expressed during the time course, were then removed from the list.

For the parameters to build the clustered gene expression tree, all input genes were used to evaluate and select the model, which corresponded to the “Penalized likelihood” framework in DREM. The maximum number of paths out of a split was set to two and the paths were not allowed to merge again after splitting. The parameters of convergence likelihood and minimum standard deviation were set to 0.01 and 0.2 respectively (section 3.7). At this step, the TF-target gene interactions were not allowed to influence the bifurcation of paths, i.e., the clustering was only based on the gene expression values. Using these parameters, the global GRN in a tree structure was constructed, which at the final time point consisted of 18 paths (which we also refer to as branches) (Figure 4.19 A). To assess whether this tree generated on the basis of gene expression values is suitable for further analysis, I checked how well each path represents the actual trajectories of genes it contains. An example of trajectories of genes comprising path 3 is shown in Figure 4.19 B. Manual inspection of gene trajectories of each of the 18 paths indicated that the generated tree accurately detects and combines genes with common expression patterns and can be used for discovering TFs that regulate transcription during the formation of mesoderm (section 3.7). The GO term analysis (with all genes at the time point ES used as the background) was performed for the 18 final time point paths to show the gene functions related to each path and to further assess the validity of the tree structure. Biologically meaningful GO terms were enriched for most of the paths (Supplementary Table 6). For instance, the genes in path 9, highly expressed at d3 and d4, include those related to gastrulation (*Eomes*, *Mesp1* and *Mixl1*), somitogenesis (*T*, *Msgn1* and *Axin2*) and mesodermal cell migration (*Fgf8* and *Mesp1*). The genes in path 18, with the expression level steadily decreasing from the beginning of differentiation, include those related to stem cell population maintenance (*Pou5f1*, *Esrrb*, and *Sox2*).

A.



B.



**Figure 4.19 The tree structure constructed using the time-series expression data**  
A. The horizontal axis indicates time points and the vertical axis indicates  $\log_2(V_x/V_{ES})$ , where  $V_x$  is the expression value for each corresponding time point. The numbers indicate the paths from 1 to 18 and the corresponding GO terms for each path are listed in Supplementary Table 6. B. An example of trajectories of genes comprising path 3.

As is evident from the gene regulatory tree, during the differentiation time course, groups of genes that exhibited similar expression pattern until a particular time point, start to separate and form smaller gene groups of divergent fates. The ultimate goal of our approach is to identify TFs that selectively control these bifurcation events. We instruct DREM, using our TF-target gene interaction dataset, to identify target genes enriched in every branch of a bifurcation event and to assign corresponding TFs to these branches. Since the TF-target gene interactions used for the computation of the TF assignment are time-point specific and the number of genes per path as well as the corresponding background are different, enrichment cutoff  $X$  (corresponding to  $p$  value  $p = 10^{-X}$ ) of various stringencies was used to assign TFs for every time point to keep the most significant and biologically meaningful TFs (ES to d1:  $X=4$ ; d1 to d2:  $X=12$ ; d2 to d3:  $X=4$ ; d3 to d4:  $X=1.5$ ; d4 to d5:  $X=12$  and  $X=3$ ). Separate figures of the regulatory trees were therefore generated to show for each time point the enriched TFs assigned to the corresponding paths (Supplementary Figure 4).

Although our approach predicts TFs that regulate the bifurcation events at all time points of the differentiation time-series, we are mainly interested in the process of mesoderm formation and EMT. Therefore, the paths which are more relevant to these processes were analyzed in more detail. The GO term analysis indicated that the genes in path 1, which undergo the strongest up-regulation during the mesoderm formation, contained cardiovascular-related terms, such as "cardiac muscle contraction" and "angiogenesis" (Supplementary Table 5). Similar terms were enriched for the genes of the up-regulated path 3. Path 4 had a clear enrichment of EMT-related terms, while the genes in path 18, which undergo the strongest downregulation, contained terms related to pluripotency maintenance. In line with this observation, we closely followed the bifurcation events leading to paths 18 (in comparison to path 10; Figure 4.20), 1 (in comparison to paths 2 and 9; Figure 4.21), 3 and 4 (in comparison to path 11; Figure 4.22) (Supplementary Table 6). We did not put emphasis on the other paths (such as 12 to 17), because, either the gene expression levels are not changing significantly during differentiation, or the GO terms are mainly metabolic-associated, indicating those genes are mainly housekeeping genes with general cellular functions.

## **Bifurcation events in the dynamic regulatory network controlling stem cell population maintenance genes**

The genes in the path 18 (Figure 4.20) were continuously repressed during the time course. The GO term analysis of the genes comprising this path revealed enrichment of the term “stem cell population maintenance” (Supplementary Table 6). Indeed, this path included 78 genes such as the pluripotency-related factors Pou5f1, Esrrb, Sox2, Nanog and Tet1. Down-regulation of pluripotency genes is essential for the ES cell differentiation. And while TFs that repress pluripotency factors are not likely to be the main effectors of mesoderm formation, we decided to treat path 18 as a control group and observed the bifurcation events leading to its formation starting from ES cells (Figure 4.20).

From the time point ES to d1, the genes comprising the tree are split in 4319 up- and 2658 down-regulated genes (path A). To determine if genes responsible for any biological processes or cell fates are enriched among the up- and down-regulated pathways, GO term analysis was performed. The upward path genes were associated, albeit with low significance, to terms “signal transduction”, “cell migration” and “cell adhesion”, while the most significant terms associated to path A were metabolism-related (Supplementary Figure 4 A).

At the ES to d1 stage, the genes are associated to up- or downward branches mainly based on their expression at later time points. Thus, Figure 4.19 B indicates that about half of the genes from path 3, which stems from the upward branch of ES to d1 stage, are in fact not affected or even down-regulated at this stage. Moreover, the absolute average values of  $\log_2(V_{d1}/V_{ES})$  for both up- and downward branches are lower than 1, raising the question how many of the 6977 genes in the tree are in fact differentially expressed from time point ES to d1. To evaluate this, I compared our ES and d1 transcriptome data. It yielded 1707 DE genes (based on the cutoff:  $\log_2FC \geq 1$ ). Although 1408 out of 1707 DE genes are in the up- or downward branches of ES to d1, the majority of genes comprising these two branches are not differentially expressed. It suggests that this tree is suboptimal to study the TF regulators of early transcriptional events. The better approach would be to build a different tree using our transcriptome data for ES, 1, 6, 12 and 24h. Despite this limitation, our method predicts that TFs such as Smad2/3, Nanog, Esrrb and Stat3 are involved in the repression of genes in the path A.

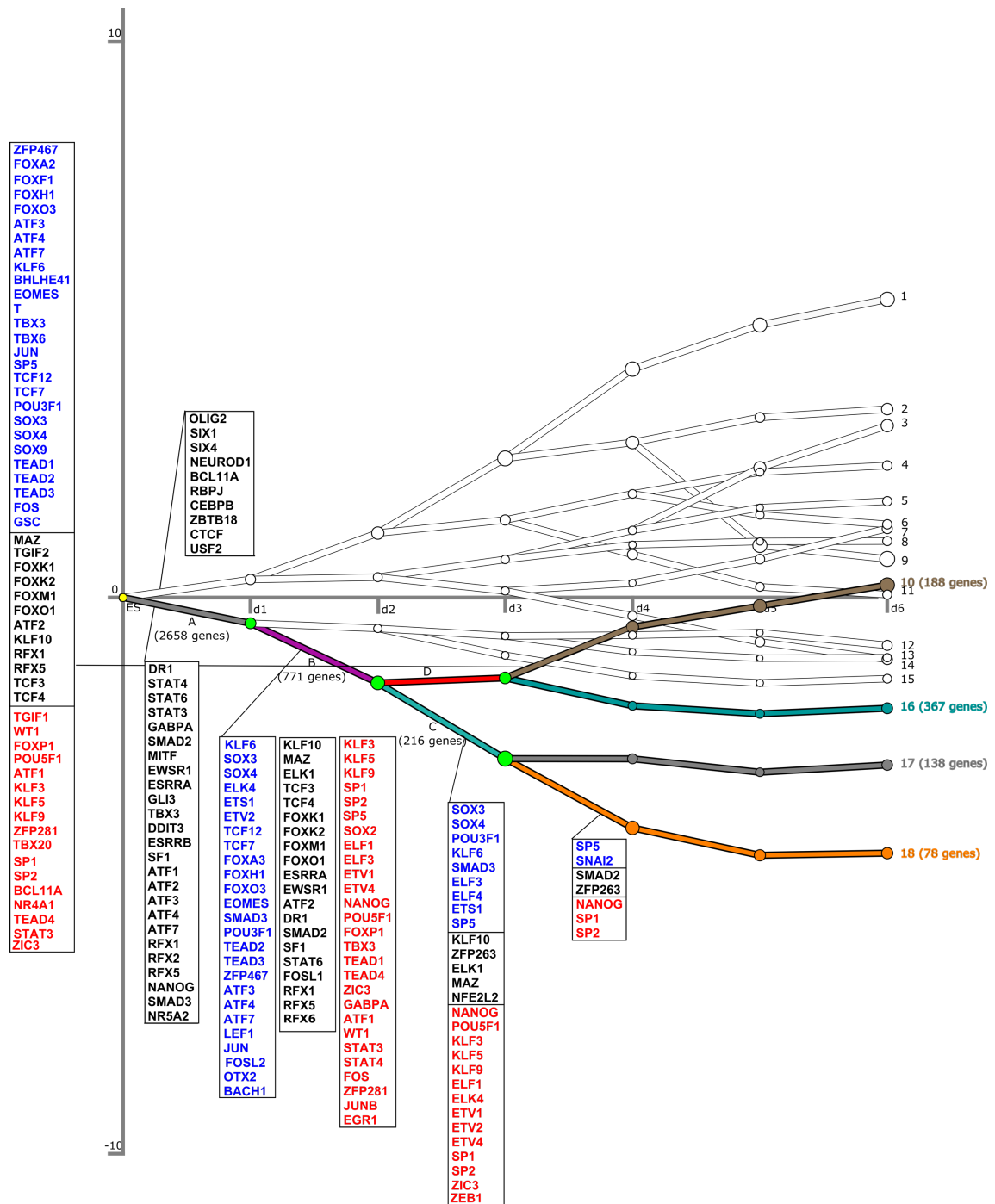
As the pluripotency genes in path 18 stem from the 2658 genes of the path A, we next looked at the fate of these genes at the stage d1 to d2. 1887 genes remained unchanged and 771 genes, including the future path 18, became further down-regulated and were segregated into a new branch, path B (Figure 4.20). The TFs responsible for the down-regulation of this branch include Foxh1, Tcfs, Lef1, Nanog, Pou5f1, Otx2 and Klf6. In addition, the GO term “stem cell population maintenance” is very significantly enriched in this path (Supplementary Figure 4 B).

From d2 to d3, the 771 genes in path B were separated in two groups. 216 genes, containing the future path 18, were further down-regulated and formed path C, which was controlled by the TFs including Nanog, Pou5f1, Klf6, Smad3 and Zeb1. The corresponding GO terms enriched in this path contain “stem cell population maintenance”, “stem cell differentiation” and “endodermal cell fate specification” (Supplementary Figure 4 C). The remaining 555 genes in the adjacent path D remained unchanged and there were no enriched TFs assigned to this path. The GO terms enriched in path D include “neural tube closure” and “negative regulation of apoptotic process”. Notably, at d3 to d4, path 10 segregates from path D. 188 genes of path 10 are enriched in GO terms “cell adhesion” and “negative regulation of cell division” and are regulated by mesodermal factor TFs Eomes and T (Figure 4.20).

From d3 to d4, the path 18 is finally segregated from path C. It is separated from 138 genes of path 17, that contain *Nodal*, *FoxD3*, *Jarid2* and other genes associated with GO terms "embryonic placenta development" and "negative regulation of transcription". Our method predicts that the separation of path 18 from path 17 is controlled by TFs such as Snai2, Smad2, Zfp263, Nanog, and Sp1/2/5.

Overall, our results show that Nanog and Smads are required for regulation of path 18 genes throughout the differentiation time course, as they are assigned to paths A, B, C and 18. In contrast, the mediators of Wnt signaling (Tcfs and Lef1) are acting transiently between d1 and d2.





**Figure 4.20 Bifurcation events in the dynamic regulatory network controlling stem cell population maintenance genes**

For the assigned TFs, the colors blue, black and red indicate “up-regulation”, “not changing” and “down-regulation” of current gene expression levels compared to ES separately.

## **Bifurcation events in the dynamic regulatory network controlling cardiovascular system development**

In contrast to path 18 with genes constantly down-regulated, the genes up-regulated during the time course, namely path 1, including genes *Hand1*, *Tgfb1*, *Actc1*, *Tnnt3*, *Tnnc1* and *Tnni1*, are highly related to “actin-mediated cell contraction”, “angiogenesis” and “cardiac muscle contraction” (Figure 4.21; Supplementary Table 6). The up-regulation of genes associated with cardiovascular system development is consistent with the differentiation process *in vivo* and is an important feature along the timeline of mesodermal development. Thus, we observed closely the bifurcation events leading to the formation of path 1 starting from ES cells (Figure 4.21).

As mentioned previously, the split at the time point ES is biased to the gene expression patterns at later time points. Genes in path 1 are originally grouped in the upward path from ES to d1 (path A, Figure 4.21), which include 4319 genes. These genes are associated with enriched GO terms “signal transduction”, “cell migration” and “cell adhesion” (Supplementary Figure 4 A). TFs such as *Olig2* and *Six1/4* were predicted to regulate their expression.

As the cardiovascular genes in path 1 stem from the 4319 genes of the path A, we next studied the fate of these genes at the stage d1 to d2. 2746 genes remained unchanged and 1573 genes, including the future path 1, became further up-regulated and were separated into a new branch, path B (Figure 4.21). Genes in path B are associated with enriched GO terms “positive regulation of cell migration”, “multicellular organism development” and “Wnt signaling pathway” (Supplementary Figure 4 B). TFs such as *Klfs*, *Sox3/4*, *Tcfs*, *Lef1*, *Eomes* and *Foxh1* were predicted to be responsible for the up-regulation of this path. The enriched TFs *Tcfs* and *Lef1* mediate Wnt signaling, which is required for mesoderm formation<sup>161</sup>. The combinatorial activities of *Foxh1* and *Eomes* are required for vertebrate mesendoderm specification via Nodal signaling pathway<sup>162</sup>.

From d2 to d3, the 1573 genes in path B were separated in two groups. 1135 genes kept unchanged and 438 genes, containing the future path 1, were further up-regulated and formed path C. This upward path C was associated with GO terms “heart morphogenesis”, “blood vessel remodeling” and “positive regulation of angiogenesis” (Supplementary Figure 4 C). Accordingly, the previously studied “heart morphogenesis” associated TFs, including

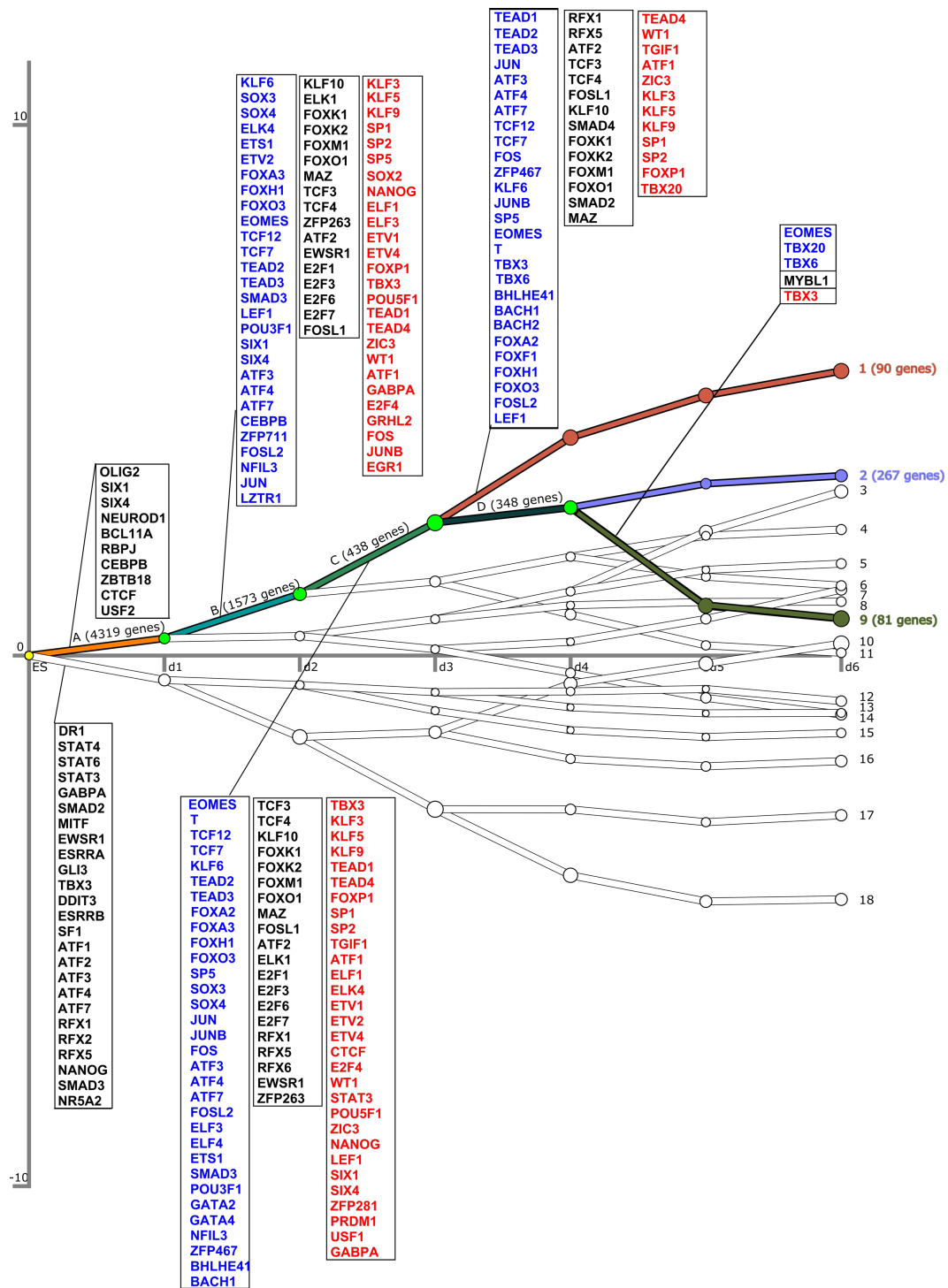
Tead1/2<sup>163</sup>, Foxh1<sup>164</sup>, Sox4<sup>165</sup>, Jun<sup>166</sup>, Smad3<sup>167</sup>, Gata4<sup>168</sup>, Atf2<sup>169</sup>, Tbx3<sup>170</sup>, Zic3<sup>171</sup> and Six1<sup>172</sup>, and “positive regulation of angiogenesis” associated TFs, including Stat3<sup>173</sup>, Ets1 and Gata2/4<sup>168</sup>, are predicted to regulate the genes in this path (Figure 4.21).

From d3 to d4, the 438 genes in path C were further separated in two groups and the path 1 was finally formed by segregating from path C. The 90 genes in path 1 are separated from 348 genes of path D, which remained unchanged and are associated with GO terms “multicellular organism development” and “heart morphogenesis”. There are no TFs assigned to path D, while our method predicts that the divergence of path 1 is controlled by TFs such as Teads, Atfs, Tcfs and T-box TFs (Eomes, T and Tbx3/6/20).

Notably, from d4 to d5, the 348 genes in path D were separated in two paths: path 2 and 9. In comparison to the continuously upward path 1, the expression levels of genes in path 2 and 9 do not show significant change from d3 to d4. At the time point d4, 81 genes of path 9 showing dramatical down-regulation are segregated from 267 genes of path 2 which are not changing significantly. The GO term analysis has shown that path 2 is related to “positive regulation of cell migration”, “outflow tract morphogenesis”, “positive regulation of smooth muscle cell proliferation” and “blood vessel development”, and path 9 is highly associated with “gastrulation” and “WNT signaling pathway”, which are required processes for the early differentiation and needed to be repressed at later time points (Supplementary Table 6). The T-box TFs Eomes, Tbx20, Tbx6, Tbx3 are predicted to regulate gene expression of path 9, even with a very stringent p value (Supplementary Figure 4 E). Since path 9 genes are down-regulated, these T-box factors are predicted to act as repressors. Many early mesodermal genes are co-expressed in this path, including *Eomes*, *Mesp1*, *Mixl1*, *T*, *Msgn1* and *Wnts*, which are only transiently active.

Similar to path 1, in the time period between d2 and d6, path 3 is steadily rising (Figure 4.21). The differences between path 3 and path 1 are that the expression levels of genes in path 3 do not shown significant change from d1 to d2 and that the degrees of their up-regulation from d2 to d6 are varying. Path 3 shares the main enriched GO terms with path 1, which are “angiogenesis” and cardiac muscle related terms with “cardiac muscle tissue development” for path 3 and “cardiac muscle contraction” for path 1. The TFs assigned to path 3 are mostly common with the TFs assigned to path 1, indicating that starting from d3, these TFs activate cardiac muscle tissue development genes in path 1 and 3.

Overall, the results show that T-box TFs (such as Eomes), Teads, Tcfs and Foxh1 are required for regulation of path 1 genes starting from the differentiation time point d1, as they are assigned to paths B, C and 1 (Figure 4.21). Some of the enriched TFs such as Tead1/2 and Foxh1 have been shown to be related to “heart morphogenesis” in previous studies, a validation of this method. Starting from d4, as genes (in path 1, 2 and 3) associated with cardiovascular system development are constantly up-regulated, genes (in path 9) associated with “gastrulation” and “WNT signaling pathway” are significantly down-regulated, since these genes are required for the early differentiation, but needed to be repressed at later time points. In general, the bifurcation events and associated TFs along paths 1, 2, 9 and 3 in the dynamic regulatory network well recapitulated the process of cardiovascular system development.



**Figure 4.21 Bifurcation events in the dynamic regulatory network controlling cardiovascular system development**

For the assigned TFs, the colors blue, black and red indicate “up-regulation”, “not changing” and “down-regulation” of current gene expression levels compared to ES separately.

### **Bifurcation events in the dynamic regulatory network controlling EMT process**

Mesoderm formation is dependent on EMT. Among the 18 paths at the final time point, path 4, which has 497 genes, is highly related to GO term “positive regulation of epithelial to mesenchymal transition”, with corresponding genes *Tgfb1i1*, *Crb2*, *Glipr2*, *Bmp2*, *Tgfb2* and *Smad3* grouped in this path (Figure 4.22; Supplementary Table 6). Thus, we observed the bifurcation events leading to the formation of path 4 closely.

Starting from the time point ES, the bifurcation events leading to path A and path B (Figure 4.22) are common between path 4 and path 1. At the time point d2, the 1573 genes in path B were separated in two groups, which are the 1135 genes in path C, containing the future path 4, and the 438 genes which included the final path 1. The genes in path C were generally unchanged, having no enriched TFs with the selected cutoff. The GO terms “signal transduction” and “negative regulation of canonical Wnt signaling pathway” are highly enriched in path C, with the corresponding genes *Bmp2*, *Notch1*, and *Wnt11*, *Wnt5a*, *Wnt9a* grouped in this path.

