Computational analysis of transcriptional responses to the Activin signal

DISSERTATION
zur Erlangung des akademischen Grades

Doctor rerum naturalium
(Dr. rer. nat.)
im Fach Biophysik

eingereicht an der
Lebenswissenschaftlichen Fakultät
der
Humboldt-Universität zu Berlin
Von

Dan Shi, M.Sc.

Präsidentin der Humboldt-Universität zu Berlin
Prof. Dr.-Ing. Dr. Sabine Kunst

Dekan der Lebenswissenschaftlichen Fakultät
der Humboldt-Universität zu Berlin
Prof. Dr. Bernhard Grimm

Gutachter/innen: Prof. Dr. Dr. h. c. Edda Klipp
Dr. Jana Wolf
Dr. Zhike Zi

Tag der mündlichen Prüfung: 25.08.2020

Abstract

Transforming growth factor-β (TGF-β) signaling pathways play a crucial role in cell proliferation, migration, and apoptosis through the activation of Smad proteins. Research has shown that the biological effects of TGF-β signaling pathway are highly cellular-context-dependent. In this thesis work, I aimed at understanding how TGF-β signaling can regulate target genes differently, how different dynamics of gene expressions are induced by TGF-β signal, and what is the role of Smad proteins in differing the profiles of target gene expression.

In this study, I focused on the transcriptional responses to the Nodal/Activin ligand, which is a member of the TGF-β superfamily and a key regulator of early embryonic development. Kinetic models were developed and calibrated with the time course data of RNA polymerase II (Pol II) and Smad2 chromatin binding profiles for the target genes. Using the Akaike information criterion (AIC) to evaluate different kinetic models, we discovered that Nodal/Activin signaling regulates target genes via different mechanisms. In the Nodal/Activin-Smad2 signaling pathway, Smad2 plays different regulatory roles on different target genes. We show how Smad2 participates in regulating the transcription or degradation rate of each target gene separately. Moreover, a series of features that can predict the transcription dynamics of target genes are selected by logistic regression.

The approach we present here provides quantitative relationships between transcription factor dynamics and transcriptional responses. This work also provides a general computational framework for studying the transcription regulations of other signaling pathways.

Zusammenfassung

Die Signalwege des transformierenden Wachstumsfaktors β (TGF-β) spielen eine entscheidende Rolle bei der Zellproliferation, -migration und -apoptose durch die Aktivierung von Smad-Proteinen. Untersuchungen haben gezeigt, dass die biologischen Wirkungen des TGF-β-Signalwegs stark vom Zellkontext abhängen. In dieser Arbeit ging es darum zu verstehen, wie TGF-β-Signale Zielgene unterschiedlich regulieren können, wie unterschiedliche Dynamiken der Genexpression durch TGF-β-Signale induziert werden und auf welche Weise Smad-Proteine zu unterschiedlichen Expressionsmustern von TGF- β-Zielgenen beitragen.

Der Fokus dieser Studie liegt auf den transkriptionsregulatorischen Effekten des Nodal / Activin-Liganden, der zur TGF-β-Superfamilie gehört und ein wichtiger Faktor in der frühen embryonalen Entwicklung ist. Um diese Effekte zu analysieren, habe ich kinetische Modelle entwickelt und mit den Zeitverlaufsdaten von RNA-Polymerase II (Pol II) und Smad2-Chromatin-Bindungsprofilen für die Zielgene kalibriert. Unter Verwendung des Akaike-Informationskriteriums (AIC) zur Bewertung verschiedener kinetischer Modelle stellten wir fest, dass der Nodal / Activin-Signalweg Zielgene über verschiedene Mechanismen reguliert. Im Nodal / Activin-Smad2-Signalweg spielt Smad2 für verschiedene Zielgene unterschiedliche regulatorische Rollen. Wir zeigen, wie Smad2 daran beteiligt ist, die Transkriptions- oder Abbaurate jedes Zielgens separat zu regulieren. Darüber hinaus werden eine Reihe von Merkmalen, die die Transkriptionsdynamik von Zielgenen vorhersagen können, durch logistische Regression ausgewählt.

Der hier vorgestellte Ansatz liefert quantitative Beziehungen zwischen der Dynamik des Transkriptionsfaktors und den Transkriptionsantworten. Diese Arbeit bietet auch einen allgemeinen mathematischen Rahmen für die Untersuchung der Transkriptionsregulation anderer Signalwege.

Acknowledgments

This paper is completed under the guidance of Prof. Dr. Edda Klipp and Dr. Zhike Zi. I am deeply impressed by the rigorous scientific attitude and good research spirit of my supervisors. They have given me a lot of helps and inspiration in the study, research, and life.

Thanks to the care and help of my colleagues in the lab who supported united and enthusiastic research environment for my studies.

I would like to give special thanks to my family and friends. Without them, I cannot concentrate on scientific research. Their care and help support me all the way.

This study was supported by a scholarship from the China Scholarship Council (CSC). Without the support of CSC, this work would be impossible to be done.

Finally, I would like to thank all the experts and professors for reviewing my graduation thesis carefully.

Table of Contents

List of Figures

List of Tables

CHAPTER 1

Background

1.1 Overview

In this chapter, I will introduce the background of biology and bioinformatics that are related to this thesis work. The chapter starts with some introduction of transcriptional regulation and TGF-β signaling, followed by a brief review on transcriptome studies, and a detailed explanation of high-throughput sequencing techniques. A brief description of the RNA-seq and ChIP-seq associated with high-throughput sequencing are given. Finally, an overview on computational modeling of transcription dynamics is provided.

1.2 Transcriptional regulation

1.2.1 The central dogma of molecular biology

Cell is recognized as the basic unit of life. Molecular biology studies the composition, structure and interactions of cellular molecules, including proteins and nucleic acids, that are important for the functions and maintenance of the cell (Alberts 2017). Deoxyribonucleic acid (DNA) is the genetic material of eukaryotic cells. It contains all the information pertaining to cellular activities. DNA strands are composed of four simple building blocks called nucleotides. The nucleotides are composed of a sugar called deoxyribose, a 5' phosphate group and one of four nucleobase: adenine (A), thymine (T), cytosine (C) and guanine (G). When forming a double strand, each nucleotide is connected to a specific partner premised on base-pairing rules (A to T and C to G). After discovering the double helical structure of DNA in 1953 (Watson and Crick 1953), Francis Crick published the first statement of the central dogma in 1958 (Crick 1958) and restated in a nature paper titled "*central dogma of molecular biology"* in 1970

(Crick 1970). The central dogma describes the flow between DNA, ribonucleic acid (RNA) and proteins (Figure 1).

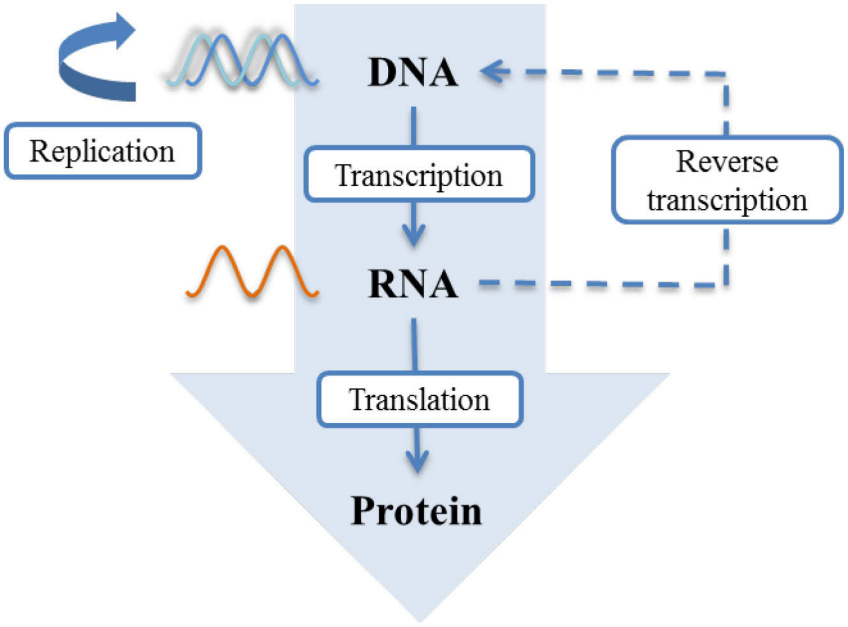


Figure 1 | Central dogma of molecular biology envisioned. This figure was modified from (Krebs, Goldstein et al. 2018).

The central dogma of molecular biology contains the four main processes of biological systems:

1) DNA is copied to produce two identical replicates in a process called **replication,** which is the primary stage in cell information maintains.
2) **Transcription** of DNA to RNA is the first step for gene expression. This RNA is a complementary and antiparallel copy of the original DNA.
3) RNA is **translated** via ribosomes into proteins. Following, the proteins perform their functions in further processes.
4) **Reverse transcription** uses RNA as a template to synthesis DNA. This process involves reverse transcriptase.