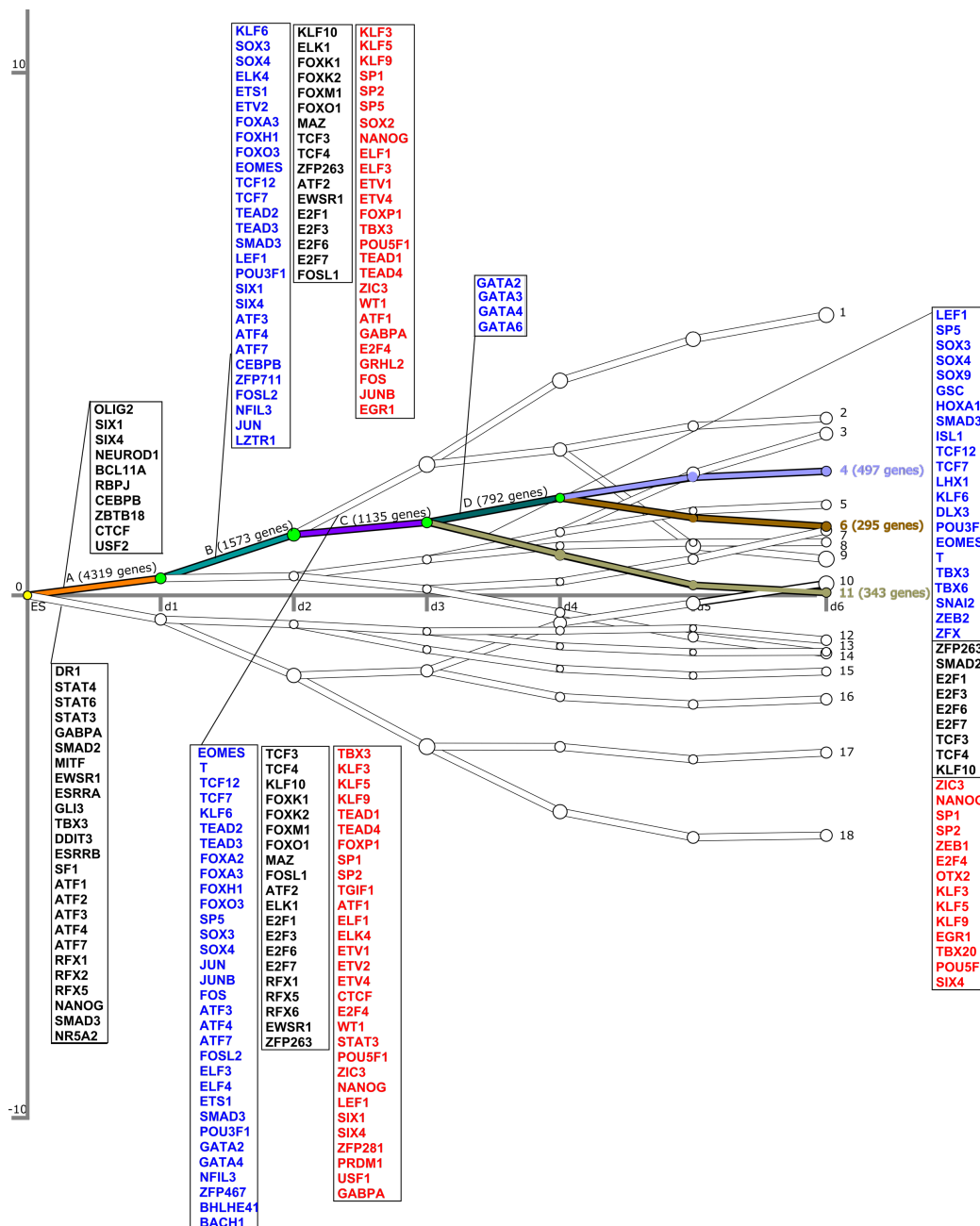
From d3 to d4, the 1135 genes in path C were further separated in two groups: the upward path D and the downward path 11. The 343 genes in path 11 were associated with “regulation of somitogenesis” and “Notch signaling pathway”. TFs such as *Lef1*, *Nanog*, *Eomes* and *T* were predicted to regulate the genes in this path. The 792 genes in path D are highly related to GO terms “positive regulation of epithelial to mesenchymal transition” and “negative regulation of canonical Wnt signaling pathway”. The GATA family was specifically assigned to path D and predicted to regulate this path.

From d4 to d5, the path 4 is finally segregated from path D. It is separated from 296 genes of path 6, which are associated with GO terms “fatty acid metabolic process” and “establishment of epithelial cell apical/basal polarity”, containing genes such as *Scd1*, *Foxf1*, *Wnt5a* and *Myo6* (Supplementary Table 6). There are no TFs assigned to path 4 and 6 with the selected cutoff. With a less stringent cutoff ( $X=1.3$ ), TFs *Smad3*, *Plagl1*, *Zfp711*, *Zeb1* were assigned to path 4, while *Ewsr1* was assigned to path 6 (data not shown). *Zeb1* is EMT-associated and probably up-regulate genes in path 4.

Overall, the GATA TFs were shown to be important along the timeline of forming path 4. From d3 to d4, they were uniquely assigned to the path containing the future path 4

(Supplementary Figure 4 D), which is path D (Figure 4.22) annotated with the GO term “positive regulation of epithelial to mesenchymal transition”. Meanwhile, path 11, separated from path D at d3, is down-regulated and associated with GO terms “regulation of somitogenesis” and “Notch signaling pathway”, indicating genes in path 11 are required for early differentiation stages, rather than the later stages.

In summary, through the bifurcation events in the gene regulatory network related to “stem cell population maintenance and differentiation”, “cardiovascular system development” and “EMT” (Figures 4.17 to 4.19), we demonstrated that the co-expressed genes were grouped into proper paths and the GO terms, together with the assigned TFs, were able to explain the bifurcation events properly. Since the TFs were predicted computationally, the validation was next performed by integrating our experimental data.



**Figure 4.22 Bifurcation events in the dynamic regulatory network controlling EMT process**

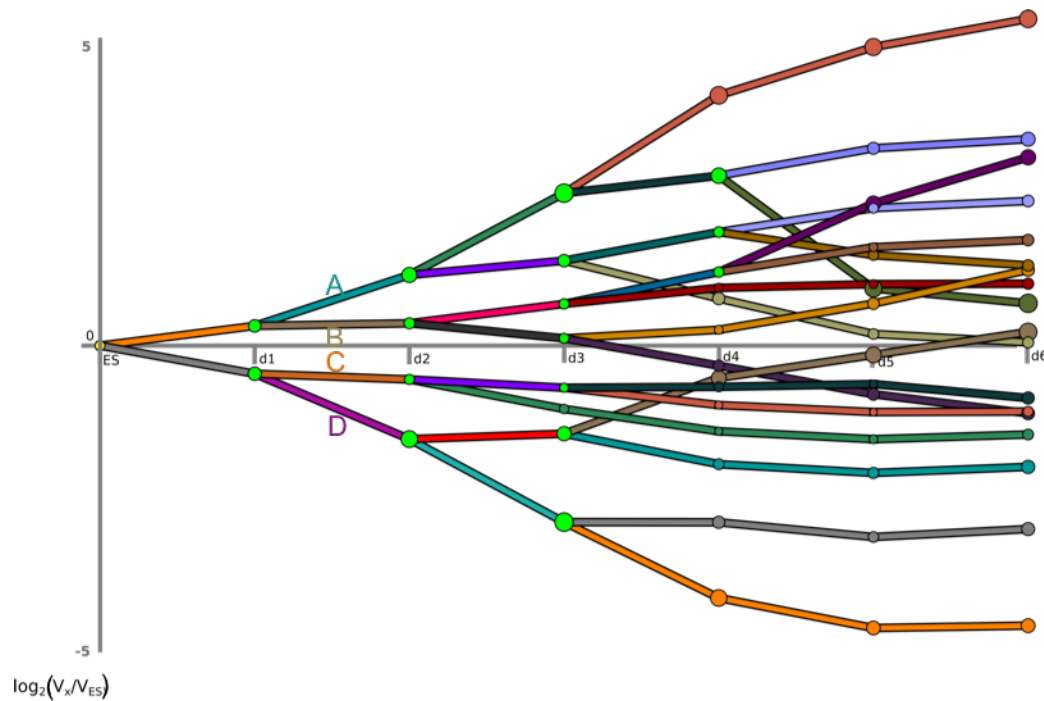
For the assigned TFs, the colors blue, black and red indicate “up-regulation”, “not changing” and “down-regulation” of current gene expression levels compared to ES separately.



### 4.3.3 Validation of the Gene Regulatory Network

The global gene regulatory network constructed in this study shows the bifurcation events (where a set of genes with a similar expression pattern diverge) in the time course of mesoderm formation. TFs that potentially regulate a certain set of genes grouped in a path were assigned to this path to explain the corresponding bifurcation events. For instance, in Figure 4.21, showing the bifurcation events controlling cardiovascular system development, Smad3 is assigned to path B, while Eomes and T are both assigned to path C and path 1. To assess whether the assigned TFs are biologically meaningful, the global network was validated by overlapping the target genes of Smads, Eomes and T with genes in specific paths of the network. The target genes used here included the DE genes from RNA-seq (WT/KO) and the direct targets from RNA-seq (WT/KO) combined with ChIP-seq binding sites.

The target genes of Smads were obtained using ChIP-seq and RNA-seq (WT/KO) assays performed at d2. Therefore, the assignment of Smads on the paths between d1 and d2 was observed. The paths from d1 to d2 were ordered alphabetically from top to bottom as from path “A” to “D” (Figure 4.23). In the global GRN, Smads were assigned to “d1-d2” paths A and D, rather than paths B and C (Supplementary Figure 4 B). By overlapping target genes of Smads with genes in these four paths, it shows that Smads target genes are specifically highly enriched in paths “A” and “D”, rather than paths “B” and “C” (Table 4-7), which is consistent with the global GRN.



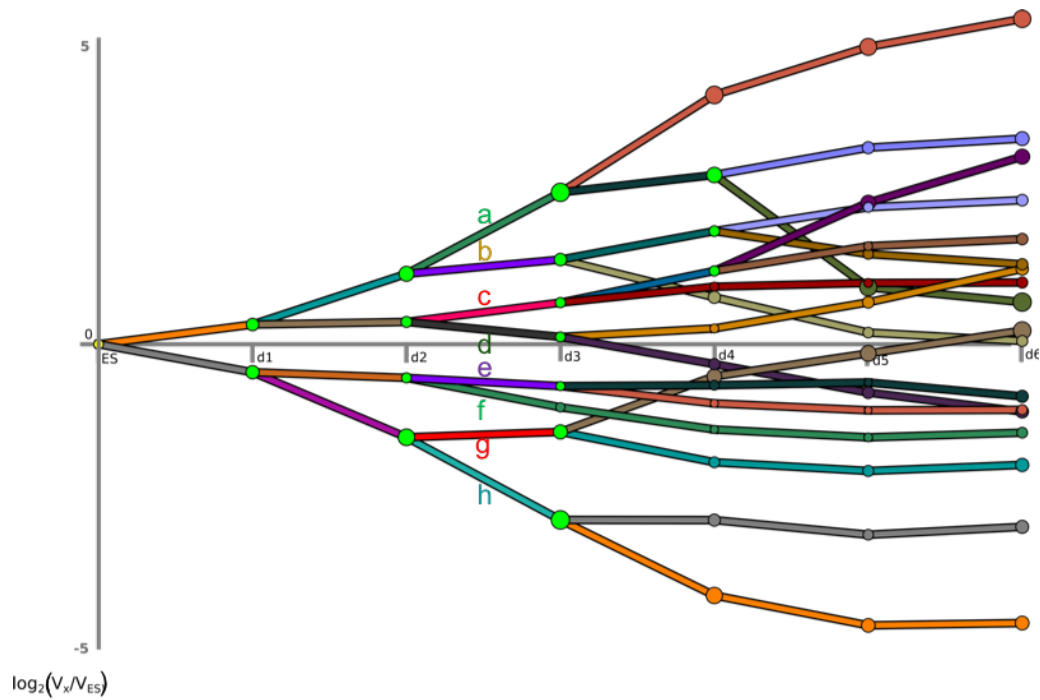
**Figure 4.23** Paths from “A” to “D” marked for the enrichment analysis showing in Table 4-7

**Table 4-7. Overlap Smad targets with “d1-d2” paths (Figure 4.23)**

“DE genes” means downstream genes from RNA-seq WT/KO comparison analysis. “direct targets” are from overlapping that with ChIP-seq results. For instance, “A (1573)” indicates that there are 1573 genes in path A. “249 (16% of the path)” indicates that 249 (or 16%) of the 1573 genes from path A are among the 1093 Smad4 DE genes. (Fisher’s exact test: \*  $p < 0.05$ , \*\*  $p < 0.01$ , \*\*\*  $p < 0.001$ , \*\*\*\*  $p < 0.0001$ ).

|               |                         | A<br>(1573)       | B<br>(2746) | C<br>(1887) | D<br>(771)        |
|---------------|-------------------------|-------------------|-------------|-------------|-------------------|
| Smad4<br>(d1) | 1093<br>(DE genes)      | 249 ****<br>(16%) | 240<br>(9%) | 80<br>(4%)  | 124 ****<br>(16%) |
|               | 288<br>(direct targets) | 89 ****<br>(6%)   | 58<br>(2%)  | 20<br>(1%)  | 35 ****<br>(5%)   |
| Smad4<br>(d2) | 1062<br>(DE genes)      | 302 ****<br>(19%) | 226<br>(8%) | 76<br>(4%)  | 182 ****<br>(24%) |
|               | 331<br>(direct targets) | 112 ****<br>(7%)  | 70<br>(3%)  | 18<br>(1%)  | 54 ****<br>(12%)  |

Our ChIP-seq and RNA-seq (WT/KO) assays of Eomes and T were performed at d2 and d3, we therefore observed the assignments of Eomes and T on the paths between d2 and d3. The paths from d2 to d3 were ordered alphabetically from top to bottom as from path “a” to “h” (Figure 4.24). From d2 to d3, the global GRN assigns Eomes and T to only path “a”, rather than any of the other paths (Supplementary Figure 4 C). By overlapping target genes of Eomes or T with genes in all 8 paths from d2 to d3, it shows that Eomes and T are specifically highly enriched in path “a”, rather than the other ones (Table 4-8), which is consistent with the global GRN. Moreover, the results above were compared with Eomes or T targets in our TF-target interaction dataset built as the input to DREM (Table 4-6), which generated a large number of common targets (Table 4-8). In conclusion, the assignment of TFs in the global GRN was shown to be valid according to the Smads, Eomes and T target data.



**Figure 4.24** Paths from “a” to “h” marked for the enrichment analysis showing in Table 4-8

**Table 4-8. Overlap Eomes and T targets with “d2-d3” paths (Figure 4.24)**

“DE genes” means downstream genes from RNA-seq WT/KO comparison analysis. “direct targets” are from overlapping that with ChIP-seq results. The numbers in black are the results of common genes. Those overlapped genes were then further overlapped with targets of T or Eomes from our TF-targets dataset in each path, which generated the results marked in blue in this table. For instance, “a (438)” indicates that there are 438 genes in path a. “137 (31% of the path)” indicates that 137 (or 31%) of the 438 genes from path a are among the 1149 T DE genes. “(74; 54% of 137)” indicates that 54% of the 137 genes are further overlapped with targets of T from our TF-targets dataset. (Fisher’s exact test: \* p < 0.05, \*\* p < 0.01, \*\*\* p < 0.001, \*\*\*\* p < 0.0001).

|       |                            | a<br>(438)                                       | b<br>(1135)              | c<br>(1949)              | d<br>(797)               | e<br>(1352)            | f<br>(535)               | g<br>(555)               | h<br>(216)              |
|-------|----------------------------|--|--------------------------|--------------------------|--------------------------|------------------------|--------------------------|--------------------------|-------------------------|
| T     | 1149<br>(DE<br>genes)      | 137 ****<br>(31% of the path)<br>(74;54% of 137) | 246<br>(22%)<br>(91;37%) | 181<br>(9%)<br>(46;25%)  | 83<br>(10%)<br>(25;30%)  | 60<br>(4%)<br>(8;13%)  | 27<br>(5%)<br>(1;4%)     | 92 (17%)<br>(18;20%)     | 41<br>(19%)<br>(15;37%) |
|       | 811<br>(direct<br>targets) | 112 ****<br>(26% of the path)<br>(66;59% of 112) | 189<br>(17%)<br>(77;41%) | 134<br>(7%)<br>(43;32%)  | 52<br>(7%)<br>(20;38%)   | 42<br>(3%)<br>(6;14%)  | 17<br>(3%)<br>(1;6%)     | 57 (10%)<br>(15;26%)     | 30<br>(14%)<br>(9;30%)  |
| Eomes | 1443<br>(DE<br>genes)      | 157 ****<br>(36% of the path)<br>(88;56% of 157) | 234<br>(21%)<br>(93;40%) | 213<br>(11%)<br>(56;26%) | 105<br>(13%)<br>(24;23%) | 72<br>(5%)<br>(13;18%) | 46 **<br>(9%)<br>(5;11%) | 125<br>(23%)<br>(32;26%) | 58<br>(27%)<br>(25;43%) |
|       | 622<br>(direct<br>targets) | 85 ****<br>(19% of the path)<br>(52;61% of 85)   | 128<br>(11%)<br>(65;51%) | 83<br>(4%)<br>(29;35%)   | 42<br>(5%)<br>(15;36%)   | 31<br>(2%)<br>(8;26%)  | 14<br>(3%)<br>(2;14%)    | 40<br>(7%)<br>(16;40%)   | 32 **<br>(15%;47%)      |

## 5 Discussion

Mesoderm formation happens early and is essential in embryogenesis. It involves complex mechanisms and tightly regulated gene expression. This study aimed to construct a global gene regulatory network (GRN) describing transcriptional regulatory events occurring dynamically during the course of mesoderm formation in the mouse. We employed an *in vitro* system using mouse ES cells differentiated to mesoderm in order to mimic the process of mesoderm formation *in vivo*. After the *in vitro* differentiation system was validated, it was used as the foundation for deciphering the GRNs mediated by the master mesodermal regulators Smads, Eomes and T. Furthermore, a global GRN was reconstructed, which reflects the whole gene regulatory process of mesoderm formation (Figure 4.13). This global network was validated by comparing its output with Smads, Eomes and T datasets, showing that these TFs are correctly assigned to their targets.

### **Analysis and Validation of Mesoderm Formation *in vitro***

To study developmental processes, it is essential to select a model system that allows to produce enough material for high-throughput experiments and to precisely score developmental stages. To fulfill these criteria, we chose to use *in vitro* mesodermal differentiation of mESCs. Applying this procedure to the mESCs resulted in formation of beating cardiomyocytes on differentiation day 8. To validate this *in vitro* system, time-series RNA-seq was performed for 10 stages of differentiation, including undifferentiated (ES), early hourly stages (1h, 6h, 12h) and later daily stages (d1 to d6). By monitoring the expression patterns of known gene markers of ES cells, PS, EMT, mesoderm precursor, and cardiac mesodermal cells, it was investigated whether the temporal aspect of gene expression mimics that of mesoderm formation *in vivo*. The results showed that the *in vitro* system recapitulated *in vivo* mesoderm formation process and roughly defined the developmental stages during the time course. These included pluripotency stage (ES to d1) with markers Pou5f1 and Sox2 highly expressed at very early stages and decreased at later stages; PS-like/nascent mesoderm stage (d2 to d3) with expression of markers Eomes, T, Fgf5, Wnt3/3a and Msgn1 peaking at d3; early cardiac mesoderm (d4) with high expression of markers Mesp1, Foxf1 and Kdr at d4; and committed cardiac cells (d5 to d6) with markers Tbx20, Hand2, Gata4, Gata5, Gata6 and Tnnt2 which were not expressed at the early stages while highly expressed at d5 or d6 (Figure 4.1). In addition, EMT started at around d3 with

the reduced expression of epithelial marker Cdh1 and increased expression of mesenchymal makers including Cdh2, Cdh11, Snai1, Snai2, Zeb1, Zeb2 and Prrx1, which is consistent with the differentiation process *in vivo*.

Checking the expression patterns of well-known markers for various developmental stages is a very straightforward way to validate the *in vitro* differentiation system. Furthermore, since co-expressed genes tend to form gene regulatory networks and share the same functions, the time-series data was observed closer by clustering the DE genes during the time course. This cluster analysis was performed to further validate our *in vitro* system by observing the gene expression patterns of each cluster, including marker genes of various differentiation stages, and to explore the elements and functions of each cluster. The hierarchical clustering was used to cluster the time-series DE genes, because it allows us to use Pearson's correlation to define the distance between genes and it does not require the prior knowledge of the number of clusters before clustering (i.e., one of the most common cluster method k-means algorithm)<sup>174</sup>.

The hierarchical clustering tree was divided into seven sub-clusters and observed that the markers of distinct differentiation stages were clustered in the same group, for instance, pluripotency markers Pou5f1 and Sox2, mesodermal markers Eomes, T, Mesp1 and Wnt3a, EMT markers Cdh2, Snai1, Snai2 and Zeb1, which further validated the *in vitro* system (Figure 4.1; Table 4-1).

Apart from observing the expression patterns of marker genes, the sub-clusters were utilized in order to functionally profile the co-expressed genes in each cluster. Since each cluster contains a lot of genes and it is not practical to explore each gene separately, GO term enrichment analysis was therefore employed. GO term analysis translates a gene list into a functional profile which offers insight of the underlying biological mechanisms associated with this gene list, including biological process, molecular function and cellular component. Although the current tools for GO term analysis are different in the aspects of data visualization, reference data source and analysis speed, their principle is to create a list of functional categories using annotation databases and to find the over-represented function categories for the input gene list<sup>175</sup>. Thus, the limitations are present in all tools based on this approach, including the tool DAVID<sup>123</sup> used in this study. Firstly, the annotation databases to generate function categories are incomplete and so previously unknown functions for

known genes cannot be discovered, which can lead to incomplete or biased enriched GO terms. Secondly, a number of annotations are often high-level GO terms which are too general to be practically useful. Thirdly, there are small chances that some inferences are incorrect. Fourthly, the gene expression levels are not considered during the GO enrichment analysis, while the various expression levels can be useful to assign different weights to the corresponding biological functions<sup>175</sup>. Despite the limitations, GO term analysis is still a useful and popular method.

Biologically meaningful GO terms were found for each cluster. Cluster 1 with genes highly expressed at early stages and started to decrease dramatically from 12h is related to “inner cell mass proliferation” and “stem cell population maintenance”. Cluster 2 with genes peaking at 1h and dropping sharply afterwards is related to “cell adhesion” and “stem cell population maintenance”. Cluster 3 with genes on average highly expressed at the start of differentiation and then dropped from d1 is related to “DNA methylation on cytosine”; “nucleosome assembly” and “protein heterotetramerization”, indicating reduced cell proliferation. Cluster 4 with genes peaking at d1 and d2 is related to “brain development”, indicating transient expression of neural genes, some of which are also important for mesoderm formation. Cluster 5 with transiently expressed genes peaking at d3 and d4 is related to “lung development” and “somitogenesis”, which are mesoderm-related terms. Cluster 6 with genes highly expressed starting at d3 and d4, reaching maximum at d5 and d6, is related to “heart development”. Cluster 7 with genes only highly expressed starting at d5 is related to “angiogenesis” and “sarcomere organization” (Table 4-1).

Since co-expressed genes are likely to be regulated by the same mechanism and are potentially involved in the same biological processes, the sub-cluster analysis provided candidates of co-regulators for the TFs in the same cluster. For example, Sall4 is in the same cluster as Pou5f1 and Sox2 and it has been shown to be required for ES cells pluripotency and early embryo development by forming an interconnected autoregulatory network with Pou5f1, Sox2 and Nanog in ES cells (Table 4-1). To find potential key TFs and co-regulators of known key TFs in each cluster, gene regulatory networks can be built for co-expressed genes in each cluster to infer TF-target gene interactions, by using methods such as Bayesian networks or motif analysis<sup>72,176</sup>. However, this analysis was not performed in this study, since the primary focus here was the validation of the *in vitro* system, rather than the details of gene regulation in each cluster.