Some parts of the central dogma may not be entirely accurate, for instance, many viruses replicate their RNA genomes and transcribe RNA into messenger RNA (mRNA) (Ahlquist 2002). The central dogma shows the two key steps of gene expression – transcription and translation. Many kinds of RNA join these two steps. For example, the ribosomal RNA (rRNA) is located in the cytoplasm. It along with proteins forms ribosome that translates mRNA into proteins. Transfer RNA (tRNA), which links the amino acids and the mRNA based on three-

nucleotide codon of rRNA, is also involved in the synthesis of proteins. Additionally, long non-coding RNAs (lncRNA) perform significant functions in transcription regulation (Geisler and Coller 2013). Small interfering RNA (siRNA) and microRNA (miRNA) can mediate gene silencing (Kim, Villeneuve et al. 2006, Fabian, Sonenberg et al. 2010). Among these RNAs, mRNA - the connector between DNA and protein - is pivotal to the transformation of gene information.

1.2.2 Eukaryotic RNA production

DNA fragment that includes information about cellular functions - *gene* - is converted to RNA molecule - *transcript* - by transcription. Opening the chromatin controls the initiation level of transcription because eukaryotic genomes are tightly condensed into a chromatin structure. Chromatin remodeling decides whether a gene is expressed. It is principally carried out by two class coactivators, one of which is the ATP-dependent chromatin-remodeling complexes (Becker and Horz 2002) while the other is the histone-modifying complexes, which covalent histone modifications (Nakajima, Uchida et al. 1997). The acetylation of histone is a characteristic of the activation of gene expression, while the methylation of histone is associated with inactive chromatin.

Transcription begins when one or more basal transcription factors (TFs) bind to the DNA with RNA polymerase II (Pol II) at the promoter region of the gene. Then, the DNA is unwound to expose the single-strand which syntheses an initial RNA product. This step is called initiation. The following step is elongation. During this step, Pol II moves along the DNA to extend the RNA copy. In the termination step, both the nascent RNA and Pol II are released from the DNA template (Figure 2).

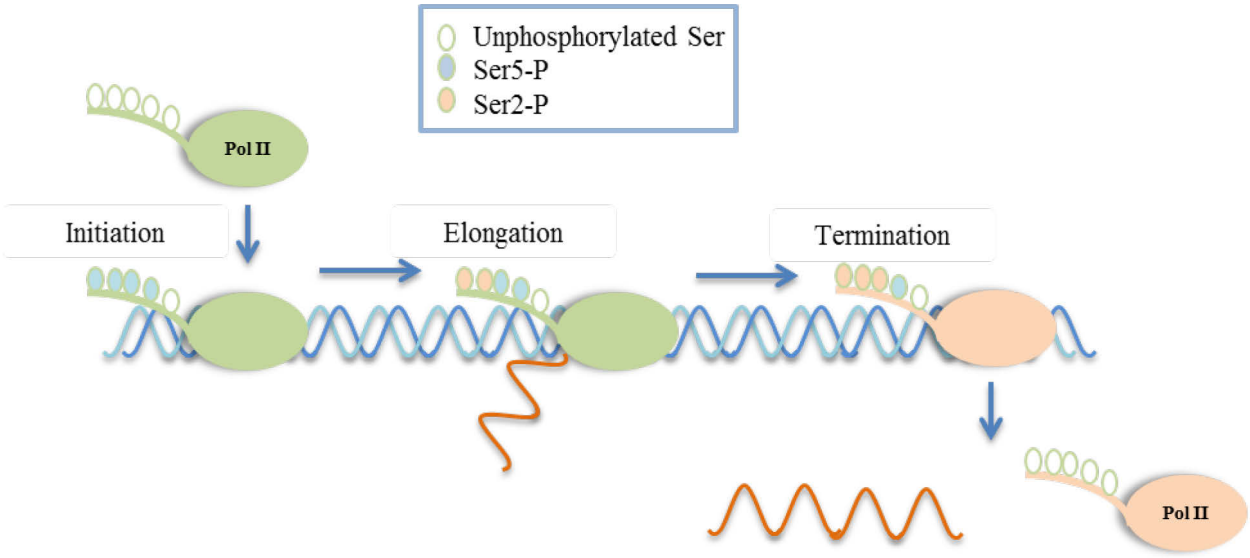


Figure 2 | Gene transcription is coordinated with Pol II CTD phosphorylation. Figure modified from (Kuehner, Pearson et al. 2011).