In conclusion, the unbiased clustering analysis showed that the known marker genes for distinct differentiation stages were clustered together and that the GO terms were consistent with *in vivo* developmental procedure, demonstrating that the *in vitro* differentiation system recapitulated the *in vivo* process. Moreover, the co-expressed genes in each cluster can be used to explore the co-regulators of known TFs.

### **Gene Regulation by TFs Smads, Eomes and T during Mesoderm Formation**

The validation of the *in vitro* differentiation system suggested that it could be used as the foundation to study the regulators underlying EMT and mesoderm formation. ChIP-seq and RNA-seq (WT/KO) for master regulators of mesoderm formation such as Smads, Eomes and T were carried out utilizing this system. While the ChIP-seq data uncovers the DNA binding sites of a TF, RNA-seq (WT/KO) data shows the DE genes (up- or down-regulated upon TF removal). For a specific TF, by combining the results of ChIP-seq and RNA-seq (WT/KO), the regulated genes with TF binding sites, i.e. direct target genes, can be discovered. With this analysis, I aimed to detect novel target genes of Smads, Eomes and T to extend our knowledge of the GRNs mediated by them. Moreover, the results were used for the validation of our reconstructed global GRN (section 4.3.2).

It has been shown that DNA binding regions of Smad1, 5 and 8 display a significant overlap, so Smad1 binding sites represent those of Smad1/5/8<sup>24</sup>. To identify direct target genes of Smad1/5/8 and Smad2/3 respectively, ChIP-seq assays on Smad1 and Smad2/3 were performed to detect their binding sites. RNA-seq (WT/KO) assay of Smad4 was performed to find DE genes. The reason for using Smad4 instead of Smad1/5/8 and Smad2/3 WT/KO cells for RNA-seq assay is that the phosphorylated Smad1/5/8 or Smad2/3 need to form a complex with Smad4 to function. The Smad4 knockout assay alone is sufficient to investigate the respective target genes of Smad1/5/8 and Smad2/3 by overlapping their respect ChIP-seq associated genes with Smad4 WT/KO DE genes<sup>26</sup>. To select a time point for high-throughput assays, the levels of phosphorylated Smad1/5/8 and Smad2/3 were detected using western blot, which showed that they reached the maximum at d1 and d2 (Figure 4.3). Thus, d2 samples were chosen to perform assays for Smads.

The ChIP-seq peaks of Smad1 and Smad2/3 preferentially bind to enhancer-associated regions (genic and intergenic) compared to promoters. My analysis showed that the binding



motif of Smad1 is a Pou5f1/Sox2 binding site<sup>34</sup>, while Smad2/3 motif is the same as Smad3 motif identified by Badis *et al.*<sup>149</sup>, which is a validation of our experiments and an indication of functional binding sites. 1062 DE genes were identified by Smad4 RNA-seq (WT/KO). The up-regulated genes by Smad4, including *Id1*, *Id2*, *Id3*, *Tdgf1*, Wnt (*Wnt3*, *Wnt4*, *Wnt5b*, *Wnt6*, *Wnt7b*, *Wnt8a*), FGF (*Fgf8*, *Fgf17*), *Notch3*, *Nodal*, *Nanog*, *Axin2*, *Mixl1*, *Eomes* and *T*, were annotated with GO terms “BMP/WNT signaling pathway”, while down-regulated genes by Smad4, including *Pax6*, were highly related to “nervous system development”.

Many factors of signaling pathways such as WNT, FGF, Nodal and NOTCH that are regulated by Smads were detected. In particular, genes *Wnt3*, *Wnt8a*, *Fgf8*, *Nodal* and *Notch3* are direct targets of both Smad1 and Smad2/3. Besides these signaling molecules, genes *Id2*, *Tdgf1*, *Nanog*, *Axin2*, *Mixl1*, *Eomes* and *T* are directly regulated by Smad1 and Smad2/3 as well, of which *Eomes* and *T* have been shown to be targets of Smad2/3 in previous studies<sup>152,153</sup>.

The same approach was used to identify *Eomes* direct targets. In the time course of mESCs differentiated to mesoderm, *Eomes* was highly expressed on the protein level on d2. Therefore, d2 samples were chosen to perform assays for *Eomes*. 10070 *Eomes* peaks were obtained from the ChIP-seq analysis, which were associated with 7346 genes. Compared to Smads and *T* (Figure 4.5, Figure 4.8, Figure 4.11), a larger proportion of *Eomes* binding sites was found in promoter regions rather than in genic or intergenic regions. The *de novo* motif analysis using all *Eomes* peaks detected *Eomes* motif as the most significant<sup>149</sup>, an indication of functional binding sites. 1443 DE genes discovered by RNA-seq (WT/KO) analysis were overlapped with ChIP-seq associated genes, resulting in 622 direct target genes of *Eomes*. The 371 target genes up-regulated by *Eomes* included *T* and *Fgf5*, while the 251 target genes down-regulated by *Eomes* included *Sox2*, *Lef1*, *Id1/3*, *Cdx1/2*, *Nkx1-2* and *Stat4*.

To identify direct target genes of *T*, d3 samples were employed for ChIP-seq and RNA-seq (WT/KO) because of the highest expression of *T* at this time point. 23714 peaks were associated with a total of 13089 genes. The *de novo* motif analysis of *T* peaks revealed both binding motifs of *T* on top of the list: the palindromic motif (depicted in the motif database as *T\_full* motif)<sup>156</sup>, and *T*-box motif, which is the DNA consensus sequence that can be bound by all members of *T*-box family. Overlapping ChIP-seq associated genes with 1149 DE genes from *T* RNA-seq (WT/KO) assays, 811 direct target genes of *T* were obtained

(Figure 4.12). The 536 target genes up-regulated by T included mesodermal marker genes *Fgf8*, *Eomes*, *Mesp1* and *Lef1*, while the 275 down-regulated target genes included *Fos*, *Id3*, *Igf2*, *Ascl2*, *Acer2*, *Heg1*, *Gata3*, *Fgf4* and *Sox2*.

In order to locate the target genes of Smads, Eomes and T during the differentiation time course, I checked the distribution of the direct target genes of Smad1 (475 genes), Smad2/3 (303 genes), Eomes (622 genes) and T (811 genes) in the seven sub-clusters of the transcriptome analysis (Table 4-1; Figure 4.2). It showed that Smad1, Smad2/3 and Eomes targets were significantly enriched in clusters 4 and 5 and that T targets were highly enriched in cluster 5 (Table 4-2; Table 4-3; Table 4-4). As SMAD signaling is the earliest response to differentiation cues, it makes sense that Smad targets get enriched in very early up-regulated genes in cluster 4 and 5. Cluster 5 contains many mesoderm-associated genes, which, apart from *Eomes* and *T*, include *Mesp1* and *Wnt3a*. The enrichment of direct target genes of Eomes and T in cluster 5 points at their roles as the master regulators for mesoderm formation.

The T-box TFs Eomes and T were shown to bind to the same genome regions during gastrulation in *Xenopus*<sup>60</sup>. In addition, the genomic binding sites of Eomes and T were shown to be very close in differentiating human ES cells<sup>35</sup>. In our study, 44% (4480 out of 10070) Eomes ChIP-seq peak summits are within 500 bp distance from T summits (Supplementary Figure 2 A) and Eomes and T share many downstream genes, such as *Cdx1/2*, *Lef1*, *Sall2*, *Zeb2*, *Stat1*, *Mixl1* and *Fgf8* (Supplementary Figure 2 B). Moreover, our results (section 4.2.2 and 4.2.3) show that the direct targets of Eomes and T are both enriched in cluster 5 of the time-series RNA-seq data (Table 4-3; Table 4-4). To study the combinatorial function of these TFs, we carried out RNA-seq assays for Eomes KO, T KO and Eomes/T double KO at d3. Different categories with various gene expression patterns were generated using k-means clustering of DE genes (Supplementary Figure 3)<sup>158,159</sup>, which showed that a great number of DE genes were regulated by the combination of Eomes and T (such as genes in cluster 1 to 4), while others depended on only one of these TFs (such as genes in cluster 5 to 7) (Supplementary Figure 3). Our preliminary basic analysis of Eomes and T peaks associated with genes in the k-means clusters did not show any significantly enriched peak patterns for any of the clusters (data not shown). More tests should be conducted in the future to characterize the mechanisms of combinatorial Eomes and T interactions. Notably, it can be insightful to perform cross comparison analysis for Smads, Eomes and T (i.e.,

common targets) to interpret how they cooperate and influence each other, which we did not carry out since this is not the focus of this study.

ChIP-seq binding peaks were associated with genes to identify the genomic binding locations and targeted genes of Smads, Eomes and T. The ChIP-seq peaks were assigned to genomic regions, such as promoter, genic and intergenic regions, as to putative target genes (RefSeq annotated genes). The peaks within -5kb/+2kb of the TSS regions were assigned to be promoter associated. The peaks within +2kb from the TSS to +5kb after the TES regions were assigned to be genic associated. The remaining peaks were defined to be intergenic. Genic and intergenic peaks are likely enhancer-related. When a ChIP-seq peak is intergenic, it is hard to define its real target genes. In this study, since the putative ChIP-seq target genes would be filtered by overlapping with DE genes from RNA-seq (WT/KO) assays, the intergenic peaks were associated with both of the closest up- and down-stream genes.

Notably, ChIP-seq assays output more TF binding sites than the direct target genes identified by overlapping TF ChIP-seq target genes with DE genes from RNA-seq WT/KO assays, suggesting that not all observed TF bindings are functional. One possibility is that the binding is not acting on transcriptional regulation, but on other processes such as chromosome structure regulation. It is also possible that TFs happen to bind to randomly occurring target sequences which are not selected against, because those binding events do not significantly affect gene expression<sup>177</sup>. In this study, most of the nonfunctional ChIP-seq binding sites were filtered by comparing with DE genes from RNA-seq (WT/KO), while keeping biologically meaningful peaks associated with transcription regulation.

### **The Global Dynamic GRN Orchestrating EMT and Mesoderm Formation**

To construct GRNs from time-series data, most of the available bioinformatical methods consider only the gene expression patterns in the time course and infer the static gene-target relationships either based on correlations between genes, such as relevance networks, or dependency between genes, such as Bayesian networks<sup>178</sup>. The method DREM utilized in this study offers the option to combine the time-series gene expression data with the dynamic regulatory data (TF-target gene interactions) to construct a global gene regulatory network in a tree structure (Figure 4.13), where the TFs responsible for the bifurcation events are assigned to the paths of the tree.

In this study, the global dynamic regulatory network underlying mesoderm formation was built by combining the time-series RNA-seq transcriptome dataset with the time-series ATAC-seq dataset. In general, this approach consists of three major steps. Firstly, the time-series RNA-seq transcriptome data was used to train the parameters associated with the tree structure and group the co-expressed genes into paths based on hidden Markov model (Figure 4.13 A). Secondly, the ATAC-seq data was used to generate a TF-target relationship table for each time point (Figure 4.13 B). Finally, the predicted TFs were assigned to the paths based on the enrichment analysis of TF targets among the genes in the paths (Figure 4.13 C). This global network allows us to predict TFs responsible for regulation of a subset of genes at every point of differentiation.

The first and third steps of this approach were performed with DREM<sup>102,160</sup>. The authors of DREM highly recommend to integrate the TF-target interactions while performing the first step to train the parameters associated with the tree structure. We did not do this because then the tree structure is highly biased to the TF-target interactions. Concerning DREM, which offers putative mouse TF-target interactions to train the parameters, it is notable that the putative TF-target data can be very noisy, especially when it is prediction-based. A better approach is to use the experimentally validated TF-target interaction data. However, this data is only available for limited number of TFs.

In this approach, only the time-series RNA-seq transcriptome data was used to train the parameters associated with the tree structure in the first step. For the second step, the time-series ATAC-seq data was used to calculate the TF-target gene interactions for each time point. The TF-target interactions were used as the input for the third step to carry out the TF enrichment analysis for genes in each path of the tree.

The first advantage of this approach is that ATAC-seq analysis was used to precisely locate the open chromatin regions potentially bound by TFs. The chromatin regions that open or close over time can be correlated to binding of certain TFs. By connecting differentially open chromatin regions, the genes associated with those regions and the potential binding TFs of those regions, the TF-target interactions were built to calculate the enriched TFs for each path of the global gene regulatory network. The second advantage of this approach is that it can be used to predict the potential binding genes for TFs with known motifs in one step, instead of performing ChIP experiments. Meanwhile, there are disadvantages of this

approach. Firstly, it is not applicable to TFs with no known motifs. In addition, TFs with similar motifs are hard to distinguish.

Time-series ATAC-seq data were used to build TF-target interactions. ATAC-seq uses a mutated hyperactive Tn5 transposase which can cut exposed DNA regions and simultaneously adapter-ligate those regions which then get amplified by PCR for NGS. The chromatin accessible open regions are potential regions where TFs can bind and exert their functions to regulate gene expression. However, it is challenging to identify the precise TF binding sites from ATAC-seq sequencing data, because the fragments might not be nucleosome-free and the Tn5 transposase insertion can cover part of the TF binding sites. Since TFs generally cannot bind to nucleosome-occupied regions, I firstly filtered out the mapped ATAC-seq fragments longer than 120 bp, because they represent DNA regions occupied by nucleosomes according to the size distribution of all mapped ATAC-seq fragments (Supplementary Figure 1). Then, since the kept nucleosome-free regions can be broad and have no binding TFs, I focused on identifying the “dips” of ATAC-seq signaling profile, which represent genomic regions protected from Tn5 binding and transposition, and since these regions are nucleosome-free, they likely represent binding sites of TFs. To detect Tn5 transposase insertion sites with a better resolution, I modified the reads by only keeping the first 10 bp from the 5' to 3' direction of each read (after shifting 3 bp right for the positive strand and 1 bp left for the negative strand, according to the binding architecture of Tn5 transposase<sup>87</sup>). Peak calling was performed after combining the modified reads of both replicates. Then only the peak pairs with a distance of less than 150 bp between two peak summits were kept and the insert regions between those peak pairs were defined as dips (section 3.11). The reason 150 bp was selected as a cutoff to define the dips was that most of the DNA regions potentially occupied by TFs are shorter than 150 bp, mostly around 60 bp, according to the global distribution of distances between peak summits (Supplementary Figure 5). Then, the “differential dips” were defined by overlapping the dips identified from two samples with their respective “differential regions”, which were identified in parallel. The advantage of integrating “differential regions” to the determination of “differential dips” is that the “differential regions” were calculated using elongated reads which are relatively broader regions and contain more signals. The TF-target interactions were built on “differential dips”.

In order to identify TF binding sites from ATAC-seq data, the published methods generally fall into two categories: computational footprinting<sup>179</sup> and peak calling followed by motif discovery. Instead of using computational footprinting, which usually requires a very high read coverage, peak calling followed by motif discovery was used after this method was improved. The difference and advantage of the improved method lies in the utilization of “dips” (section 3.11), which best suits this study. Another relatively different strategy to identify TF binding sites I have tried was using Homer with the parameter *-nfr*. This resulted in genomic regions that contain both high signal and a gap in between. The resulting regions were generally too broad and did not pass my further validation, so this method was not employed in the study.

Since the identified dips are the potential TF binding sites, to assess the validity of them, I utilized our Eomes and T ChIP-seq data. A dip contains ATAC-seq signals at the sides and a gap in between, so it was assumed that the TF binding motifs generally locate at the center of dips. The validation here was divided into two steps. Firstly, the precise locations of Eomes and T motifs, which were top motifs from the *de novo* motif analysis of Eomes and T ChIP-seq assays, were obtained. Then, the profile plot was generated to show the distributions of merged d1, d2, d3 ATAC-seq signals around Eomes motifs and d2, d3, d4 ATAC-seq signals around T motifs. The result demonstrated that the motifs were on average located within the dips and indicated that the dips were well defined (Figure 4.18).

In this study, only the differential open chromatin regions from ATAC-seq was used. In principle, all open regions, combined with either motif analysis or ChIP-seq data analysis, are useful to predict which TFs are involved in transcriptional regulation or which TFs are regulating the genes of interest. In this study, considering only the differential open regions helped us to narrow down the candidates of TFs which play more important roles in the differentiation process.

The parameters to enforce the tree structure of the global GRN were trained by only the time-series transcriptome data. The final tree structure includes 18 branches at the final time point, which was supported by appropriate groups of co-expressed genes and biologically meaningful GO terms enriched for most of the paths (Supplementary Table 6). For example, the genes in path 9 are highly expressed at d3/4 and annotated with GO terms gastrulation (*Eomes*, *Mesp1* and *Mixl1*), somitogenesis (*T* and *Msox1*) and mesodermal cell migration (*Fgf8*

and *Mesp1*). Compared to the seven sub-clusters from our transcriptome analysis (Table 4-1), this path fits the pattern of sub-cluster 5 and all the genes above are in this sub-cluster. Notably, while determining the tree structure, there is a tradeoff between the number of paths (variety of gene expression patterns) and the number of genes in each path. As shown in Figure 4.19 B, illustrating the trajectories of genes comprising path 3, there are some outliers which do not follow the pattern of this path. They can be assigned to a more appropriate path if there are more paths allowed during data training, but this will cause a problem that some paths contain too few genes to carry out further analysis.

After the tree structure of the global GRN was determined, enrichment analysis was used to assign the TFs to the paths out of the splits to interpret how genes of each path were regulated by specific TFs. We are mainly interested in the process of mesoderm formation and EMT. Thus, according to the GO terms enriched in the paths at the final time point, I followed the timeline of differentiation and analyzed the bifurcation events related to the processes of stem cell maintenance (Figure 4.20), cardiovascular system development (Figure 4.21) and EMT (Figure 4.22).

At the final time point, path 18 with genes continuously repressed during the time course is highly related to the GO term “stem cell population maintenance” (Figure 4.20). As a control group, the bifurcation events resulting in the formation of path 18 were firstly studied. My results show that Nanog and Smads are assigned to all paths leading to the future path 18, indicating they are required for regulation of path 18 throughout the differentiation time course. In addition, the TFs Klf8 are shown to be very important for regulation of path 18 from d1 to d3 (Figure 4.20).

In contrast to path 18, the continuously rising path 1 is related to cardiovascular system development (Figure 4.21). The results show that T-box TFs (such as *Eomes*), *Teads*, *Tcfs* and *Foxh1* are required for regulation of path 1 during the differentiation time course starting from d1. Starting from d4, genes (in path 9) associated with “gastrulation” and “WNT signaling pathway” are significantly down-regulated, since they are required for the early differentiation, but need to be repressed at later time points (Figure 4.21). Genes in path 9 contained many early mesodermal genes, such as *Eomes*, *Mesp1*, *Mixl1*, *T*, *Msox1* and *Wnts*. T-box TFs *Eomes* and *Tbx3* are predicted to regulate gene expression of this path, even with a very stringent p value. With a less stringent p value, the mediators of Wnt signaling (*Tcfs*

and Lef1) and EMT-associated TF Snai2 are also enriched in this path (Supplementary Figure 4 E).

The bifurcation events that resulted in the formation of path 4 (Figure 4.22) were then observed, since it is related to the GO term “positive regulation of epithelial to mesenchymal transition”. The main finding is that the GATA TFs are predicted to be important for regulation of path 4, because, among all paths from d3 to d4, they are uniquely assigned to the path containing the future path 4 (Supplementary Figure 4 D).

Among the enriched TFs assigned to the global network, many of them have been validated in previous studies. For instance, among the TFs assigned to path C (in Figure 4.21), which is associated with the GO term “heart morphogenesis” and contains the future path 1, the involvement of TFs such as Tead1/2, Gata4, Foxh1, Sox4, Jun, Smad3, Atf2, Tbx3, Zic3 and Tbx3 in the process of heart development was reported in early studies<sup>163–172</sup>, while the relations of some other TFs, such as Etv1<sup>180,181</sup> and TCF family TF Tcf7l1<sup>182</sup>, to heart development were established only in recent studies. These promising results indicate that the enriched TFs which have not yet been experimentally validated are good candidates to test in future studies. For most of the splits of the global network, a relatively stringent p value cutoff was used to identify the enriched TFs, less stringent cutoffs can be selected if more TF candidates are required. Apart from comparing the enriched TFs with published studies, I assessed the validity of the global network by overlapping the target genes of Smads, Eomes and T with genes in specific paths, d1 to d2 paths for Smads and d2 to d3 paths for Eomes and T, of the network. The results show that their target genes are enriched in the paths where they were assigned, rather than the other paths where they were not assigned (Table 4-7; Table 4-8). Overall, it was demonstrated that, in the global network, the co-expressed genes were grouped into biologically meaningful paths and that the enriched GO terms, combined with the enriched TFs, can well explain the bifurcation events. Furthermore, this global network was validated using the target genes of Smads, Eomes and T. The findings in general indicate that the global dynamic regulatory network recapitulates the regulatory process of mesoderm formation.

In this study, I have developed a bioinformatical approach that combines time-series RNA-seq transcriptome data and time-series ATAC-seq data to investigate gene regulation during mesoderm formation and EMT in the mouse. The *in vitro* system of the mESCs differentiated



to mesoderm was used. The computational and experimental validation of this system allowed me to use it as the foundation to identify target genes of crucial mesodermal TFs, such as Smads, Eomes and T, and to reconstruct a global gene regulatory network in a tree structure. This global network shows the bifurcation events (where a set of genes with a similar expression pattern diverge) in the time course of mesoderm formation. With enrichment analysis, TFs potentially controlling the bifurcation events can be assigned to the network, which requires the construction of a TF-target genes database in advance. I propose an original approach to build this database using the time-series ATAC-seq data, which can identify TF binding sites with greater accuracy than other methods. The assigned TFs shown in published studies and the further validation using our Smads, Eomes and T data support the predictive power of our global regulatory network. In this study, we introduced a bioinformatical approach of utilizing time-series transcriptome data combined with time-series ATAC-seq data to construct a global gene regulatory network. This method can be applied to future studies designed to characterize molecular mechanisms underlying specific developmental processes.

## 6 Bibliography

1. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of human transcription factors: Function, expression and evolution. *Nature Reviews Genetics* (2009). doi:10.1038/nrg2538
2. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).
3. Johnson, P. F. & McKnight, S. L. Eukaryotic transcriptional regulatory proteins. *Annu. Rev. Biochem.* **58**, 799–839 (1989).
4. Reiter, F., Wienerroither, S. & Stark, A. Combinatorial function of transcription factors and cofactors. *Current Opinion in Genetics and Development* (2017). doi:10.1016/j.gde.2016.12.007
5. Boeva, V. Analysis of genomic sequence motifs for deciphering transcription factor binding and transcriptional regulation in Eukaryotic cells. *Frontiers in Genetics* (2016). doi:10.3389/fgene.2016.00024
6. Reményi, A., Schöler, H. R. & Wilmanns, M. Combinatorial control of gene expression. *Nat. Struct. Mol. Biol.* **11**, 812–815 (2004).
7. Lee, T. I. & Young, R. A. Transcriptional regulation and its misregulation in disease. *Cell* **152**, 1237–1251 (2013).
8. Arya, G., Maitra, A. & Grigoryev, S. A. A structural perspective on the where, how, why, and what of nucleosome positioning. *J. Biomol. Struct. Dyn.* (2010). doi:10.1080/07391102.2010.10508585
9. Venkatesh, S. & Workman, J. L. Histone exchange, chromatin structure and the regulation of transcription. *Nature Reviews Molecular Cell Biology* (2015). doi:10.1038/nrm3941
10. Zaret, K. S. & Carroll, J. S. Pioneer transcription factors: Establishing competence for gene expression. *Genes and Development* (2011). doi:10.1101/gad.176826.111
11. Boyle, A. P. *et al.* High-Resolution Mapping and Characterization of Open Chromatin across the Genome. *Cell* **132**, 311–322 (2008).
12. Giresi, P. G., Kim, J., McDaniell, R. M., Iyer, V. R. & Lieb, J. D. FAIRE (Formaldehyde-Assisted Isolation of Regulatory Elements) isolates active regulatory elements from human chromatin. *Genome Res.* **17**, 877–885 (2007).
13. Buenrostro, J. D., Giresi, P. G., Zaba, L. C., Chang, H. Y. & Greenleaf, W. J. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat. Methods* **10**, 1213–1218 (2013).
14. Pajoro, A., Muiño, J. M., Angenent, G. C. & Kaufmann, K. Profiling nucleosome occupancy by MNase-seq: Experimental protocol and computational analysis. *Methods Mol. Biol.* **1675**, 167–181 (2018).
15. Tsompana, M. & Buck, M. J. Chromatin accessibility: A window into the genome. *Epigenetics and Chromatin* (2014). doi:10.1186/1756-8935-7-33
16. Nahaboo, W. & Migeotte, I. Cleavage and Gastrulation in the Mouse Embryo. *eLS* (2018). doi:10.1002/9780470015902.a0001068.pub3
17. Arnold, S. J. & Robertson, E. J. Making a commitment: Cell lineage allocation and axis patterning in the early mouse embryo. *Nat. Rev. Mol. Cell Biol.* **10**, 91–103 (2009).
18. Lamouille, S., Xu, J. & Derynck, R. Molecular mechanisms of epithelial-mesenchymal transition. *Nat. Rev. Mol. Cell Biol.* **15**, 178–196 (2014).
19. Arnold, S. J., Hofmann, U. K., Bikoff, E. K. & Robertson, E. J. Pivotal roles for eomesodermin during axis formation, epithelium-to-mesenchyme transition and

- endoderm specification in the mouse. *Development* **135**, 501–511 (2008).
20. Kitajima, S., Takagi, A., Inoue, T. & Saga, Y. MesP1 and MesP2 are essential for the development of cardiac mesoderm. *Development* **127**, 3215–3226 (2000).
  21. Dunn, N. R. Combinatorial activities of Smad2 and Smad3 regulate mesoderm formation and patterning in the mouse embryo. *Development* **131**, 1717–1728 (2004).
  22. Morikawa, M., Koinuma, D., Miyazono, K. & Heldin, C. H. Genome-wide mechanisms of Smad binding. *Oncogene* **32**, 1609–1615 (2013).
  23. Hill, C. S. Transcriptional control by the SMADs. *Cold Spring Harb. Perspect. Biol.* **8**, (2016).
  24. Malkoski, S. P. & Wang, X. J. Two sides of the story? Smad4 loss in pancreatic cancer versus head-and-neck cancer. *FEBS Lett.* **586**, 1984–1992 (2012).
  25. Rahman, M. S., Akhtar, N., Jamil, H. M., Banik, R. S. & Asaduzzaman, S. M. TGF- $\beta$ /BMP signaling and other molecular events: Regulation of osteoblastogenesis and bone formation. *Bone Res.* **3**, (2015).
  26. Massagué, J., Seoane, J. & Wotton, D. Smad transcription factors. *Genes Dev.* **19**, 2783–2810 (2005).
  27. Brennan, J. *et al.* Nodal signalling in the epiblast patterns the early mouse embryo. *Nature* **411**, 965–969 (2001).
  28. Yuan, Z., Richardson, J. A., Parada, L. F. & Graft, J. M. Smad3 mutant mice develop metastatic colorectal cancer. *Cell* **94**, 703–714 (1998).
  29. Funayama, N. S. *et al.*  $\beta$ -Catenin Regulates Primitive Streak Induction through Collaborative Interactions with SMAD2/SMAD3 and OCT4. *Cell Stem Cell* **16**, 639–652 (2015).
  30. Nelson, A. C. *et al.* Global identification of smad2 and eomesodermin targets in zebrafish identifies a conserved transcriptional network in mesendoderm and a novel role for eomesodermin in repression of ectodermal gene expression. *BMC Biol.* **12**, (2014).
  31. Mullen, A. C. *et al.* Master transcription factors determine cell-type-specific responses to TGF- $\beta$  signaling. *Cell* **147**, 565–576 (2011).
  32. Arnold, S. J., Maretto, S., Islam, A., Bikoff, E. K. & Robertson, E. J. Dose-dependent Smad1, Smad5 and Smad8 signaling in the early mouse embryo. *Dev. Biol.* **296**, 104–118 (2006).
  33. Morikawa, M. *et al.* ChIP-seq reveals cell type-specific binding patterns of BMP-specific Smads and a novel binding motif. *Nucleic Acids Res.* **39**, 8712–8727 (2011).
  34. Chen, X. *et al.* Integration of External Signaling Pathways with the Core Transcriptional Network in Embryonic Stem Cells. *Cell* **133**, 1106–1117 (2008).
  35. Faial, T. *et al.* Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development* **142**, 2121–2135 (2015).
  36. Russ, A. P. *et al.* Eomesodermin is required for mouse trophoblast development and mesoderm formation. *Nature* **404**, 95–99 (2000).
  37. Costello, I. *et al.* The T-box transcription factor Eomesodermin acts upstream of Mesp1 to specify cardiac mesoderm during mouse gastrulation. *Nat. Cell Biol.* **13**, 1084–1092 (2011).
  38. Probst, S. & Arnold, S. J. Eomesodermin—At Dawn of Cell Fate Decisions During Early Embryogenesis. *Curr. Top. Dev. Biol.* **122**, 93–155 (2017).
  39. Bondue, A. *et al.* Mesp1 acts as a master regulator of multipotent cardiovascular progenitor specification. *Cell Stem Cell* **3**, 69–84 (2008).
  40. Lindsley, R. C. *et al.* Mesp1 coordinately regulates cardiovascular fate restriction and

- epithelial-mesenchymal transition in differentiating ESCs. *Cell Stem Cell* **3**, 55–68 (2008).
41. Wardle, F. C. & Papaioannou, V. E. Teasing out T-box targets in early mesoderm. *Curr. Opin. Genet. Dev.* **18**, 418–425 (2008).
  42. Herrmann, B. G. & Kispert, A. The T genes in embryogenesis. *Trends Genet.* **10**, 280–286 (1994).
  43. Smith, J. T-box genes: What they do and how they do it. *Trends Genet.* **15**, 154–158 (1999).
  44. Chesley, P. Development of the short-tailed mutant in the house mouse. *J. Exp. Zool.* **70**, 429–459 (1935).
  45. Gluecksohn-Schoenheimer, S. The Development of Normal and Homozygous Brachy (T/T) Mouse Embryos in the Extraembryonic Coelom of the Chick. *Proc. Natl. Acad. Sci.* **30**, 134–140 (1944).
  46. Herrmann, B. G., Labeit, S., Poustka, A., King, T. R. & Lehrach, H. Cloning of the T gene required in mesoderm formation in the mouse. *Nature* **343**, 617–622 (1990).
  47. Wilkinson, D. G., Bhatt, S. & Herrmann, B. G. Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature* **343**, 657–659 (1990).
  48. Koch, F. *et al.* Antagonistic Activities of Sox2 and Brachyury Control the Fate Choice of Neuro-Mesodermal Progenitors. *Dev. Cell* **42**, 514–526 (2017).
  49. Smith, J. C., Price, B. M. J., Green, J. B. A., Weigel, D. & Herrmann, B. G. Expression of a *Xenopus* homolog of Brachyury (T) is an immediate-early response to mesoderm induction. *Cell* **67**, 79–87 (1991).
  50. Kispert, A., Ortner, H., Cooke, J. & Herrmann, B. G. The chick Brachyury gene: Developmental expression pattern and response to axial induction by localized activin. *Dev. Biol.* **168**, 406–415 (1995).
  51. Cunliffe, V. & Smith, J. C. Ectopic mesoderm formation in *Xenopus* embryos caused by widespread expression of a Brachyury homologue. *Nature* **358**, 427–430 (1992).
  52. Kispert, A. & Herrmann, B. G. The Brachyury gene encodes a novel DNA binding protein. *EMBO J.* **12**, 3211–3220 (1993).
  53. Casey, E. S., O'Reilly, M. A., Conlon, F. L. & Smith, J. C. The T-box transcription factor Brachyury regulates expression of eFGF through binding to a non-palindromic response element. *Development* **125**, 3887–3894 (1998).
  54. Katikala, L. *et al.* Functional Brachyury Binding Sites Establish a Temporal Read-out of Gene Expression in the *Ciona* Notochord. *PLoS Biol.* **11**, (2013).
  55. Yamaguchi, T. P., Takada, S., Yoshikawa, Y., Wu, N. & McMahon, A. P. T (Brachyury) is a direct target of Wnt3a during paraxial mesoderm specification. *Genes Dev.* **13**, 3185–3190 (1999).
  56. Schulte-Merker, S. & Smith, J. C. Mesoderm formation in response to Brachyury requires FGF signalling. *Curr. Biol.* **5**, 62–67 (1995).
  57. Naiche, L. A., Holder, N. & Lewandoski, M. FGF4 and FGF8 comprise the wavefront activity that controls somitogenesis. *Proc. Natl. Acad. Sci.* **108**, 4018–4023 (2011).
  58. Ciruna, B. & Rossant, J. FGF Signaling Regulates Mesoderm Cell Fate Specification and Morphogenetic Movement at the Primitive Streak. *Dev. Cell* **1**, 37–49 (2001).
  59. Morley, R. H. *et al.* A gene regulatory network directed by zebrafish No tail accounts for its roles in mesoderm formation. *Proc. Natl. Acad. Sci.* **106**, 3829–3834 (2009).
  60. Gentsch, G. E. *et al.* In Vivo T-Box Transcription Factor Profiling Reveals Joint Regulation of Embryonic Neuromesodermal Bipotency. *Cell Rep.* **4**, 1185–1196 (2013).
  61. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell

- differentiation. *Nature* **518**, 344–349 (2015).
62. Lolas, M., Valenzuela, P. D. T., Tjian, R. & Liu, Z. Charting Brachyury-mediated developmental pathways during early mouse embryogenesis. *Proc. Natl. Acad. Sci.* **111**, 4478–4483 (2014).
  63. Gadue, P., Huber, T. L., Paddison, P. J. & Keller, G. M. Wnt and TGF- $\beta$  signaling are required for the induction of an in vitro model of primitive streak formation using embryonic stem cells. *Proc. Natl. Acad. Sci.* **103**, 16806–16811 (2006).
  64. Willems, E. & Leyns, L. Patterning of mouse embryonic stem cell-derived pan-mesoderm by Activin A/Nodal and Bmp4 signaling requires Fibroblast Growth Factor activity. *Differentiation* **76**, 745–759 (2008).
  65. Zhang, P. *et al.* Short-term BMP-4 treatment initiates mesoderm induction in human embryonic stem cells. *Blood* **111**, 1933–1941 (2008).
  66. Ernst, J., Vainas, O., Harbison, C. T., Simon, I. & Bar-Joseph, Z. Reconstructing dynamic regulatory maps. *Mol. Syst. Biol.* **3**, (2007).
  67. Schlitt, T. & Brazma, A. Current approaches to gene regulatory network modelling. *BMC Bioinformatics* **8**, (2007).
  68. Lee, T. I. *et al.* Transcriptional regulatory networks in *Saccharomyces cerevisiae*. *Science* **298**, 799–804 (2002).
  69. Williams, S. Pearson's correlation coefficient. *N. Z. Med. J.* **109**, 38 (1996).
  70. Cover, T. M. & Thomas, J. A. *Elements of Information Theory. Elements of Information Theory* (2005). doi:10.1002/047174882X
  71. Baba, K., Shibata, R. & Sibuya, M. Partial correlation and conditional correlation as measures of conditional independence. *Aust. New Zeal. J. Stat.* **46**, 657–664 (2004).
  72. Pournara, I. & Wernisch, L. Reconstruction of gene networks using Bayesian learning and manipulation experiments. *Bioinformatics* **20**, 2934–2942 (2004).
  73. Friedman, N. Inferring Cellular Networks Using Probabilistic Graphical Models. *Science* **303**, 799–805 (2004).
  74. Kauffman, S. Homeostasis and differentiation in random genetic control networks. *Nature* **224**, 177–178 (1969).
  75. Pinney, J. W. J. W., Westhead, D. R. D. R. & McConkey, G. A. G. A. Petri Net representations in systems biology. *Biochem. Soc. Trans.* **31**, 1513–1515 (2003).
  76. HARDY, S. & ROBILLARD, P. N. Modeling and simulation of molecular biology systems using petri nets: modeling goals of various approaches. *J. Bioinform. Comput. Biol.* **2**, 619–637 (2004).
  77. Wahde, M. *et al.* Modeling Genetic Regulatory Dynamics in Neural Development. *J. Comput. Biol.* **8**, 429–442 (2001).
  78. Ren, B. *et al.* Genome-wide location and function of DNA binding proteins. *Science* **290**, 2306–2309 (2000).
  79. Barski, A. *et al.* High-Resolution Profiling of Histone Methylations in the Human Genome. *Cell* **129**, 823–837 (2007).
  80. Pettersson, E., Lundeberg, J. & Ahmadian, A. Generations of sequencing technologies. *Genomics* **93**, 105–111 (2009).
  81. Park, P. J. ChIP-seq: Advantages and challenges of a maturing technology. *Nat. Rev. Genet.* **10**, 669–680 (2009).
  82. Wang, Z., Gerstein, M. & Snyder, M. RNA-Seq: A revolutionary tool for transcriptomics. *Nat. Rev. Genet.* **10**, 57–63 (2009).
  83. Clark, T. A., Sugnet, C. W. & Ares, M. Genomewide analysis of mRNA processing in yeast using splicing-specific microarrays. *Science* **296**, 907–910 (2002).
  84. Sexton, T. & Cavalli, G. The role of chromosome domains in shaping the functional

- genome. *Cell* **160**, 1049–1059 (2015).
85. Song, L. & Crawford, G. E. DNase-seq: A high-resolution technique for mapping active gene regulatory elements across the genome from mammalian cells. *Cold Spring Harb. Protoc.* **5**, (2010).
  86. Simon, J. M., Giresi, P. G., Davis, I. J. & Lieb, J. D. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat. Protoc.* **7**, 256–267 (2012).
  87. Adey, A. *et al.* Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol.* **11**, (2010).
  88. Hauke, J. & Kossowski, T. Comparison of values of pearson’s and spearman’s correlation coefficients on the same sets of data. *Quaest. Geogr.* **30**, 87–93 (2011).
  89. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Bioinformatics enrichment tools: Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Res.* **37**, 1–13 (2009).
  90. Hess, A. S. & Hess, J. R. Understanding tests of the association of categorical variables: the Pearson chi-square test and Fisher’s exact test. *Transfusion* **57**, 877–879 (2017).
  91. Fisher, R. A. On the Interpretation of  $\chi^2$  from Contingency Tables, and the Calculation of P. *J. R. Stat. Soc.* **85**, 87 (1922).
  92. Benjamini, Y. Simultaneous and selective inference: Current successes and future challenges. *Biometrical J.* **52**, 708–721 (2010).
  93. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B* **57**, 289–300 (1995).
  94. Pace, R. K. Maximum likelihood estimation. in *Handbook of Regional Science* 1553–1569 (2014). doi:10.1007/978-3-642-23430-9\_88
  95. Do, C. B. & Batzoglu, S. What is the expectation maximization algorithm? *Nat. Biotechnol.* **26**, 897–899 (2008).
  96. Jurafsky, D. & Martin, J. H. Hidden Markov Model. in *Speech and Language Processing (3rd ed. draft)* (2018).
  97. Baum, L. E. & Petrie, T. Statistical Inference for Probabilistic Functions of Finite State Markov Chains. *Ann. Math. Stat.* **37**, 1554–1563 (1966).
  98. Markov, A. A. Classical text in translation: An example of statistical investigation of the text eugene onegin concerning the connection of samples in chains. *Sci. Context* **19**, 591–600 (2006).
  99. Eisner, J. An interactive spreadsheet for teaching the forward-backward algorithm. *Proc. ACL-02 Work. Eff. tools Methodol. Teach. Nat. Lang. Process. Comput. Linguist.* - (2002). doi:10.3115/1118108.1118110
  100. Viterbi, A. J. Error Bounds for Convolutional Codes and an Asymptotically Optimum Decoding Algorithm. *IEEE Trans. Inf. Theory* **13**, 260–269 (1967).
  101. Baum, L. E. & Eagon, J. A. An inequality with applications to statistical estimation for probabilistic functions of Markov processes and to a model for ecology. *Bull. Am. Math. Soc.* **73**, 360–364 (1967).
  102. Schulz, M. H. *et al.* DREM 2.0: Improved reconstruction of dynamic regulatory networks from time-series expression data. *BMC Syst. Biol.* **6**, (2012).
  103. Barabási, A. L. & Oltvai, Z. N. Network biology: Understanding the cell’s functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
  104. Krishnapuram, B., Carin, L., Figueiredo, M. A. T. & Hartemink, A. J. Sparse multinomial logistic regression: Fast algorithms and generalization bounds. *IEEE Trans. Pattern Anal. Mach. Intell.* **27**, 957–968 (2005).
  105. Li, H. & Durbin, R. Fast and accurate short read alignment with Burrows-Wheeler

- transform. *Bioinformatics* **25**, 1754–1760 (2009).
106. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
  107. Li, R. *et al.* SOAP2: An improved ultrafast tool for short read alignment. *Bioinformatics* **25**, 1966–1967 (2009).
  108. Ning, Z., Cox, A. J. & Mullikin, J. C. SSAHA: A fast search method for large DNA databases. *Genome Res.* **11**, 1725–1729 (2001).
  109. Zhang, Y. *et al.* Model-based analysis of ChIP-Seq (MACS). *Genome Biol.* **9**, R137 (2008).
  110. Feng, J., Liu, T., Qin, B., Zhang, Y. & Liu, X. S. Identifying ChIP-seq enrichment using MACS. *Nat. Protoc.* **7**, 1728–1740 (2012).
  111. Trapnell, C., Pachter, L. & Salzberg, S. L. TopHat: discovering splice junctions with RNA-Seq. *Bioinformatics* **25**, 1105–11 (2009).
  112. Trapnell, C. *et al.* Transcript assembly and quantification by RNA-Seq reveals unannotated transcripts and isoform switching during cell differentiation. *Nat. Biotechnol.* **28**, 511–515 (2010).
  113. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).
  114. Shen, L. *et al.* diffReps: Detecting Differential Chromatin Modification Sites from ChIP-seq Data with Biological Replicates. *PLoS One* **8**, (2013).
  115. Ramírez, F. *et al.* deepTools2: a next generation web server for deep-sequencing data analysis. *Nucleic Acids Res.* **44**, w160–w165 (2016).
  116. Roth, F. P., Hughes, J. D., Estep, P. W. & Church, G. M. Finding DNA regulatory motifs within unaligned noncoding sequences clustered by whole-genome mRNA quantitation. *Nat. Biotechnol.* **16**, 939–945 (1998).
  117. Thijs, G. *et al.* A Gibbs Sampling Method to Detect Overrepresented Motifs in the Upstream Regions of Coexpressed Genes. *J. Comput. Biol.* **9**, 447–464 (2002).
  118. Heinz, S. *et al.* Simple Combinations of Lineage-Determining Transcription Factors Prime cis-Regulatory Elements Required for Macrophage and B Cell Identities. *Mol. Cell* **38**, 576–589 (2010).
  119. Bailey, T. L. & Elkan, C. Unsupervised Learning of Multiple Motifs in Biopolymers Using Expectation Maximization. *Mach. Learn.* **21**, 51–80 (1995).
  120. Lawrence, C. E. *et al.* Detecting subtle sequence signals: A gibbs sampling strategy for multiple alignment. *Science* **262**, 208–214 (1993).
  121. Hu, J., Li, B. & Kihara, D. Limitations and potentials of current motif discovery algorithms. *Nucleic Acids Res.* **33**, 4899–4913 (2005).
  122. Ashburner, M. *et al.* Gene ontology: Tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
  123. Dennis, G. *et al.* DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol.* **4**, R60 (2003).
  124. Maere, S., Heymans, K. & Kuiper, M. BiNGO: A Cytoscape plugin to assess overrepresentation of Gene Ontology categories in Biological Networks. *Bioinformatics* **21**, 3448–3449 (2005).
  125. Rivals, I., Personnaz, L., Taing, L. & Potier, M. C. Enrichment or depletion of a GO category within a class of genes: Which test? *Bioinformatics* **23**, 401–407 (2007).
  126. Hosack, D. A., Dennis Jr., G., Sherman, B. T., Lane, H. C. & Lempicki, R. A. Identifying biological themes within lists of genes with EASE. *Genome Biol.* **4**, R70 (2003).
  127. Huang, D. W., Sherman, B. T. & Lempicki, R. A. Systematic and integrative analysis

- of large gene lists using DAVID bioinformatics resources. *Nat. Protoc.* **4**, 44–57 (2009).
128. Spies, D. & Ciaudo, C. Dynamics in Transcriptomics: Advancements in RNA-seq Time Course and Downstream Analysis. *Comput. Struct. Biotechnol. J.* **13**, 469–477 (2015).
  129. Boyer, L. A. *et al.* Core Transcriptional Regulatory Circuitry in Human Embryonic Stem Cells. *Cell* **122**, 947–956 (2005).
  130. Pfendler, K. C., Catuar, C. S., Meneses, J. J. & Pedersen, R. a. Overexpression of Nodal promotes differentiation of mouse embryonic stem cells into mesoderm and endoderm at the expense of neuroectoderm formation. *Stem Cells Dev.* **14**, 162–172 (2005).
  131. Winnier, G., Blessing, M., Labosky, P. A. & Hogan, B. L. M. Bone morphogenetic protein-4 is required for mesoderm formation and patterning in the mouse. *Genes Dev.* **9**, 2105–2116 (1995).
  132. Bondue, A. & Blanpain, C. *Mesp1*: A key regulator of cardiovascular lineage commitment. *Circ. Res.* **107**, 1414–1427 (2010).
  133. Ren, X. *et al.* FOXF1 transcription factor is required for formation of embryonic vasculature by regulating VEGF signaling in endothelial cells. *Circ. Res.* **115**, 709–720 (2014).
  134. Yang, L. *et al.* Human cardiovascular progenitor cells develop from a KDR+embryonic-stem-cell-derived population. *Nature* **453**, 524–528 (2008).
  135. Kattman, S. J., Huber, T. L. & Keller, G. M. M. Multipotent Flk-1+Cardiovascular Progenitor Cells Give Rise to the Cardiomyocyte, Endothelial, and Vascular Smooth Muscle Lineages. *Dev. Cell* **11**, 723–732 (2006).
  136. Zhao, R. *et al.* Loss of both GATA4 and GATA6 blocks cardiac myocyte differentiation and results in acardia in mice. *Dev. Biol.* **317**, 614–619 (2008).
  137. Richards, A. A. & Garg, V. Genetics of congenital heart disease. *Curr. Cardiol. Rev.* **6**, 91–97 (2010).
  138. Ahmad, F. *et al.* The role of cardiac troponin T quantity and function in cardiac development and dilated cardiomyopathy. *PLoS One* **3**, (2008).
  139. Takahashi, Y. *et al.* Paired related homoeobox 1, a new EMT inducer, is involved in metastasis and poor prognosis in colorectal cancer. *Br. J. Cancer* **109**, 307–311 (2013).
  140. Cserjesi, P. *et al.* *MHox*: a mesodermally restricted homeodomain protein that binds an essential site in the muscle creatine kinase enhancer. *Development* **115**, 1087–1101 (1992).
  141. Bell, C. E. & Watson, A. J. *SNAI1* and *SNAI2* are asymmetrically expressed at the 2-cell stage and become segregated to the TE in the mouse blastocyst. *PLoS One* **4**, (2009).
  142. Gheldof, A., Hulpiau, P., van Roy, F., De Craene, B. & Berx, G. Evolutionary functional analysis and molecular regulation of the ZEB transcription factors. *Cell. Mol. Life Sci.* **69**, 2527–2541 (2012).
  143. Cho, S., Park, J. S., Kwon, S. & Kang, Y. K. Dynamics of *Setdb1* expression in early mouse development. *Gene Expr. Patterns* **12**, 213–218 (2012).
  144. Elling, U., Klasen, C., Eisenberger, T., Anlag, K. & Treier, M. Murine inner cell mass-derived lineages depend on *Sall4* function. *Proc. Natl. Acad. Sci.* **103**, 16319–16324 (2006).
  145. Lim, C. Y. *et al.* *Sall4* Regulates Distinct Transcription Circuitries in Different Blastocyst-Derived Stem Cell Lineages. *Cell Stem Cell* **3**, 543–554 (2008).
  146. Zhao, Q. *et al.* Developmental ablation of *Id1* and *Id3* genes in the vasculature leads to postnatal cardiac phenotypes. *Dev. Biol.* **349**, 53–64 (2011).