Gene transcription is mediated by Pol II, whose carboxy-terminal domain (CTD) phosphorylation status is associated with the production steps (Figure 2). Pol II CTD contains a repeat amino acid consensus sequence (Tyr1-Ser2-Pro3-Thr4-Ser5-Pro6-Ser7). It is phosphorylated by the transcription initiation factor II H (TFIIH) on Ser5 during initiation (Holstege, Fiedler et al. 1997). The phosphorylation of Ser2 by the positive transcription elongation factor (P-TEFb) activates the elongation process (Price 2000). Though the Pol II CTD can be phosphorylated by an unknown kinase on Ser7, its function is not clear yet.

In general, Ser5 phosphorylation is enriched at the 5' end of genes and associated with early transcription events (Komarnitsky, Cho et al. 2000). In contrast, Ser2 phosphorylation is enriched at the 3' end of genes and involved in 3' end processing (Ahn, Kim et al. 2004). The serine phosphorylation is regulated by many kinases such as the suppressor of Sua72 (Ssu72) and Fcp1, which can dephosphorylate Ser5 and Ser2 (Werner-Allen, Lee et al. 2011, Fuda, Buckley et al. 2012).

1.2.3 Post-transcriptional modification

Post-transcriptional modification, which is also called co-transcriptional modification, is a set of biological processes that modify an RNA primary transcript to produce a mature, functional RNA molecule after transcription (Kiss 2001). The nascent mRNA is altered in the

nucleus through three main processes: 5' capping, 3' polyadenylation and RNA splicing, which are described as below (Figure 3).

1) **5' capping**. A cap is formed by adding 7-methylguanosine (m7G) to the 5' end. This methylated cap is recognized by the cell's translational machinery (e.g. nuclear export proteins and ribosome) and can affect mRNA splicing, export, translation and stability (Jacobson and Peltz 1996).
2) **3' polyadenylation**. The sequence AAUAAA provides the signal for cleavage and for receiving a poly(A) tail at the 3' end. The major functions of the poly(A) tail are to protect the mRNA from degradation and facilitate exportation of mRNA from the nucleus (Drummond, Armstrong et al. 1985).
3) **RNA splicing**. Splicing occurs through breaking exon-intron junctions and joining exons' end. The sequence that is removed by splicing during mature mRNA forming is called intron. However, splicing can also result in different mRNA products when alternative splicing junctions are used. This is known as alternative splicing. Alternative splicing allows structural and functional variation of gene products. Following splicing, nascent mRNA becomes mature and functional. It will be transported from the nucleus to the cytoplasm and then translated into proteins.