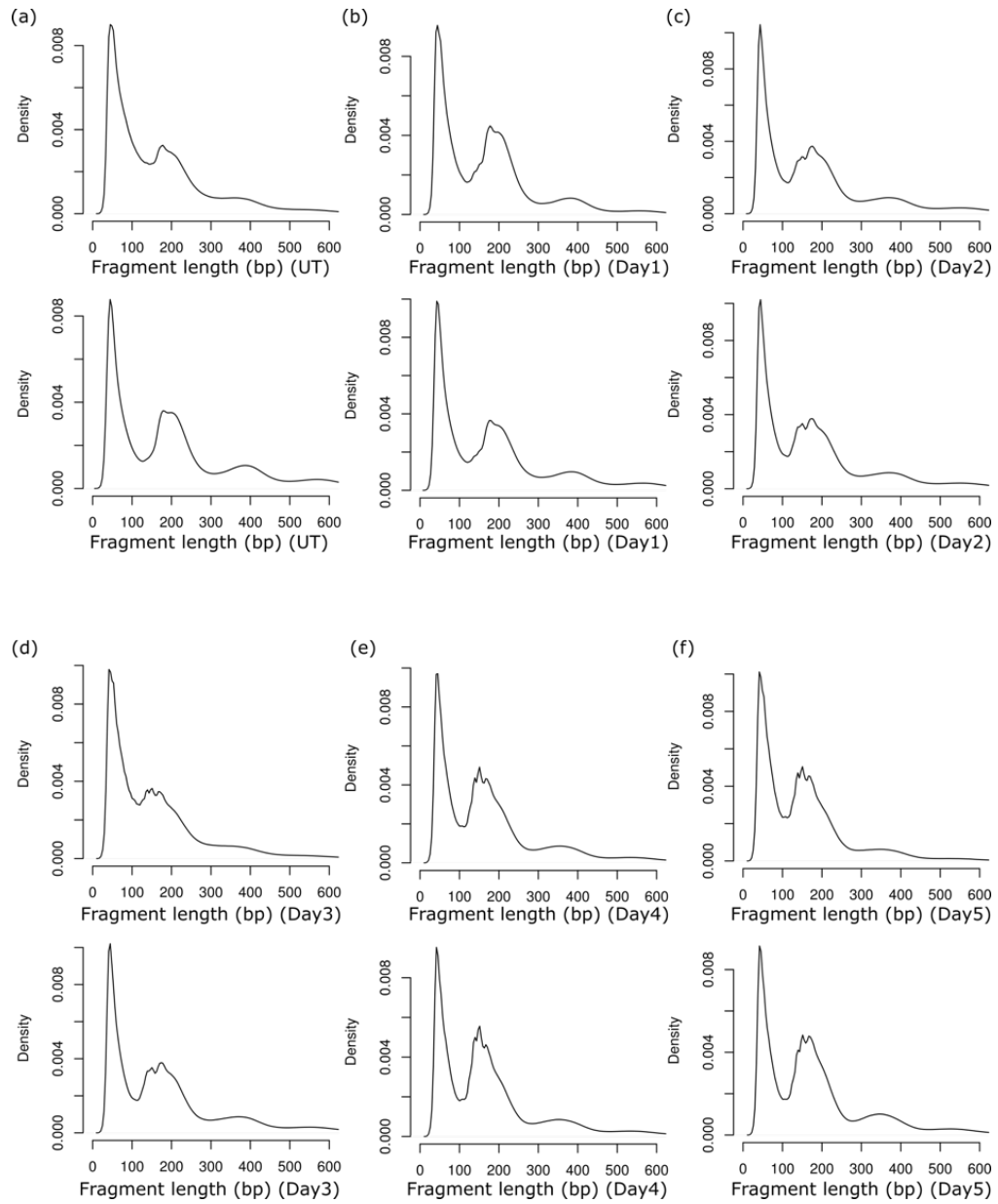
147. Qian, L., Mahaffey, J. P., Alcorn, H. L. & Anderson, K. V. Tissue-specific roles of Axin2 in the inhibition and activation of Wnt signaling in the mouse embryo. *Proc. Natl. Acad. Sci.* **108**, 8692–8697 (2011).
148. Yasuda, T. *et al.* PAX6 mutation as a genetic factor common to aniridia and glucose intolerance. *Diabetes* **51**, 224–230 (2002).
149. Badis, G. *et al.* Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).
150. Pearce, J. J. H. & Evans, M. J. Mml, a mouse Mix-like gene expressed in the primitive streak. *Mech. Dev.* **87**, 189–192 (1999).
151. Izzi, L. *et al.* Foxh1 recruits Gsc to negatively regulate Mixl1 expression during early mouse development. *EMBO J.* **26**, 3132–3143 (2007).
152. Simon, C. S. *et al.* Functional characterisation of cis -regulatory elements governing dynamic Eomes expression in the early mouse embryo. *Development* **144**, 1249–1260 (2017).
153. Dahle, Ø., Kumar, A. & Kuehn, M. R. Nodal signaling recruits the histone demethylase Jmjd3 to counteract Polycomb-mediated repression at target genes. *Sci. Signal.* **3**, (2010).
154. Herrmann, B. G., Labeit, S., Poustka, A., King, T. R. & Lehrach, H. Cloning of the T gene required in mesoderm formation in the mouse. *Nature* **343**, 617–622 (1990).
155. Wilkinson, D. G., Bhatt, S. & Herrmann, B. G. Expression pattern of the mouse T gene and its role in mesoderm formation. *Nature* **343**, 657–659 (1990).
156. Kispert, A., Koschorz, B. & Herrmann, B. G. The T protein encoded by Brachyury is a tissue-specific transcription factor. *EMBO J.* **14**, 4763–4772 (1995).
157. Gouti, M. *et al.* A Gene Regulatory Network Balances Neural and Mesoderm Specification during Vertebrate Trunk Development. *Dev. Cell* **41**, 243–261 (2017).
158. de Hoon, M. J. L., Imoto, S., Nolan, J. & Miyano, S. Open source clustering software. *Bioinformatics* **20**, 1453–1454 (2004).
159. Saldanha, A. J. Java Treeview - Extensible visualization of microarray data. *Bioinformatics* **20**, 3246–3248 (2004).
160. Geier, F., Timmer, J. & Fleck, C. Reconstructing gene-regulatory networks from time series, knock-out data, and prior knowledge. *BMC Syst. Biol.* **1**, (2007).
161. Lindsley, R. C. Canonical Wnt signaling is required for development of embryonic stem cell-derived mesoderm. *Development* **133**, 3787–3796 (2006).
162. Slagle, C. E., Aoki, T. & Burdine, R. D. Nodal-dependent mesendoderm specification requires the combinatorial activities of FoxH1 and eomesodermin. *PLoS Genet.* **7**, (2011).
163. Sawada, A. *et al.* Redundant Roles of Tead1 and Tead2 in Notochord Development and the Regulation of Cell Proliferation and Survival. *Mol. Cell. Biol.* **28**, 3177–3189 (2008).
164. von Both, I. *et al.* Foxh1 is essential for development of the anterior heart field. *Dev. Cell* **7**, 331–345 (2004).
165. Bhattaram, P. *et al.* Organogenesis relies on SoxC transcription factors for the survival of neural and mesenchymal progenitors. *Nat. Commun.* **1**, (2010).
166. Eferl, R. *et al.* Functions of c-Jun in liver and heart development. *J. Cell Biol.* **145**, 1049–1061 (1999).
167. Liu, Y. *et al.* Smad2 and Smad3 coordinately regulate craniofacial and endodermal development. *Dev. Biol.* **270**, 411–426 (2004).
168. Heineke, J. *et al.* Cardiomyocyte GATA4 functions as a stress-responsive regulator of angiogenesis in the murine heart. *J. Clin. Invest.* **117**, 3198–3210 (2007).

169. Zhou, W. *et al.* Modulation of morphogenesis by noncanonical Wnt signaling requires ATF/CREB family-mediated transcriptional activation of TGF $\beta$ 2. *Nat. Genet.* **39**, 1225–1234 (2007).
170. Hoogaars, W. M. H. *et al.* Tbx3 controls the sinoatrial node gene program and imposes pacemaker function on the atria. *Genes Dev.* **21**, 1098–1112 (2007).
171. Ware, S. M., Harutyunyan, K. G. & Belmont, J. W. Heart defects in X-linked heterotaxy: Evidence for a genetic interaction of Zic3 with the Nodal signaling pathway. *Dev. Dyn.* **235**, 1631–1637 (2006).
172. Guo, C. *et al.* A Tbx1-Six1/Eya1-Fgf8 genetic pathway controls mammalian cardiovascular and craniofacial morphogenesis. *J. Clin. Invest.* **121**, 1585–1595 (2011).
173. Zhang, Y., Liu, Y., Zhang, H., Wang, M. & Zhang, J. Mmu-MIR-351 attenuates the survival of cardiac arterial endothelial cells through targeting STAT3 in the atherosclerotic mice. *Biochem. Biophys. Res. Commun.* **468**, 300–305 (2015).
174. Wilks, D. S. *Cluster Analysis. International Geophysics* (2011). doi:10.1016/B978-0-12-385022-5.00015-4
175. Khatri, P. & Drăghici, S. Ontological analysis of gene expression data: Current tools, limitations, and open problems. *Bioinformatics* **21**, 3587–3595 (2005).
176. Janky, R. *et al.* iRegulon: From a Gene List to a Gene Regulatory Network Using Large Motif and Track Collections. *PLoS Comput. Biol.* **10**, (2014).
177. Li, X. Y. *et al.* Transcription factors bind thousands of active and inactive regions in the Drosophila blastoderm. *PLoS Biol.* **6**, (2008).
178. Sima, C., Hua, J. & Jung, S. Inference of Gene Regulatory Networks Using Time-Series Data: A Survey. *Curr. Genomics* **10**, 416–429 (2009).
179. Piper, J. *et al.* Wellington: A novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res.* **41**, (2013).
180. Shekhar, A. *et al.* Transcription factor ETV1 is essential for rapid conduction in the heart. *J. Clin. Invest.* **126**, 4444–4459 (2016).
181. Shekhar, A. *et al.* ETV1 activates a rapid conduction transcriptional program in rodent and human cardiomyocytes. *Sci. Rep.* **8**, (2018).
182. Liang, R. & Liu, Y. Tcf7l1 directly regulates cardiomyocyte differentiation in embryonic stem cells. *Stem Cell Res. Ther.* **9**, (2018).
183. Hoffman, B. G. & Jones, S. J. M. Genome-wide identification of DNA-protein interactions using chromatin immunoprecipitation coupled with flow cell sequencing. *J. Endocrinol.* **201**, 1–13 (2009).
184. Schbath, S. *et al.* Mapping Reads on a Genomic Sequence: An Algorithmic Overview and a Practical Comparative Analysis. *J. Comput. Biol.* **19**, 796–813 (2012).
185. Valouev, A. *et al.* Genome-wide analysis of transcription factor binding sites based on ChIP-Seq data. *Nat. Methods* **5**, 829–834 (2008).

# 7 Appendices

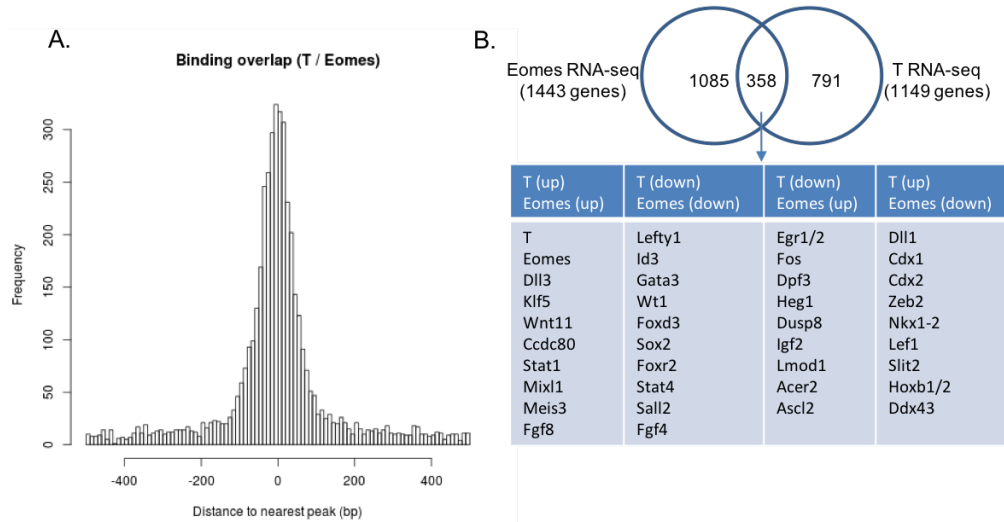
## A. Supplementary Figures

### Supplementary Figure 1 Size distribution of the mapped pair-end fragments



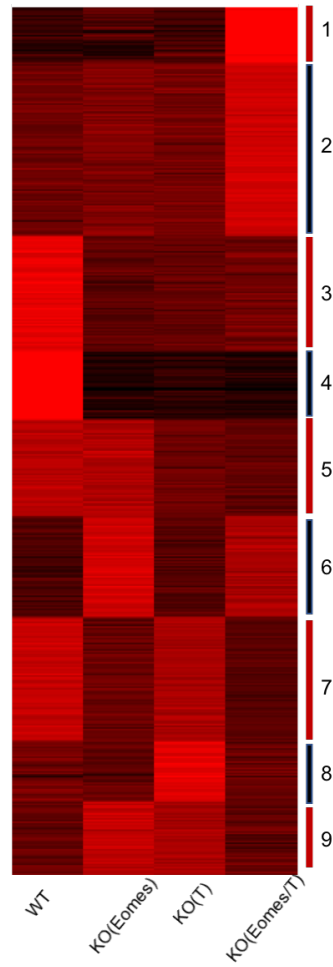
## Supplementary Figure 2 Combinatorial function of Eomes and T

A. Histogram showing the distance between Eomes and T ChIP peaks. B. Venn diagram showing the overlap of differential genes from Eomes and T WT/KO RNA-seq. Selected common downstream genes of Eomes and T are shown in the table.



**Supplementary Figure 3 Heatmap of differentially expressed genes (k-means clustering)**

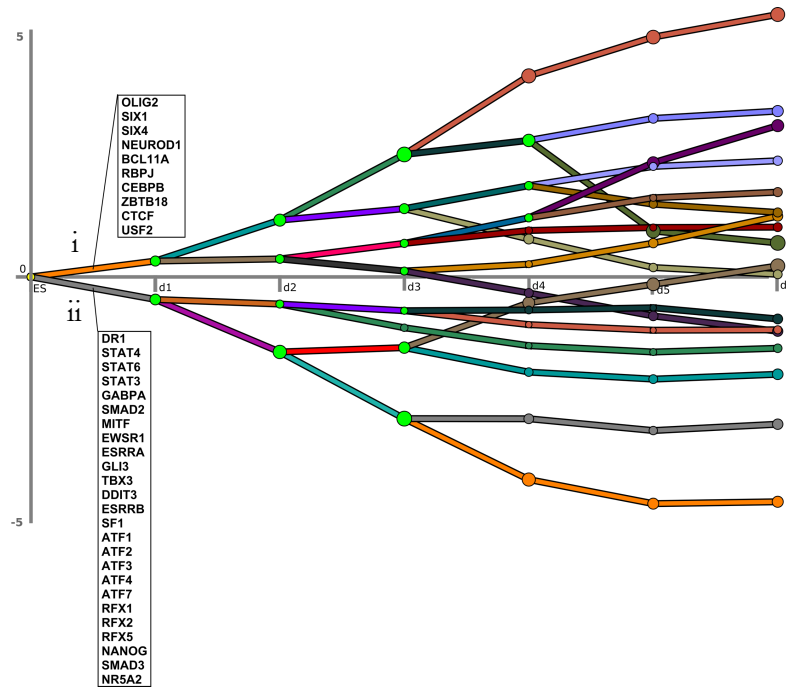
2145 DE genes obtained by pairwise comparisons of WT, T knockout, Eomes/T double knockout RNA-seq transcriptome data at d3 and grouped into 9 clusters by k-means clustering.



### Supplementary Figure 4 Enriched TFs assigned to the corresponding paths for each time point

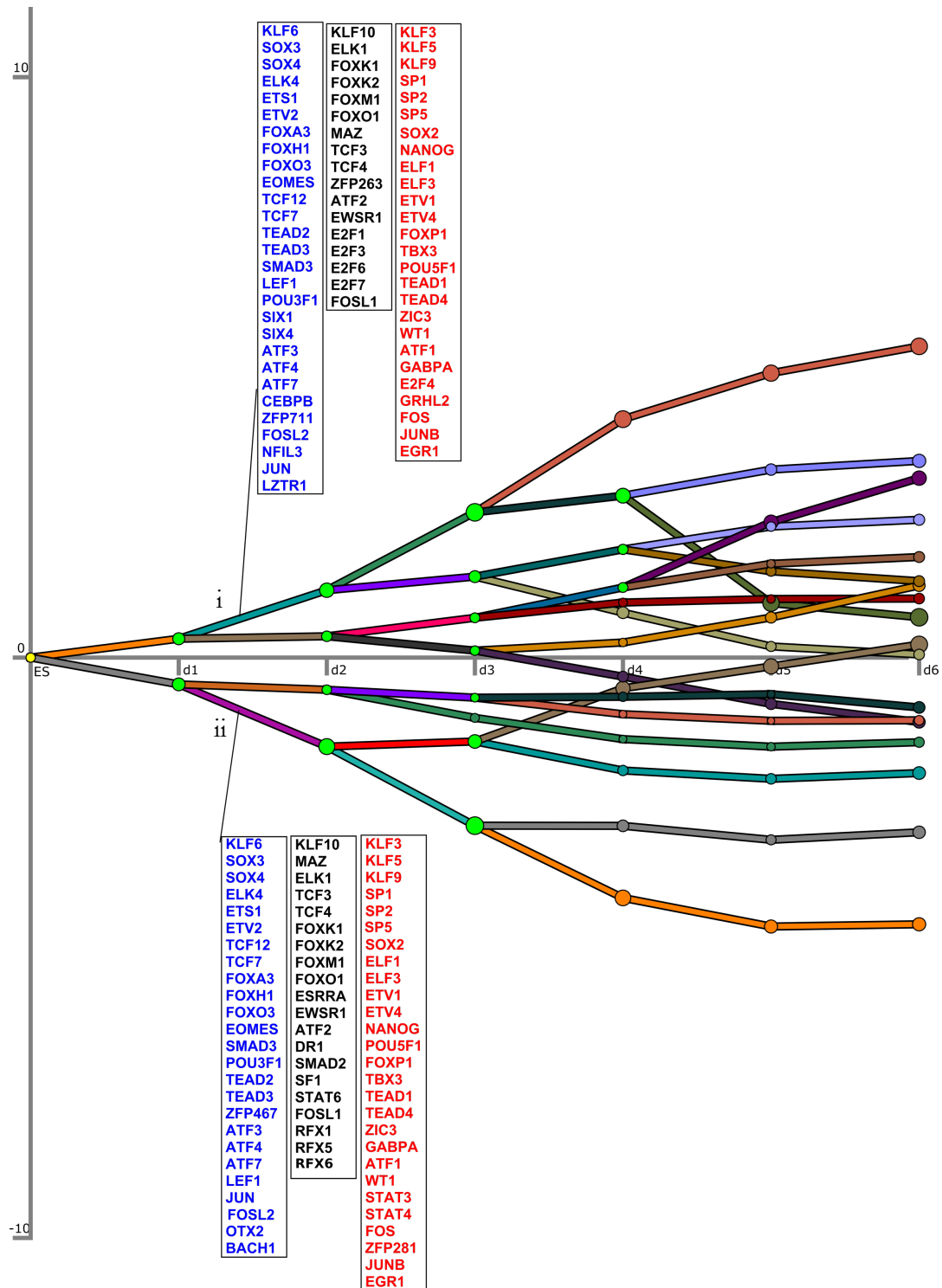
For the assigned TFs, the colors blue, black and red indicate “up-regulation”, “not changing” and “down-regulation” of current gene expression levels compared to ES separately. The horizontal axis indicates time points and the vertical axis indicates  $\log_2(V_x/V_{ES})$ , where  $V_x$  is the expression value for each corresponding time point.

**(A) Stage: ES to d1.** Assigned TFs (cutoff: X=4) and top 10 GO terms corresponding to the two paths diverging at the time point ES.



| Path no. | GO terms  | P-Value  | Benjamini |
|----------|---|----------|-----------|
| i        | lipid metabolic process                           | 3.20E-06 | 2.30E-02  |
|          | signal transduction                               | 1.80E-05 | 6.10E-02  |
|          | inflammatory response                             | 1.90E-05 | 4.50E-02  |
|          | positive regulation of cell migration             | 2.50E-05 | 4.40E-02  |
|          | positive regulation of osteoblast differentiation | 2.80E-05 | 3.90E-02  |
|          | cell migration                                    | 5.40E-05 | 6.20E-02  |
|          | cell adhesion                                     | 5.80E-05 | 5.70E-02  |
|          | cell-cell signaling                               | 5.90E-05 | 5.10E-02  |
|          | ossification                                      | 6.10E-05 | 4.70E-02  |
|          | palate development                                | 6.50E-05 | 4.50E-02  |
| ii       | mRNA processing                                   | 7.10E-26 | 3.80E-22  |
|          | translation                                       | 1.00E-21 | 2.80E-18  |
|          | RNA splicing                                      | 1.20E-20 | 2.10E-17  |
|          | rRNA processing                                   | 3.50E-20 | 4.60E-17  |
|          | ribosome biogenesis                               | 4.80E-20 | 5.10E-17  |
|          | DNA repair  | 8.20E-17 | 9.80E-14  |
|          | cellular response to DNA damage stimulus          | 2.80E-14 | 2.20E-11  |
|          | mRNA splicing, via spliceosome                    | 3.70E-13 | 2.40E-10  |
|          | transcription, DNA-templated                      | 1.30E-12 | 7.60E-10  |
|          | regulation of transcription, DNA-templated        | 1.80E-12 | 9.60E-10  |

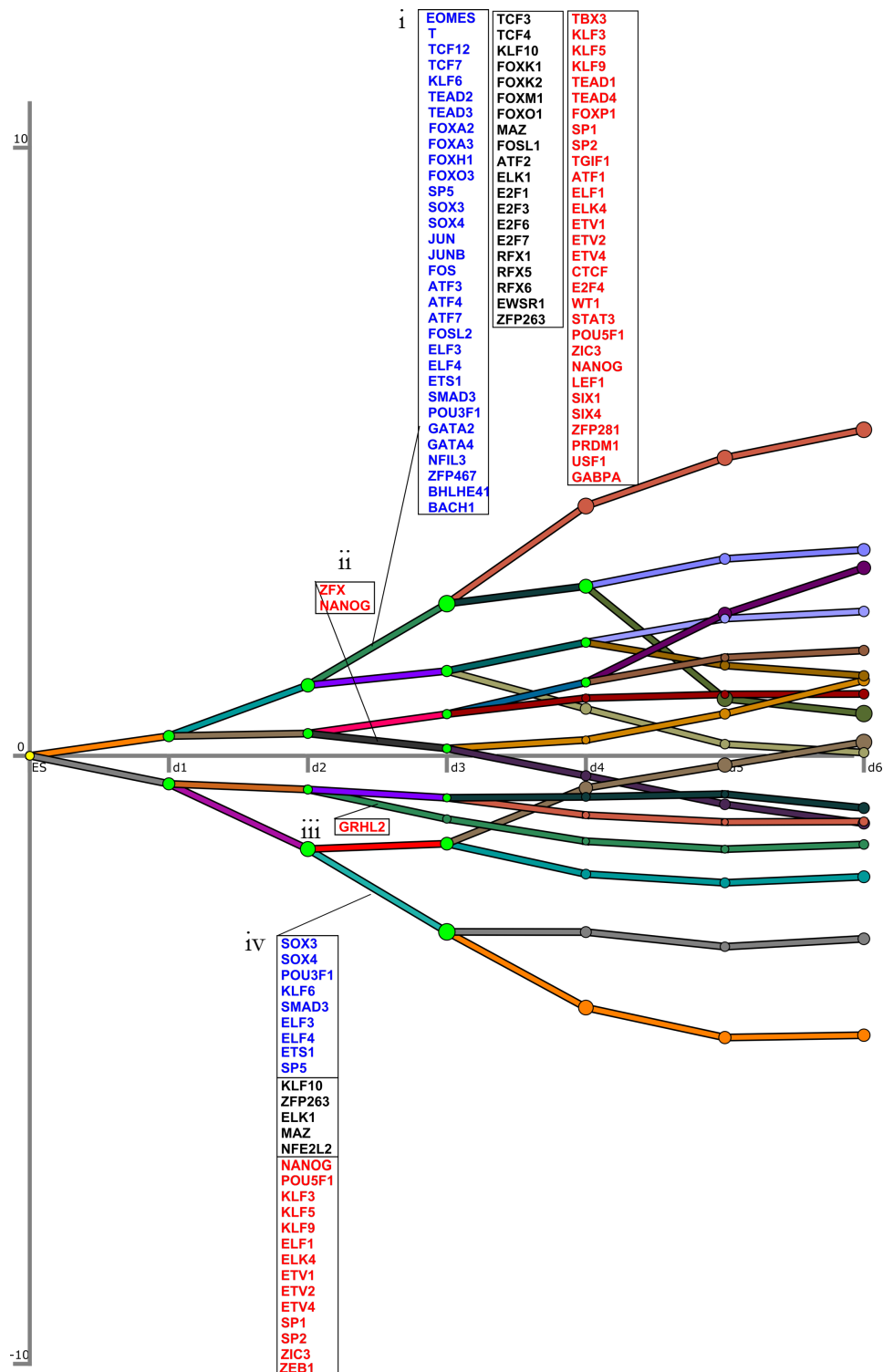
**(B) Stage: d1 to d2.** Assigned TFs (cutoff: X=12) and top 10 GO terms corresponding to the paths diverging at the time point d1.



| Path no. | GO terms   | P-Value  | Benjamini |
|----------|--|----------|-----------|
| i        | positive regulation of cell migration                                | 1.70E-07 | 7.80E-04  |
|          | multicellular organism development                                   | 7.30E-06 | 1.70E-02  |
|          | signal transduction  | 7.60E-06 | 1.20E-02  |
|          | positive regulation of MAPK cascade                                  | 1.30E-05 | 1.50E-02  |
|          | outflow tract morphogenesis  | 4.70E-05 | 4.30E-02  |
|          | negative regulation of canonical Wnt signaling pathway               | 4.80E-05 | 3.70E-02  |
|          | positive regulation of cell proliferation                            | 6.90E-05 | 4.50E-02  |
|          | signal transduction involved in regulation of gene expression        | 1.10E-04 | 6.30E-02  |
|          | Wnt signaling pathway  | 1.60E-04 | 8.10E-02  |
|          | cell differentiation   | 1.90E-04 | 8.40E-02  |
| ii       | stem cell population maintenance                                     | 1.50E-06 | 4.40E-03  |
|          | regulation of gene expression  | 8.60E-05 | 1.20E-01  |
|          | response to retinoic acid  | 4.00E-04 | 3.30E-01  |
|          | multicellular organism development                                   | 5.70E-04 | 3.50E-01  |
|          | cell differentiation   | 6.30E-04 | 3.20E-01  |
|          | meiotic cell cycle   | 6.60E-04 | 2.80E-01  |
|          | positive regulation of transcription from RNA polymerase II promoter | 1.30E-03 | 4.20E-01  |
|          | DNA methylation involved in gamete generation                        | 1.30E-03 | 3.80E-01  |
|          | spermatogenesis  | 1.50E-03 | 4.00E-01  |
|          | neural tube closure  | 2.50E-03 | 5.40E-01  |

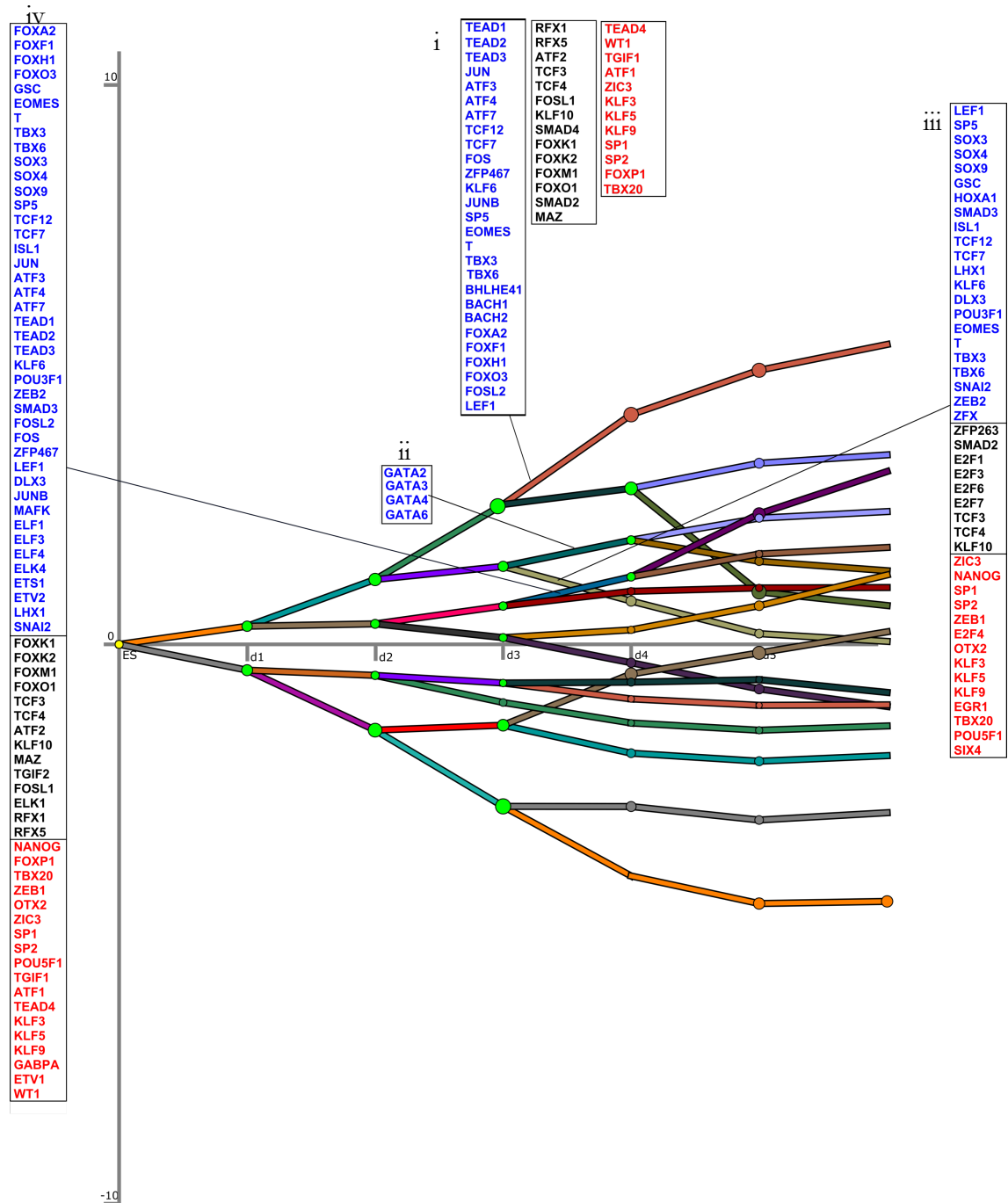


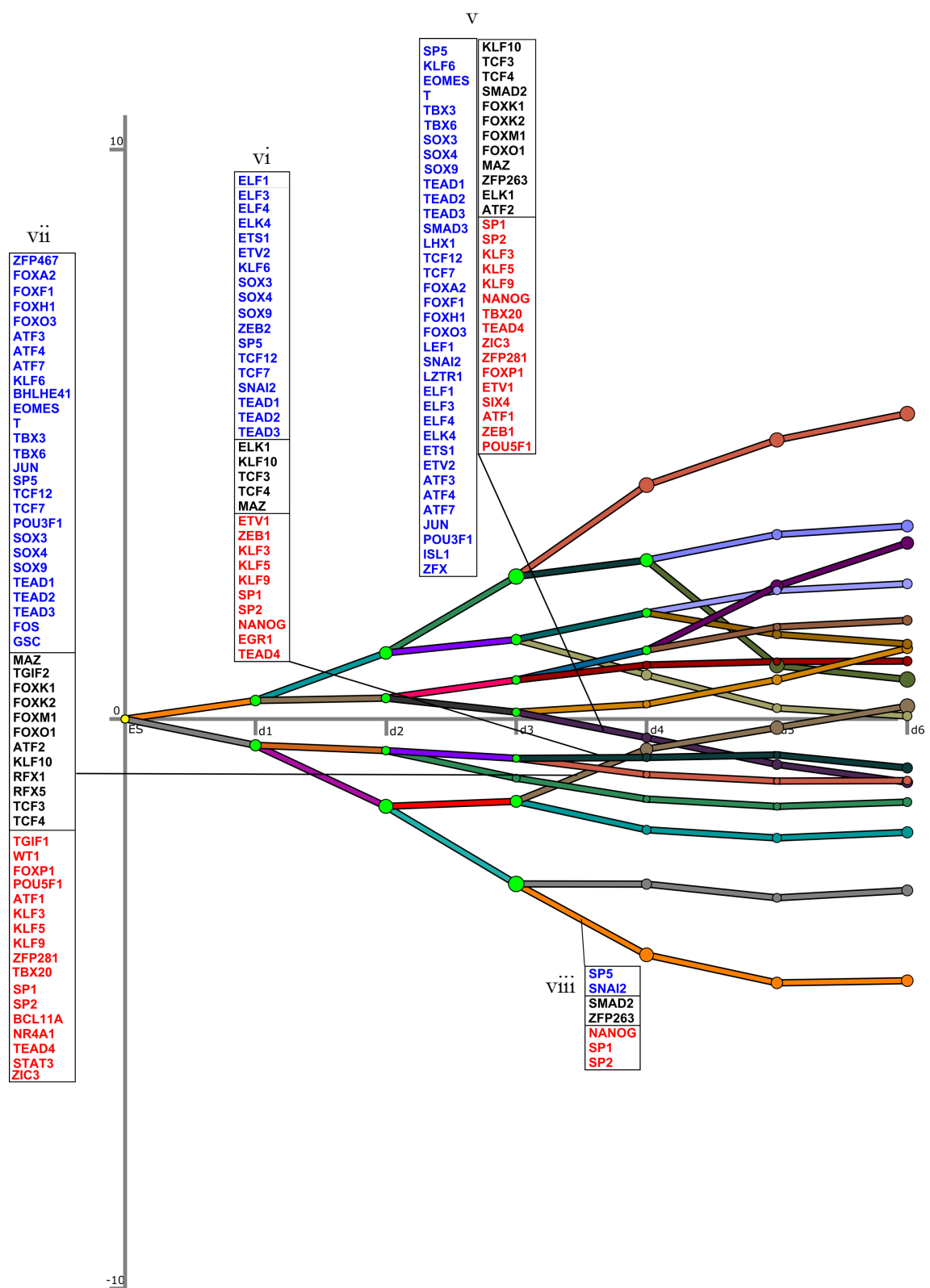
(C) Stage: d2 to d3. Assigned TFs (cutoff: X=4) and top 10 GO terms corresponding to the paths diverging at the time point d2.



| Path no.                                | GO terms   | P-Value  | Benjamini |
|---|--|----------|-----------|
| i                                       | multicellular organism development   | 7.90E-09 | 1.90E-05  |
|   | heart morphogenesis  | 8.60E-07 | 1.00E-03  |
|   | positive regulation of cell migration  | 2.00E-06 | 1.60E-03  |
|   | signal transduction involved in regulation of gene expression                            | 2.80E-06 | 1.70E-03  |
|   | positive regulation of cell proliferation  | 1.90E-05 | 9.20E-03  |
|   | odontogenesis  | 5.70E-05 | 2.30E-02  |
|   | outflow tract morphogenesis  | 5.90E-05 | 2.00E-02  |
|   | blood vessel remodeling  | 1.90E-04 | 5.60E-02  |
|   | collagen fibril organization   | 1.90E-04 | 5.60E-02  |
| positive regulation of angiogenesis     | 2.30E-04   | 6.00E-02 |           |
| ii                                      | positive regulation of gene expression, epigenetic                                       | 7.60E-05 | 1.80E-01  |
|   | DNA replication-dependent nucleosome assembly  | 7.60E-05 | 1.80E-01  |
|   | negative regulation of megakaryocyte differentiation                                     | 9.30E-05 | 1.10E-01  |
|   | DNA methylation on cytosine  | 1.50E-04 | 1.20E-01  |
|   | nucleosome assembly  | 3.70E-04 | 2.20E-01  |
|   | glutathione biosynthetic process   | 1.60E-03 | 5.80E-01  |
|   | DNA replication-independent nucleosome assembly  | 2.10E-03 | 6.00E-01  |
|   | protein heterotetramerization  | 2.60E-03 | 6.20E-01  |
|   | response to interleukin-1  | 3.50E-03 | 6.80E-01  |
| DNA-templated transcription, initiation | 4.60E-03   | 7.40E-01 |           |
| iii                                     | ribosome biogenesis  | 1.80E-09 | 3.10E-06  |
|   | rRNA processing  | 1.90E-06 | 1.70E-03  |
|   | maturation of SSU-rRNA from tricistronic rRNA transcript (SSU-rRNA, 5.8S rRNA, LSU-rRNA) | 2.10E-05 | 1.20E-02  |
|   | DNA replication  | 2.90E-04 | 1.20E-01  |
|   | translation  | 5.60E-04 | 1.80E-01  |
|   | mRNA processing  | 8.70E-04 | 2.30E-01  |
|   | positive regulation of protein targeting to mitochondrion                                | 1.20E-03 | 2.70E-01  |
|   | cellular response to DNA damage stimulus   | 1.30E-03 | 2.40E-01  |
|   | rRNA base methylation  | 1.60E-03 | 2.70E-01  |
| ribosomal large subunit assembly        | 1.80E-03   | 2.70E-01 |           |
| iv                                      | stem cell population maintenance   | 8.80E-06 | 8.80E-03  |
|   | multicellular organism development   | 3.80E-05 | 1.90E-02  |
|   | stem cell differentiation  | 1.50E-04 | 4.80E-02  |
|   | regulation of transcription, DNA-templated   | 1.30E-03 | 2.70E-01  |
|   | endodermal cell fate specification   | 5.30E-03 | 6.50E-01  |
|   | negative regulation of transposition   | 5.30E-03 | 6.50E-01  |
|   | regulation of MAPK cascade   | 5.40E-03 | 5.90E-01  |
|   | response to organic substance  | 7.20E-03 | 6.40E-01  |
|   | negative regulation of cell differentiation  | 7.20E-03 | 6.40E-01  |
| regulation of gene expression           | 9.20E-03   | 6.80E-01 |           |

**(D) Stage: d3 to d4.** Assigned TFs (cutoff: X=1.5) and top 10 GO terms corresponding to the paths diverging at the time point d3.

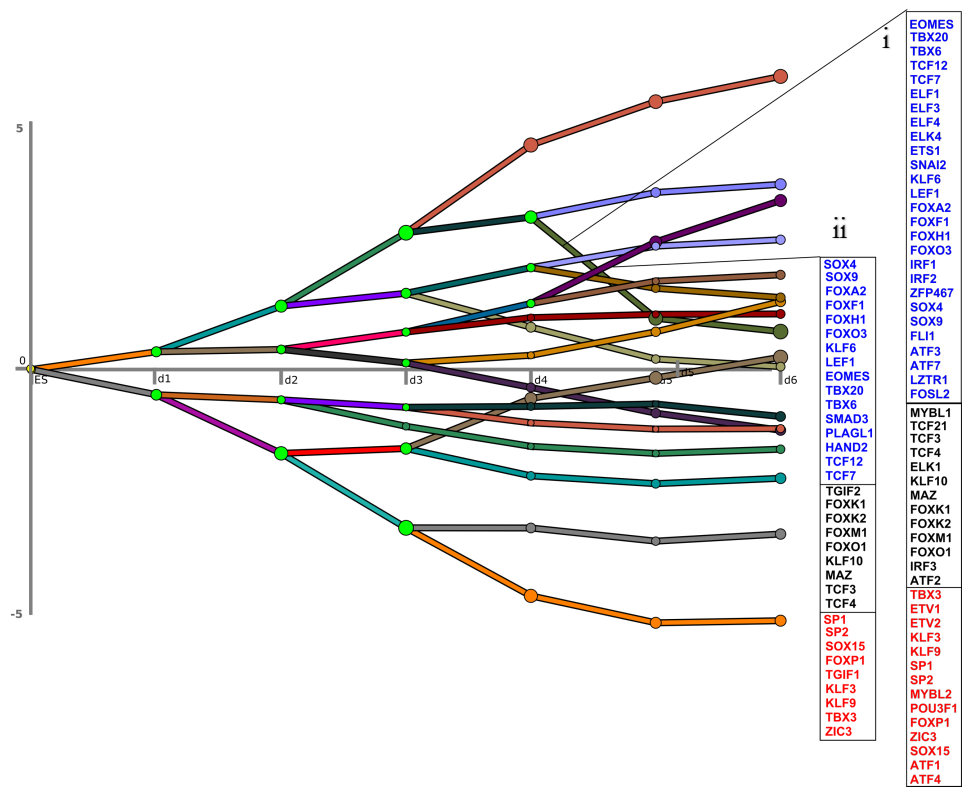
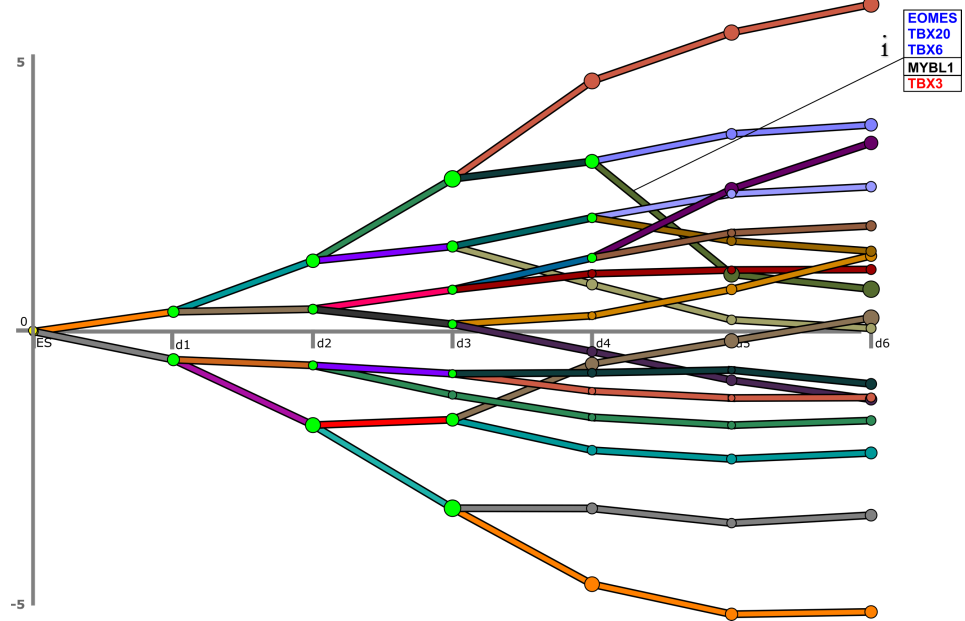




| Path no.  | GO terms  | P-Value  | Benjamini |
|---|---|----------|-----------|
| i   | actin-mediated cell contraction   | 5.60E-04 | 3.50E-01  |
|   | angiogenesis  | 1.20E-03 | 3.70E-01  |
|   | bleb assembly   | 1.80E-03 | 3.70E-01  |
|   | muscle contraction  | 4.20E-03 | 5.50E-01  |
|   | cardiac muscle contraction  | 4.70E-03 | 5.10E-01  |
|   | ossification  | 4.90E-03 | 4.60E-01  |
|   | extracellular matrix organization                                       | 6.00E-03 | 4.80E-01  |
|   | hepatocyte apoptotic process  | 6.40E-03 | 4.60E-01  |
|   | skeletal muscle contraction   | 1.10E-02 | 6.20E-01  |
| regulation of muscle contraction                | 1.50E-02  | 6.90E-01 |           |
| ii  | positive regulation of MAPK cascade                                     | 7.20E-04 | 8.90E-01  |
|   | dephosphorylation   | 2.10E-03 | 9.60E-01  |
|   | positive regulation of epithelial to mesenchymal transition             | 3.40E-03 | 9.70E-01  |
|   | positive regulation of I-kappaB kinase/NF-kappaB signaling              | 3.50E-03 | 9.30E-01  |
|   | receptor-mediated endocytosis   | 3.50E-03 | 8.90E-01  |
|   | neuron projection development   | 4.10E-03 | 8.80E-01  |
|   | negative regulation of canonical Wnt signaling pathway                  | 4.10E-03 | 8.80E-01  |
|   | unsaturated fatty acid biosynthetic process                             | 4.20E-03 | 8.40E-01  |
|   | aorta development   | 7.20E-03 | 9.40E-01  |
| phagocytosis, recognition                       | 7.70E-03  | 9.30E-01 |           |
| iii   | regulation of somitogenesis   | 4.20E-04 | 5.20E-01  |
|   | Notch signaling pathway   | 8.20E-04 | 5.10E-01  |
|   | signal transduction   | 3.10E-03 | 8.40E-01  |
|   | locomotory exploration behavior   | 3.30E-03 | 7.60E-01  |
|   | negative regulation of auditory receptor cell differentiation           | 6.60E-03 | 9.00E-01  |
|   | somite rostral/caudal axis specification                                | 1.00E-02 | 9.50E-01  |
|   | neuronal stem cell population maintenance                               | 1.10E-02 | 9.40E-01  |
|   | negative regulation of tumor necrosis factor-mediated signaling pathway | 1.30E-02 | 9.40E-01  |
|   | positive regulation of neuron apoptotic process                         | 1.70E-02 | 9.60E-01  |
| nervous system development                      | 1.90E-02  | 9.60E-01 |           |
| iv  | extracellular matrix organization                                       | 2.00E-04 | 4.00E-01  |
|   | decidualization   | 3.30E-04 | 3.40E-01  |
|   | cell adhesion   | 3.10E-03 | 9.20E-01  |
|   | cardiac muscle hypertrophy in response to stress                        | 3.60E-03 | 9.00E-01  |
|   | Angiogenesis  | 1.10E-02 | 1.00E+00  |
|   | negative regulation of peptidase activity                               | 1.40E-02 | 1.00E+00  |
|   | canonical Wnt signaling pathway   | 1.60E-02 | 1.00E+00  |
|   | positive regulation of endothelial cell proliferation                   | 2.30E-02 | 1.00E+00  |
|   | regulation of cell proliferation  | 2.70E-02 | 1.00E+00  |
| positive regulation of myoblast differentiation | 3.00E-02  | 1.00E+00 |           |
| v   | nucleosome assembly   | 1.70E-08 | 2.00E-05  |
|   | DNA replication-dependent nucleosome assembly                           | 4.40E-08 | 2.60E-05  |
|   | positive regulation of gene expression, epigenetic                      | 4.40E-08 | 2.60E-05  |
|   | DNA methylation on cytosine   | 9.30E-08 | 3.60E-05  |
|   | negative regulation of megakaryocyte differentiation                    | 1.50E-07 | 4.20E-05  |
|   | DNA replication-independent nucleosome assembly                         | 4.50E-06 | 1.00E-03  |
|   | DNA-templated transcription, initiation                                 | 1.10E-05 | 2.20E-03  |
|   | protein heterotetramerization   | 1.30E-05 | 2.20E-03  |
|   | chromatin silencing at rDNA   | 4.10E-05 | 5.90E-03  |
| tRNA aminoacylation for protein translation     | 2.60E-03  | 2.80E-01 |           |

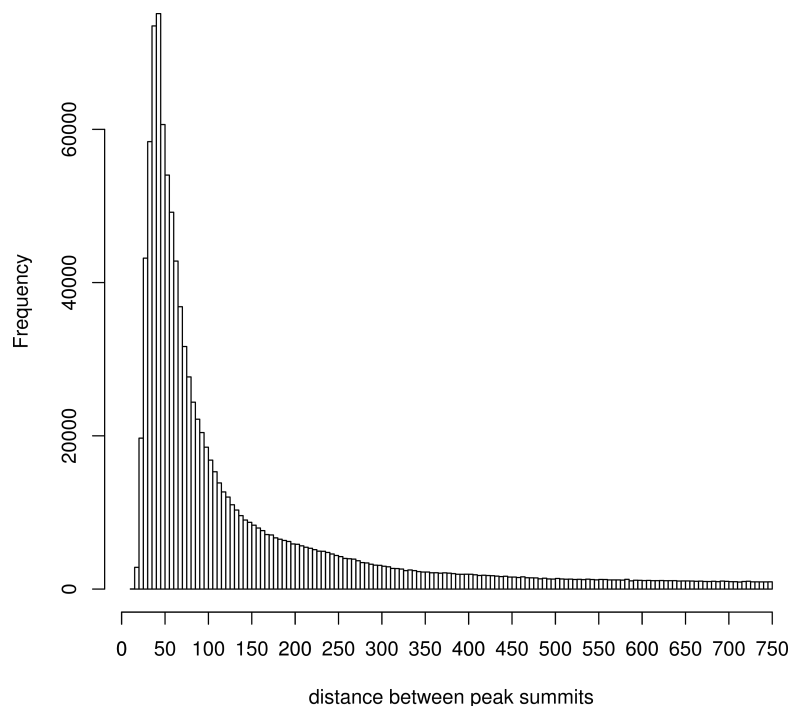
|      |   |          |          |
|------|---|----------|----------|
| vi   | mitotic nuclear division                              | 3.80E-08 | 5.00E-05 |
|      | cell division   | 5.30E-06 | 3.50E-03 |
|      | chromosome segregation                                | 1.90E-05 | 8.20E-03 |
|      | DNA recombination                                     | 1.50E-03 | 3.80E-01 |
|      | cytoplasmic translation                               | 1.90E-03 | 3.90E-01 |
|      | cell cycle  | 3.70E-03 | 5.50E-01 |
|      | DNA repair  | 5.10E-03 | 6.10E-01 |
|      | translation   | 6.20E-03 | 6.40E-01 |
|      | regulation of mitotic centrosome separation           | 1.30E-02 | 8.40E-01 |
|      | GDP-mannose biosynthetic process                      | 1.30E-02 | 8.40E-01 |
| vii  | cell adhesion   | 5.90E-05 | 7.30E-02 |
|      | negative regulation of cell division                  | 3.10E-04 | 1.80E-01 |
|      | positive regulation of epidermal cell differentiation | 9.20E-03 | 9.80E-01 |
|      | neural tube closure                                   | 1.00E-02 | 9.60E-01 |
|      | wound healing   | 1.00E-02 | 9.60E-01 |
|      | response to peptide hormone                           | 1.10E-02 | 9.40E-01 |
|      | negative regulation of cell proliferation             | 1.10E-02 | 9.10E-01 |
|      | regulation of mitotic cell cycle                      | 1.20E-02 | 8.90E-01 |
|      | response to drug                                      | 1.30E-02 | 8.80E-01 |
|      | cellular response to estrogen stimulus                | 1.70E-02 | 9.10E-01 |
| viii | stem cell population maintenance                      | 1.90E-06 | 7.70E-04 |
|      | multicellular organism development                    | 7.40E-05 | 1.50E-02 |
|      | endodermal cell fate specification                    | 6.80E-04 | 9.00E-02 |
|      | stem cell differentiation                             | 8.70E-04 | 8.60E-02 |
|      | regulation of genetic imprinting                      | 1.70E-03 | 1.30E-01 |
|      | regulation of transcription, DNA-templated            | 2.30E-03 | 1.50E-01 |
|      | response to organic substance                         | 2.60E-03 | 1.40E-01 |
|      | negative regulation of cell differentiation           | 2.60E-03 | 1.40E-01 |
|      | response to retinoic acid                             | 5.20E-03 | 2.40E-01 |
|      | spermatogenesis                                       | 6.00E-03 | 2.40E-01 |

**(E) Stage: d4 to d5.** Assigned TFs (cutoff: X=12 (upper figure), X=3 (lower figure)) and top 10 GO terms corresponding to the paths diverging at the time point d4.



| Path no.  | GO terms  | P-Value         | Benjamini |
|---|---|-----------------|-----------|
| i   | signal transduction involved in regulation of gene expression | 7.80E-07        | 5.60E-04  |
|   | gastrulation  | 3.80E-05        | 1.30E-02  |
|   | multicellular organism development                            | 6.10E-05        | 1.40E-02  |
|   | anterior/posterior pattern specification                      | 3.10E-04        | 5.30E-02  |
|   | heart morphogenesis   | 6.30E-04        | 8.60E-02  |
|   | somitogenesis   | 9.70E-04        | 1.10E-01  |
|   | Wnt signaling pathway   | 1.10E-03        | 1.00E-01  |
|   | immune system process   | 1.20E-03        | 1.00E-01  |
|   | neuron differentiation  | 3.40E-03        | 2.40E-01  |
|   | camera-type eye development                                   | 4.20E-03        | 2.60E-01  |
|   | ii  | decidualization | 1.30E-03  |
| angiogenesis  |   | 1.70E-03        | 5.60E-01  |
| cell-cell signaling                                     |   | 3.60E-03        | 6.90E-01  |
| positive regulation of endothelial cell proliferation   |   | 5.90E-03        | 7.60E-01  |
| cellular response to fibroblast growth factor stimulus  |   | 6.70E-03        | 7.30E-01  |
| positive regulation of leukocyte migration              |   | 6.90E-03        | 6.80E-01  |
| neutrophil chemotaxis                                   |   | 7.70E-03        | 6.60E-01  |
| positive regulation of smooth muscle cell proliferation |   | 9.80E-03        | 7.00E-01  |
| chemokine-mediated signaling pathway                    |   | 1.40E-02        | 7.90E-01  |
| positive regulation of ERK1 and ERK2 cascade            |   | 1.70E-02        | 8.10E-01  |

**Supplementary Figure 5 Distribution of the distances between peak summits**





## B. Supplementary Tables

**Supplementary Table 1. Percentages of aligned reads (10 samples of time-series RNA-seq transcriptome analysis)**

|                    | <i>ES</i>           | <i>Hour 1</i>       | <i>Hour 6</i>       | <i>Hour 12</i>      | <i>Day 1</i>        |
|--------------------|---------------------|---------------------|---------------------|---------------------|---------------------|
| <i>Replicate 1</i> | 19314823<br>(54.8%) | 17189196<br>(50.5%) | 20137110<br>(55.8%) | 19304223<br>(52.6%) | 19402668<br>(54.0%) |
| <i>Replicate 2</i> | 24957462<br>(58.1%) | 20167200<br>(55.4%) | 21614314<br>(59.2%) | 23345278<br>(57.4%) | 20866930<br>(57.2%) |
|                    | <i>Day 2</i>        | <i>Day 3</i>        | <i>Day 4</i>        | <i>Day 5</i>        | <i>Day 6</i>        |
| <i>Replicate 1</i> | 19450099<br>(55.8%) | 20817155<br>(56.2%) | 20666929<br>(55.8%) | 19915476<br>(58.0%) | 24045613<br>(61.3%) |
| <i>Replicate 2</i> | 21322284<br>(58.2%) | 23365189<br>(56.2%) | 21291829<br>(59.4%) | 20780005<br>(57.5%) | 21951764<br>(59.9%) |

**Supplementary Table 2. Pearson's correlation between 2 replicates**

Pearson's correlation was calculated for two replicates of 10 samples. The results show that the replicates are significantly related.

| Two RNA-seq Replicates | Pearson's Correlation | p-value  |
|------------------------|-----------------------|----------|
| ES                     | 0.977                 | <2.2e-16 |
| 1h                     | 0.930                 | <2.2e-16 |
| 6h                     | 0.978                 | <2.2e-16 |
| 12h                    | 0.965                 | <2.2e-16 |
| d1                     | 0.989                 | <2.2e-16 |
| d2                     | 0.963                 | <2.2e-16 |
| d3                     | 0.978                 | <2.2e-16 |
| d4                     | 0.994                 | <2.2e-16 |
| d5                     | 0.997                 | <2.2e-16 |
| d6                     | 0.992                 | <2.2e-16 |

**Supplementary Table 3. Marker genes differentially expressed in the time course**

FPKM values are shown. Different colors separate the marker genes for pluripotent cells, PS-like/nascent mesoderm, early cardiac mesoderm, EMT cells and committed cardiac cells (Figure 4.1).

|        | ES      | 1h      | 6h      | 12h     | d1      | d2      | d3     | d4     | d5    | d6     |
|--------|---------|---------|---------|---------|---------|---------|--------|--------|-------|--------|
| Pou5f1 | 1689.92 | 1651.68 | 1786.46 | 1545.65 | 1298.75 | 1264.17 | 804.87 | 192.23 | 32.71 | 23.24  |
| Sox2   | 261.37  | 231.7   | 203.31  | 170.47  | 105.97  | 22.12   | 21.31  | 13.31  | 7.8   | 5.39   |
| Dppa3  | 160.32  | 143.94  | 155.3   | 107.83  | 59.48   | 25      | 11.56  | 5.51   | 6.17  | 5.42   |
| Klf4   | 9.85    | 9.93    | 1.4     | 0.59    | 0.37    | 0.45    | 0.73   | 0.62   | 1.11  | 3.83   |
| Esrrb  | 76.08   | 81.3    | 35.25   | 18.24   | 6.01    | 2.39    | 1.57   | 1.08   | 0.67  | 1.07   |
| Nodal  | 14.64   | 24.29   | 10.4    | 7.56    | 10.51   | 33.95   | 32.88  | 2.37   | 1.3   | 0.64   |
| Nanog  | 106.08  | 103.29  | 65.64   | 23.59   | 11.31   | 86.41   | 86.02  | 17.89  | 5.1   | 3.99   |
| Eomes  | 1.08    | 1.27    | 1.09    | 1.77    | 2.7     | 11.45   | 21.57  | 10.1   | 0.22  | 0.19   |
| T      | 0.13    | 0.03    | 0.09    | 0.17    | 1.38    | 54.13   | 419.17 | 71.47  | 2.64  | 0.37   |
| Fgf5   | 0.97    | 1.47    | 2.36    | 5.35    | 9.14    | 12.41   | 28.69  | 4.66   | 0.26  | 0.61   |
| Wnt3   | 0.12    | 0.49    | 0.33    | 0.63    | 2.62    | 17.47   | 27.61  | 6.46   | 1.05  | 1.51   |
| Wnt3a  | 0.44    | 0.31    | 0.08    | 0       | 0.01    | 0.04    | 2.01   | 0.06   | 0     | 0.04   |
| Msgn1  | 0       | 0       | 0       | 0       | 0.03    | 0       | 23.75  | 14.85  | 0.44  | 0.15   |
| Mesp1  | 0       | 0       | 0.03    | 0       | 0.19    | 0       | 5.47   | 18.31  | 0.77  | 0.03   |
| Wnt5a  | 0.71    | 1.01    | 0.73    | 0.47    | 0.59    | 0.09    | 4.18   | 10.64  | 2.76  | 1.98   |
| Foxf1  | 0       | 0.1     | 0       | 0.08    | 0.15    | 0.1     | 0.82   | 3.33   | 2.98  | 1.13   |
| Kdr    | 0.21    | 0.2     | 0.14    | 0.01    | 0.06    | 0.09    | 0.5    | 3.67   | 4.09  | 1.68   |
| Cdh1   | 146.71  | 165.78  | 140.3   | 146.99  | 157.6   | 177.03  | 149.51 | 115.72 | 85.87 | 135.96 |
| Cdh2   | 1.37    | 2.15    | 1.36    | 1.2     | 1.8     | 1.19    | 6.2    | 12.38  | 4.78  | 8.93   |
| Cdh11  | 0.34    | 0.56    | 0.57    | 0.71    | 0.74    | 0.49    | 1.55   | 1.76   | 5.73  | 5.65   |
| Snai1  | 0.07    | 0.12    | 0       | 0       | 0.12    | 0.12    | 0.64   | 1.14   | 0.72  | 1.63   |
| Snai2  | 0.16    | 0.17    | 0.33    | 0.14    | 0.02    | 0.06    | 0.86   | 2.94   | 8.62  | 5.27   |
| Zeb1   | 3.13    | 3.2     | 2.95    | 1.85    | 0.95    | 0.79    | 2.32   | 8.63   | 7.97  | 6.47   |
| Zeb2   | 0.54    | 0.65    | 0.42    | 0.34    | 0.32    | 1.52    | 6.44   | 11.65  | 6.8   | 5.66   |
| Prrx1  | 0       | 0       | 0.01    | 0       | 0       | 0.05    | 0.27   | 0.35   | 1.89  | 2.22   |
| Tbx20  | 0.51    | 0.6     | 0.3     | 0.17    | 0.6     | 0.9     | 0.36   | 2.16   | 3.36  | 1.87   |
| Hand2  | 0       | 0       | 0.02    | 0       | 0.03    | 0       | 0      | 0.84   | 19.02 | 11.58  |
| Gata4  | 0.29    | 0.44    | 0.24    | 0.5     | 0.57    | 0.58    | 1.3    | 3.39   | 3.9   | 4.88   |
| Gata5  | 0       | 0       | 0       | 0       | 0       | 0.01    | 0      | 0.23   | 2.68  | 1.5    |
| Gata6  | 0       | 0.08    | 0.15    | 0.04    | 0.03    | 0.01    | 0.48   | 1.78   | 7.38  | 8.03   |
| Myh6   | 0.66    | 0.86    | 0.76    | 1.11    | 1.88    | 0.54    | 0.2    | 0.24   | 0.94  | 2.69   |
| Tnnt2  | 0.1     | 0.09    | 0.16    | 0.05    | 0       | 0.03    | 1.17   | 1.87   | 14.63 | 39.67  |

**Supplementary Table 4. Percentages of aligned reads (ATAC-seq samples)**

|                    | <i>ES</i>            | <i>Day 1</i>         | <i>Day 2</i>         | <i>Day 3</i>         | <i>Day 4</i>         | <i>Day 5</i>         |
|--------------------|----------------------|----------------------|----------------------|----------------------|----------------------|----------------------|
| <i>Replicate 1</i> | 26603898<br>(51.09%) | 36078686<br>(55.57%) | 39762760<br>(64.90%) | 29054098<br>(52.61%) | 41666860<br>(68.17%) | 20108105<br>(27.55%) |
| <i>Replicate 2</i> | 31665953<br>(63.77%) | 31796275<br>(65.79%) | 33428551<br>(64.80%) | 45090932<br>(66.38%) | 41537823<br>(68.75%) | 47344619<br>(70.00%) |

**Supplementary Table 5. TF groups**

| TF groups | Names of TFs from Homer  |
|-----------|--|
| AP        | AP-2alpha, AP-2gamma   |
| ATF       | Atf1, Atf2, Atf3, Atf4, Atf7   |
| Bach      | Bach1, Bach2   |
| E2F       | E2F1, E2F3, E2F4, E2F6, E2F7   |
| EBF       | EBF, EBF1, EBF2  |
| ELK       | ELK1, ELK3, ELK5, ETS, ETS1, ETV1, ETV4, Elf4, Elk1, Elk4, Ets1-distal, Etv2,                |
| TBOX      | Eomes, T, Tbr1, Tbx20, Tbx21, Tbx5, Tbx6   |
| FOX       | FOXA1, FOXK1, FOXK2, FOXM1, FOXP1, Fox:Ebox, FoxL2, Foxa2, Foxa3, Foxf1, Foxh1, Foxo1, Foxo3 |
| GATA      | GATA, GATA3, GATA:SCL, Gata1, Gata2, Gata4, Gata6,   |
| HOX       | HOXA1, HOXB13, Hoxa10, Hoxa9, Hoxb4, Hoxc9, Hoxd10   |
| IRF       | IRF1, IRF2, IRF3, IRF4, IRF8   |
| KLF       | KLF10, KLF14, KLF3, KLF5, KLF6, KLF4, KLF9   |
| WNT       | LEF1, TCF4, TCFL2, Txf12, Tcf21, Tcf3, Tcf7  |
| LHX       | Lhx1, Lhx2, Lhx3   |
| MEF2      | Mef2a, Mef2b, Mef2c, Mef2d   |
| OCT       | OCT:OCT, OCT:OCT-short, Oct2, Oct4, Oct6, Oct4:Sox17   |
| RFX       | Rfx1, Rfx2, Rfx5, Rfx6   |
| STAT      | STAT1, STAT4, STAT5, STAT6   |
| SIX       | Six1, Six2, Six4   |
| SOX       | Sox2, Sox3, Sox4, Sox6, Sox9, Sox10, Sox15, Sox17  |
| SP        | Sp1, Sp2, Sp5  |
| TEAD      | TEAD, TEAD1, TEAD2, TEAD3, TEAD4   |

**Supplementary Table 6. GO terms for each path of the gene regulatory tree constructed by using the time-series expression data (Figure 4.19)**

The GO terms were selected on the basis of “p-value < 0.05”.

| <b>GO terms (selected) for each path</b>                    | <b>p-value</b> | <b>related genes (selected)</b>                |
|---|----------------|--|
| <b>1</b>  |                |  |
| actin-mediated cell contraction                             | 5.64E-4        | ACTC1, EMP2, PARVA                             |
| angiogenesis  | 1.21E-3        | COL4A2, COL4A1, FLT1, HAND1, TGFB1             |
| cardiac muscle contraction                                  | 4.73E-3        | ACTC1, MYL4, TNNC1, TNNI1                      |
| extracellular matrix organization                           | 5.99E-3        | COL4A2, APP, COL4A1, TGFB1, CCDC80             |
| ventricular cardiac muscle tissue morphogenesis             | 2.25E-2        | HAND1, TNNC1, TNNI1                            |
|   |                |  |
| <b>2</b>  |                |  |
| positive regulation of cell migration                       | 8.47E-7        | EGFR, IRS2, PDGFB, PODXL, FURIN                |
| outflow tract morphogenesis                                 | 5.55E-5        | DHRS3, JUN, VEGFA, TGFB3, SEMA3C               |
| positive regulation of smooth muscle cell proliferation     | 3.59E-4        | EGFR, CYBA, PDGFB, ID2, JUN                    |
| blood vessel development                                    | 5.68E-4        | DLX3, PDGFB, VEGFA, PDGFRB, TGFB3              |
| atrial septum morphogenesis                                 | 6.72E-4        | SMO, ISL1, CFC1, TGFB2, CYR61                  |
|   |                |  |
| <b>3</b>  |                |  |
| angiogenesis  | 1.69E-3        | FGFR2, ARHGAP22, CCL2, NRP1, HAND2             |
| positive regulation of smooth muscle cell proliferation     | 9.77E-3        | FGFR2, AKT1, PTGS2, HBEGF, ITGB3               |
| positive regulation of ERK1 and ERK2 cascade                | 1.69E-2        | FGFR2, CCL2, NRP1, C3, HAND2                   |
| positive regulation of canonical Wnt signaling pathway      | 2.95E-2        | WNT2, FGFR2, COL1A1, BAMBI                     |
| cardiac muscle tissue development                           | 3.55E-2        | SIN3B, GATA6, CSRP3                            |
|   |                |  |
| <b>4</b>  |                |  |
| positive regulation of epithelial to mesenchymal transition | 7.30E-4        | GLIPR2, DAB2, BMP2, TGFB2, SMAD3, TGFB11, CRB2 |
| erythrocyte differentiation                                 | 4.62E-3        | THRA, LYN, GATA3, JAK2, HEPH                   |
|   |                |  |
| <b>5</b>  |                |  |
| extracellular matrix organization                           | 1.21E-3        | MPZL3, RECK, FBLN1, LAMA4, LAMB2               |
| cardiac muscle hypertrophy in response to stress            | 1.61E-2        | MEF2C, GATA4, MYH7, PPP3CA                     |
|   |                |  |
| <b>6</b>  |                |  |
| fatty acid metabolic process                                | 1.99E-2        | SCD1, PRKAR2B, CD36, CPT2, STAT5B              |
| establishment of epithelial cell apical/basal polarity      | 3.03E-2        | WNT5A, FOXF1, CRB3                             |
| inner ear morphogenesis                                     | 3.56E-2        | WNT5A, FGFR1, MYO6, COL2A1, FRZB               |
|   |                |  |
| <b>7</b>  |                |  |
| inflammatory response                                       | 1.76E-3        | NFKBIZ, S100A8, LY96, RELA, IL19               |
| positive regulation of chondrocyte differentiation          | 2.10E-3        | RELA, SOX5, ZBTB16, SOX6, MUSTN1               |

|   |          |                                     |
|---|----------|-------------------------------------|
| blood coagulation                           | 2.03E-2  | PTPRJ, FGG, THBD, PROCR, PDGFA      |
| cartilage development                       | 3.84E-2  | SMAD9, EDN1, SOX5, PRRX1, ZBTB16    |
| skeletal system development                 | 4.35E-2  | HAPLN1, TBX3, HEXB, COL3A1, EDN1    |
|   |          |                                     |
| 8   |          |                                     |
| protein transport                           | 4.14E-5  | COPA, AP1G2, RAB5B, SLC15A2, LMAN2L |
| cholesterol biosynthetic process            | 9.42E-4  | CYP51, MVD, DHCR7, INSIG1, HMGCS1   |
| protein phosphorylation                     | 1.42E-3  | RNASEL, CDK18, NUA2, RORC, LATS1    |
| determination of left/right symmetry        | 3.43E-3  | MEGF8, KIF3A, FOXJ1, DRC1, DYNC2L1  |
| carbohydrate metabolic process              | 4.86E-3  | PHKA2, LDHA, GNPDA2, PHKB, GALK1    |
|   |          |                                     |
| 9   |          |                                     |
| gastrulation                                | 3.77E-5  | CER1, APLNR, EOMES, MESP1, MIXL1    |
| anterior/posterior pattern specification    | 3.05E-4  | CER1, ALDH1A2, T, WNT3, FOXA2       |
| heart morphogenesis                         | 6.31E-4  | ALDH1A2, T, FGF8, FOXC1, MESP1      |
| somitogenesis                               | 9.70E-4  | T, DLL3, FOXC1, MSGN1, AXIN2        |
| Wnt signaling pathway                       | 1.08E-3  | WNT3, WNT5B, AXIN2, WNT8A, PITX2    |
| neuron differentiation                      | 3.37E-3  | ALDH1A2, WNT5B, WNT8A, DDIT4, PITX2 |
| neural crest cell development               | 9.85E-3  | ALDH1A2, CYP26A1, FOXC1             |
| mesodermal cell migration                   | 2.35E-2  | FGF8, MESP1                         |
|   |          |                                     |
| 10  |          |                                     |
| cell adhesion                               | 5.89E-5  | ICAM1, PTPRK, NID1, ACKR3, CDH3     |
| neural tube closure                         | 1.03E-2  | DLC1, BMP4, SFRP1, PTCH1, GRHL3     |
| regulation of mitotic cell cycle            | 1.19E-2  | PIM3, PTCH1, SIK1, MYC              |
| substrate adhesion-dependent cell spreading | 2.01E-2  | SFRP1, TEK, LAMC1, FN1              |
| extracellular matrix disassembly            | 2.09E-2  | LAMA1, NID1, LAMC1                  |
|   |          |                                     |
| 11  |          |                                     |
| regulation of somitogenesis                 | 4.20E-4  | CDX1, NOTCH1, CDX2, DLL1            |
| Notch signaling pathway                     | 8.23E-4  | DTX4, NOTCH3, HES1, S1PR3, NOTCH1   |
| neuronal stem cell population maintenance   | 1.13E-2  | HES1, NOTCH1, DLL1, FOXO3, PROX1    |
| nervous system development                  | 1.91E-2  | NES, MAGI2, TRNP1, BHLHE22, FGF13   |
| hypothalamus development                    | 3.02E-2  | SOX3, CRH, LEF1                     |
|   |          |                                     |
| 12  |          |                                     |
| mitotic nuclear division                    | 3.83E-8  | KIF11, TADA3, NEK2, BORA, PAPD5     |
| chromosome segregation                      | 1.89E-5  | KIF2C, CEP85, KIF11, NEK2, CDCA2    |
| cell cycle                                  | 3.66E-3  | CKAP2, KIF11, NEK2, BORA, PAPD5     |
|   |          |                                     |
| 13  |          |                                     |
| mRNA processing                             | 1.23E-28 | NCBP1, APOBEC1, PRPF4B, SCAF4, RNMT |
| translation                                 | 4.20E-16 | EIF6, TARS2, RPL13, EIF5, EIF5B     |
| mRNA export from nucleus                    | 1.93E-9  | NUP133, NCBP1, SMG5, SMG7, DDX39B   |
| ribosome biogenesis                         | 1.66E-8  | EIF6, FASTKD2, SURF6, GAR1, GTPBP10 |
|   |          |                                     |

|  |         |   |
|--|---------|---|
| 14   |         |   |
| nucleosome assembly  | 1.74E-8 | HIST1H4N, HIST2H3B, HIST1H2BM, HIST1H4K |
| DNA methylation on cytosine  | 9.34E-8 | HIST1H4N, HIST2H3B, HIST1H4K, HIST1H4B  |
| chemical synaptic transmission                                       | 9.37E-3 | MYO5A, DOC2A, SNCA, CLSTN1, APBA2       |
| regulation of neuron apoptotic process                               | 3.24E-2 | GABRB3, SNCA, TRP73, SIGMAR1            |
|  |         |   |
| 15   |         |   |
| ribosome biogenesis  | 1.75E-9 | NAF1, KRR1, SDAD1, NOC4L, TSR1          |
| DNA replication  | 2.89E-4 | RECQL4, TICRR, NASP, POLE, POLA1        |
| translation  | 5.56E-4 | TUFM, SLC25A5, MRPS12, MRPS24, RPL27    |
| mRNA processing  | 8.69E-4 | SRSF1, PDCD11, PPIL1, SYNCRIP, SRSF2    |
| regulation of transcription, DNA-templated                           | 5.88E-3 | XRCC5, E2F2, 5730507C01RIK, TAF1A, E2F4 |
|  |         |   |
| 16   |         |   |
| circadian rhythm   | 2.79E-4 | TRP53, DDC, DHX9, KLF9, MAT2A           |
| spermatogenesis  | 1.81E-3 | DNMT3A, MAEL, MOV10L1, ARNTL, SIRT1     |
| positive regulation of neuron projection development                 | 3.82E-2 | TWF2, RRN3, ENC1, NGFR, SIRT1           |
| neural tube closure  | 4.55E-2 | ENAH, RARG, SALL4, SALL1, ZIC5          |
|  |         |   |
| 17   |         |   |
| negative regulation of transcription from RNA polymerase II promoter | 3.55E-2 | CTBP2, ZFP57, JARID2, ARID5B, NODAL     |
| embryonic placenta development                                       | 4.17E-2 | NODAL, TTPA, FOXD3                      |
|  |         |   |
| 18   |         |   |
| stem cell population maintenance                                     | 1.86E-6 | NANOG, POU5F1, ESRRB, SOX2, TET1        |
| endodermal cell fate specification                                   | 6.76E-4 | NANOG, POU5F1, SOX2                     |
| regulation of genetic imprinting                                     | 1.67E-3 | ZFP42, DPPA3, TET1                      |
| spermatogenesis  | 5.97E-3 | RPL10L, DNMT3L, HORMAD1, HSF2BP         |

### C. Supplementary Notes

#### Supplementary Note 1. Adapter sequences of ATAC-seq reads

```
# adapter.fa
>1
TCGTCGGCAGCGTCAGATGTGTATAAGAGACAG
>2
CTGTCTCTTATACACATCTGACGCTGCCGACGA
>3
GTCTCGTGGGCTCGGAGATGTGTATAAGAGACAG
>4
CTGTCTCTTATACACATCTCCGAGCCCACGAGAC
```

## Supplementary Note 2. Artefact regions

| chromosome | Start     | End       |
|------------|-----------|-----------|
| 1          | 24611520  | 24616200  |
| 1          | 56782052  | 56782440  |
| 1          | 102628176 | 102628344 |
| 1          | 122357123 | 122357318 |
| 1          | 183299000 | 183299700 |
| 1          | 195241453 | 195241934 |
| 2          | 3050000   | 3055000   |
| 2          | 5379000   | 5379400   |
| 2          | 22587283  | 22590547  |
| 2          | 69355530  | 69355677  |
| 2          | 90395030  | 90395240  |
| 2          | 98662100  | 98667600  |
| 2          | 181917300 | 181919500 |
| 2          | 181926500 | 181933000 |
| 3          | 8245690   | 8246640   |
| 3          | 3000000   | 3035000   |
| 3          | 5860300   | 5860900   |
| 4          | 3049000   | 3258000   |
| 4          | 34935690  | 34935910  |
| 4          | 70378040  | 70378320  |
| 4          | 80002398  | 80005206  |
| 5          | 146260900 | 146261400 |
| 6          | 3200500   | 3202000   |
| 6          | 9889890   | 9890178   |
| 6          | 49236447  | 49236621  |
| 6          | 79818087  | 79818352  |
| 6          | 103649000 | 103649350 |
| 7          | 20779500  | 20792000  |
| 7          | 21256500  | 21269000  |
| 8          | 15519790  | 15520030  |
| 8          | 19784000  | 19785400  |
| 8          | 20320500  | 20378500  |
| 9          | 3000000   | 3038320   |
| 9          | 24541940  | 24542200  |
| 9          | 35305000  | 35305700  |
| 9          | 110281220 | 110281400 |
| 9          | 123461750 | 123462250 |

|    |           |           |
|----|-----------|-----------|
| 10 | 22142530  | 22143070  |
| 10 | 96077520  | 96077900  |
| 10 | 130594400 | 130594960 |
| 11 | 3122500   | 3201000   |
| 11 | 54139940  | 54140740  |
| 11 | 109011640 | 109012110 |
| 12 | 3109850   | 3110150   |
| 12 | 78350850  | 78351180  |
| 12 | 97061423  | 97061800  |
| 13 | 44869540  | 44870050  |
| 13 | 77438870  | 77439090  |
| 13 | 85126554  | 85127518  |
| 13 | 97190460  | 97190690  |
| 13 | 119595000 | 119603400 |
| 13 | 119609000 | 119617700 |
| 14 | 19415700  | 19419750  |
| 15 | 75085300  | 75087150  |
| 16 | 11143909  | 11144324  |
| 16 | 57391357  | 57391690  |
| 17 | 36231250  | 36231650  |
| 17 | 39842900  | 39848900  |
| 17 | 70936660  | 70963897  |
| 18 | 3005550   | 3006050   |
| 18 | 12949190  | 12949400  |
| 18 | 40307970  | 40308340  |
| 18 | 68691990  | 68692230  |
| 19 | 45650030  | 45650310  |
| 19 | 61199640  | 61199880  |
| 19 | 61266550  | 61267210  |
| X  | 143483000 | 143483150 |



## D. List of Figures

|   |    |
|---|----|
| Figure 1.1 The typical structure of a TF .....  | 2  |
| Figure 1.2 The process of mouse embryogenesis .....   | 4  |
| Figure 1.3 Cell lineage commitment during blastulation.....   | 5  |
| Figure 1.4 Cellular mechanism of EMT .....  | 8  |
| Figure 1.5 Mesoderm formation through EMT during gastrulation.....  | 9  |
| Figure 1.6 Smads function as intracellular signaling mediators .....  | 11 |
| Figure 1.7 Different types of gene regulatory network (GRN) models.....   | 15 |
| Figure 2.1 Overview of the chromatin immunoprecipitation (ChIP) experiment.....   | 18 |
| Figure 2.2 Overview of the ChIP-seq process .....   | 20 |
| Figure 2.3 ATAC-seq process and advantages .....  | 22 |
| Figure 3.1 A Markov chain for three different weather conditions .....  | 29 |
| Figure 3.2 An example of hidden Markov model .....  | 30 |
| Figure 3.3 An illustration of forward algorithm.....  | 32 |
| Figure 3.4 An example of calculating Viterbi path probability .....   | 34 |
| Figure 3.5 An illustration of a DREM (version 2.0) analysis .....   | 38 |
| Figure 3.6 Hashing algorithm .....  | 43 |
| Figure 3.7 Burrows-Wheeler transform .....  | 43 |
| Figure 3.8 Workflow of MACS .....   | 45 |
| Figure 3.9 Modeling the shift size of ChIP-seq tags.....  | 46 |
| Figure 4.1 Heatmap showing differential expression of marker genes during the time course of<br>mesoderm formation..... | 56 |
| Figure 4.2 Hierarchical clustering of 9888 differentially expressed genes .....   | 57 |
| Figure 4.3 Phosphorylation levels of Smad1/5/8 and Smad2/3 proteins detected by Western blotting .                      | 63 |
| Figure 4.4 Genes up- or down-regulated by Smad4 KO .....  | 64 |
| Figure 4.5 ChIP-seq analysis of Smads.....  | 65 |
| Figure 4.6 Venn diagram showing direct target genes of Smad1 and Smad2/3 .....  | 66 |
| Figure 4.7 Expression levels of Eomes during the differentiation time course detected by Western<br>blotting.....       | 68 |
| Figure 4.8 ChIP-seq analysis of Eomes .....   | 68 |
| Figure 4.9 Venn diagram showing direct target genes of Eomes .....  | 69 |
| Figure 4.10 Expression levels of T during the differentiation time course detected by Western blotting                  | 70 |
| Figure 4.11 ChIP-seq analysis of T .....  | 71 |
| Figure 4.12 Venn diagram showing direct target genes of T.....  | 71 |
| Figure 4.13 Three major steps to build the dynamic gene regulatory network/tree.....                                    | 73 |
| Figure 4.14 Spearman's rank correlation coefficients between samples.....   | 75 |

|   |     |
|---|-----|
| Figure 4.15 An example of the identified dips.....  | 76  |
| Figure 4.16 An example of the identified differential regions.....  | 76  |
| Figure 4.17 Schematic procedure of ATAC-seq analysis to build TF-target gene interactions .....   | 77  |
| Figure 4.18 The distributions of merged d1, d2, d3 ATAC-seq signals around Eomes motifs and d2, d3, d4<br>ATAC-seq signals around T motifs..... | 78  |
| Figure 4.19 The tree structure constructed using the time-series expression data.....   | 81  |
| Figure 4.20 Bifurcation events in the dynamic regulatory network controlling stem cell population<br>maintenance genes.....                     | 85  |
| Figure 4.21 Bifurcation events in the dynamic regulatory network controlling cardiovascular system<br>development.....                          | 89  |
| Figure 4.22 Bifurcation events in the dynamic regulatory network controlling EMT process.....   | 92  |
| Figure 4.23 Paths from “A” to “D” marked for the enrichment analysis showing in.....  | 94  |
| Figure 4.24 Paths from “a” to “h” marked for the enrichment analysis showing in.....  | 96  |
| <br>  |     |
| Supplementary Figure 1 Size distribution of the mapped pair-end fragments .....   | 119 |
| Supplementary Figure 2 Combinatorial function of Eomes and T.....   | 120 |
| Supplementary Figure 3 Heatmap of differentially expressed genes (k-means clustering) .....   | 121 |
| Supplementary Figure 4 Enriched TFs assigned to the corresponding paths for each time point .....   | 122 |
| Supplementary Figure 5 Distribution of the distances between peak summits.....  | 132 |

## E. List of Tables

|   |     |
|---|-----|
| Table 2-1. Datasets from experiments .....  | 23  |
| Table 3-1. 2 × 2 contingency table.....   | 26  |
| Table 3-2. Components of a Hidden Markov model .....  | 30  |
| Table 3-3 Tools used in this study .....  | 52  |
| Table 4-1. GO term analysis for seven sub-clusters .....  | 60  |
| Table 4-2. Enrichment analysis for target genes of Smad1 and Smad2/3.....   | 67  |
| Table 4-3. Enrichment analysis for target genes of Eomes .....  | 69  |
| Table 4-4. Enrichment analysis for target genes of T .....  | 72  |
| Table 4-5. Differential regions and differential dips between samples.....  | 76  |
| Table 4-6. Final list of TF-gene interactions.....  | 79  |
| Table 4-7. Overlap Smad targets with “d1-d2” paths (Figure 4.23) .....  | 94  |
| Table 4-8. Overlap Eomes and T targets with “d2-d3” paths (Figure 4.24).....  | 96  |
| <br>  |     |
| Supplementary Table 1. Percentages of aligned reads (10 samples of time-series RNA-seq transcriptome analysis) .....                              | 133 |
| Supplementary Table 2. Pearson’s correlation between 2 replicates.....  | 133 |
| Supplementary Table 3. Marker genes differentially expressed in the time course.....  | 134 |
| Supplementary Table 4. Percentages of aligned reads (ATAC-seq samples).....   | 135 |
| Supplementary Table 5. TF groups.....   | 135 |
| Supplementary Table 6. GO terms for each path of the gene regulatory tree constructed by using the time-series expression data (Figure 4.19)..... | 136 |

## F. Abbreviations

|                     |   |
|---------------------|---|
| A-P                 | Anterior-posterior  |
| APS                 | Anterior primitive streak   |
| ATAC-seq            | Assay for transposase-accessible chromatin using sequencing       |
| AVE                 | Anterior visceral endoderm  |
| BWT                 | Burrows-Wheeler Transform   |
| ChIP                | Chromatin immunoprecipitation                                     |
| ChIP-chip           | Chromatin immunoprecipitation combined with microarray            |
| ChIP-seq            | Chromatin immunoprecipitation followed by sequencing              |
| DE                  | Differentially expressed  |
| DNase-seq           | DNase I hypersensitive sites sequencing                           |
| DREM                | Dynamic regulatory events miner                                   |
| E6.0                | Embryonic state 6.0 days after fertilization                      |
| EM                  | Expectstion maximization  |
| EMT                 | Epithelial-mesenchymal transition                                 |
| EPI                 | Epiblast  |
| ES cells            | Embryonic stem cells  |
| ExE                 | Extra embryonic ectoderm  |
| FAIRE-seq           | Formaldehyde assisted isolation of regulatory elements            |
| FDR                 | False discovery rate  |
| FPKM                | Fragments per kilobase of transcript per million fragments mapped |
| GO                  | Gene ontology   |
| GRN                 | Gene regulatory network   |
| HMM                 | Hidden Markov model   |
| ICM                 | Inner cell mass   |
| IOHMM               | Input-output hidden Markov model                                  |
| KO                  | Knockout  |
| lncRNAs             | Long noncoding RNAs   |
| log <sub>2</sub> FC | log <sub>2</sub> fold change                                      |
| MEME                | Multiple Expectation Maximization Estimation                      |
| mESCs               | Mouse embryonic stem cells  |

|           |   |
|-----------|---|
| MLE       | Maximum likelihood estimation               |
| MNase-seq | Direct sequencing following MNase digestion |
| NGS       | Next-generation sequencing                  |
| P-D       | Proximal-distal                             |
| PCR       | Polymerase chain reaction                   |
| PE        | Primitive endoderm                          |
| Pol II    | RNA polymerase II                           |
| PS        | Primitive streak                            |
| R-Smads   | Receptor-mediated Smads                     |
| RNA-seq   | RNA sequencing                              |
| TE        | Trophectoderm                               |
| TES       | Transcription end site                      |
| TF        | Transcription factor                        |
| TSS       | Transcription start site                    |
| VE        | Visceral endoderm                           |
| WT        | Wild-type                                   |



## Zusammenfassung

Die Embryonalentwicklung ist ein komplexer mehrstufiger Vorgang, der auf der genetischen Ebene eine präzise Kontrolle durch Genregulationsnetzwerke (GRNs) erfordert. Während der Differenzierung von Vorläuferzellen in ihre Nachkommen aktivieren oder unterdrücken verschiedene Gruppen von Transkriptionsfaktoren (TFs) auf jeder Stufe der Musterbildung und der Organogenese ihre Zielgene um bestimmte Zellschicksale festzulegen. Eine Fehlregulation verschiedener Entwicklungsvorgänge kann zu schweren Krankheiten oder zum Tode führen, während deren ektopische Aktivierung im adulten Organismus die Ausbildung von Tumoren induzieren kann. Aus diesem Grund ist es von großer Bedeutung die entsprechenden Transkriptionsfaktoren zu entschlüsseln und herauszufinden, wie sie zum einen interagieren und zum anderen ein GRN bilden das die Entwicklungsprozesse kontrolliert.

Die Entstehung des Mesoderms ist bei der Embryonalentwicklung von großer Bedeutung. Sie findet während der Gastrulation statt und ist abhängig von der epithelial-mesenchymalen Transition (EMT). In Wirbeltieren entstehen aus dem Mesoderm verschiedene Gewebe: das axiale Skelett, die Skelettmuskulatur, das Herz, die Nieren, die Blutgefäße und das Blut. In einer Fülle von Studien wurde erläutert, welche Gene die Entstehung des Mesoderms beeinflussen. So ist bekannt, dass die WNT-, BMP- und FGF-Signalwege, zusammen mit TFs, vor allem Smads, Eomes und T, eine grundlegende Rolle bei diesen Vorgängen spielen. Allerdings gibt es bis jetzt noch keine umfassende und mechanistische Beschreibung des mesodermalen GRN.

Das Ziel dieser Arbeit ist es, ein globales Genregulationsnetzwerk zu erstellen, welches die transkriptionellen regulatorischen Ereignisse, die dynamisch während der Entstehung des Mesoderms in der Maus auftreten, zu beschreiben. Wir konnten nachweisen, dass die *in-vitro* Differenzierung von murinen embryonalen Stammzellen die Entstehung des Mesoderms *in-vivo* nachahmen kann. Aus diesem Grund verwenden wir die *in-vitro* Differenzierung als Modellsystem. Durch die kombinierte Anwendung von ChIP-Seq- und RNA-Seq-Techniken habe ich zuerst GRNs rekonstruiert, welche durch die für die Mesodermentwicklung wichtigen TFs Smads, Eomes und T gesteuert werden. Um ein globales Genregulationsnetzwerk, das die EMT und die Mesodermentwicklung steuert, zu erstellen, haben wir des weiteren Genexpression-Zeitreihen und Datensätze von Zielgenen bekannter TFs miteinander integriert. Letztere wurden durch einen originären Ansatz erzielt mit dem die funktional aktiven TFs aus ATAC-Seq-Daten ermittelt und mit ihren mutmaßlichen Zielgenen assoziiert wurden. Zusammen mit einem bioinformatischen Programm, das auf einem „hidden Markov-Modell“ basiert, konnte ich so Gruppen von koexprimierten Genen identifizieren und die TFs vorhersagen, welche deren Expression regulieren.

Wir konnten die Vorhersagekraft unseres Ansatzes bestätigen und beweisen, dass er die TFs ihren Zielen korrekt zuordnet, indem wir die Ergebnisse mit unseren Datensätzen von Smads, Eomes und T verglichen haben. Mittels dieses *de novo* Ansatzes haben wir sowohl neue Kandidaten für mesodermale TFs identifiziert als auch die sich dynamisch ändernden Gruppen von Zielgenen von schon bekannten TFs charakterisiert. Diese Arbeit erweitert unser Verständnis der der EMT und der Entstehung des Mesoderms zugrundeliegenden genregulatorischen Prozesse in der Maus und stellt eine Liste an neuen potentiellen Regulatoren des Mesoderms für deren zukünftige detaillierte Beschreibungen zur Verfügung. Dieser bioinformatische Ansatz ist daher ein vielversprechender Ansatz für zukünftige Studien, deren Ziel die Charakterisierung molekularer Mechanismen anderer wichtiger Entwicklungsprozesse ist.





## Publications

1. **Liu, J.**, Wang, X., Wang, H., Wei, G., Yan, J. (2014). Reconstruction of the gene regulatory network involved in the sonic hedgehog pathway with a potential role in early development of the mouse brain. *PLoS Comput Biol.* 10(10):e1003884. doi: 10.1371/journal.pcbi.1003884
2. Sudheer, S., **Liu, J.**, Marks, M., Koch, F., Anurin, A., Scholze, M., Senft, A. D., Wittler, L., Macura, K., Grote, P. and Herrmann, B. G. (2016). Different Concentrations of FGF Ligands, FGF2 or FGF8 Determine Distinct States of WNT-Induced Presomitic Mesoderm. *Stem Cells*, 34: 1790–1800. doi:10.1002/stem.2371
3. Lange, L., Marks, M., **Liu, J.**, Wittler, L., Bauer, H., Piehl, S., Bläß, G., Timmermann, B. and Herrmann, B. G. (2017). Patterning and gastrulation defects caused by the *tw18* lethal are due to loss of *Ppp2r1a*. *Biology Open*, 6(6), 752-764. doi:10.1242/bio.023200
4. Dynamic Gene Regulatory Network Controlling EMT and Mesoderm Formation (In preparation; Authors: Liu, J., Tsaytler, P., Koch, F., Bläß, G., Herrmann, B. G. *et al.*)



## **Selbstständigkeitserklärung**

Hiermit erkläre ich, dass ich diese Arbeit selbständig verfasst und keine anderen als die angegebenen Hilfsquellen und Quellen verwendet habe. Ich erkläre weiterhin, dass ich die vorliegende Arbeit oder deren Inhalt nicht in einem früheren Promotionsverfahren eingereicht habe.

Jinhua Liu

Berlin, 2019