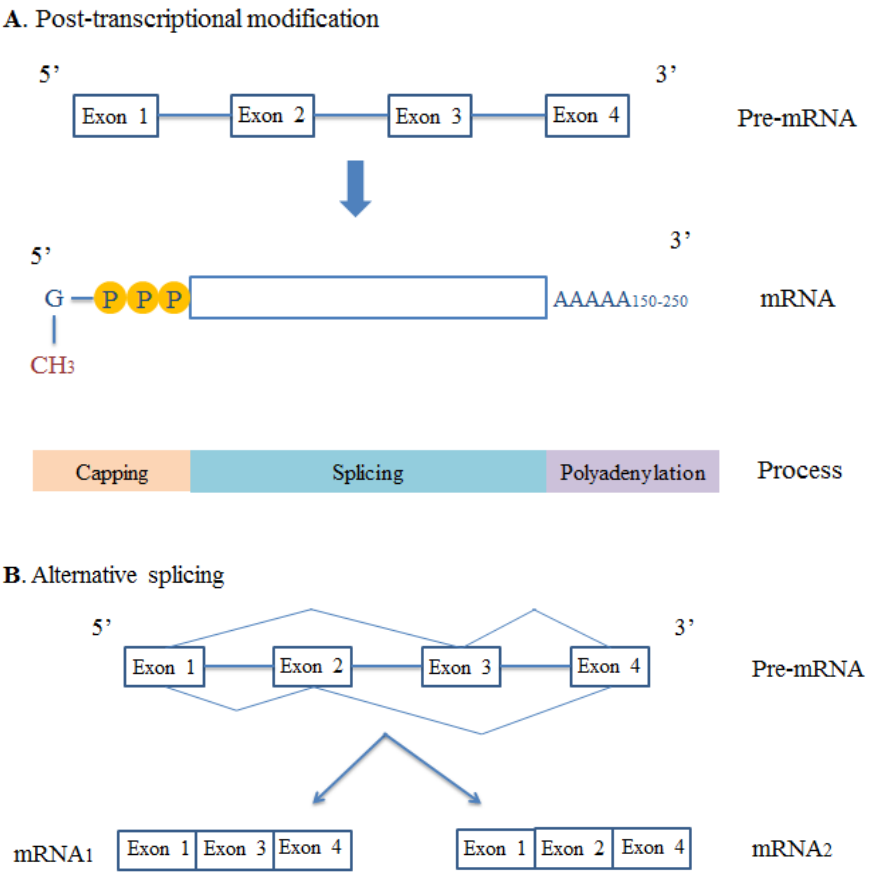


Figure 3 | Post-transcriptional modification and alternative splicing. (A) RNA is modified by adding m7G and poly(A) tail to the 5' and 3' ends in the nucleus. The introns are removed by splicing. **(B)** Different combinations of exons from a single pre-mRNA cause various mRNAs to be translated into different proteins. Thus, a single gene can encode multiple proteins.

1.2.4 Eukaryotic mRNA degradation

Most eukaryotic mRNA degradation is initiated as poly(A) tails are shorten by a poly(A) nuclease until it reaches a length of around 10A (Krebs, Goldstein et al. 2018). Then, the mRNA may be degraded either by the 5' to 3' pathway, or by the 3' to 5' pathway. The 5' to 3' pathway involves Dcp1/2 for decapping and Xrn1 for 5' to 3' exonuclease digestion. The 3' to 5' pathway involves the exosome complex for digestion. There are four other pathways for mRNA degradation in eukaryotic cells: deadenylation-independent decapping pathway, endonucleolytic pathway, histone mRNA pathway and miRNA pathway. Degradation of some specific mRNAs may be initiated by decapping. Few specific mRNAs can be cut by

endonuclease cleavage before 5' to 3' and 3' to 5' exonuclease digestion. Histone mRNA in mammals receive a short poly(U) tail before normal degradation. MicroRNA is able to target some RNA for silencing or degradation. Due to different degradation pathways, mRNAs exhibit a wide range of stabilities that contribute to differential mRNA abundance in a cell. Therefore, the spectrum of proteins made in a cell is also related to the stability of mRNAs.

1.2.5 Eukaryotic transcription regulation

Although the regulation of RNA processing may be taken in many steps during gene expression, eukaryotic gene expression is mainly controlled at the initiation level of transcription (Lemon and Tjian 2000, White 2009). Transcription initiation can be regulated by multiple proteins that bind to specific regulatory sequences and modulate Pol II activity. Different transcriptional regulatory proteins contribute to the regulation of various gene expressions in the different cell types. Moreover, chromatin remodeling and histone modification in the modification of chromatin structure also influence eukaryotic gene expression. In the following section, I will elaborate these transcriptional regulations.

1.2.5.1 *Cis*-regulatory modules

Cis-regulatory module (CRM) is a 100-1000 base pairs (bp) DNA that regulates transcription rates of target genes and their expressions by binding numerous TFs, such as enhancers, promoters, silencers, boundary elements and locus control regions (Davidson 2006, Jeziorska, Jordan et al. 2009) (Figure 4). Among these CRMs, the promoter leads to the initiation of transcription of a particular gene with the aid of the pre-initiation complex. Commonly, promoters are located upstream of the transcription start sites (TSS) of genes. Promoters lack universal structural features, while two functional parts are always present (Juven-Gershon and Kadonaga 2010). One functional part is the basal promoter (or core promoter) situated at about -35 bp to the TSS. The basal promoter provides a binding site for the transcription complex and localizes the transcription initiation site relative to the coding sequences (Reinberg, Orphanides et al. 1998, Lee and Young 2000). For many genes, it contains a TATA box, to which the TATA-binding protein (TBP) binds (Lifton, Goldberg et al. 1978). The other functional part of promoters is the binding site for a specific TF at approximately 250 bp or further upstream from the TSS. The TFs are necessary to activate or repress genes under various conditions (Lemon and Tjian 2000). Another CRM that regulates transcription is the enhancer. It can be located at a great distance (sometimes more than 10

kilobases) upstream or downstream from the TSS of regulated genes. Like promoters, the enhancers also function by binding TFs. Enhancer-bound TF can interact with the proteins at the promoter through DNA looping (Cooper, Hausman et al. 2000),

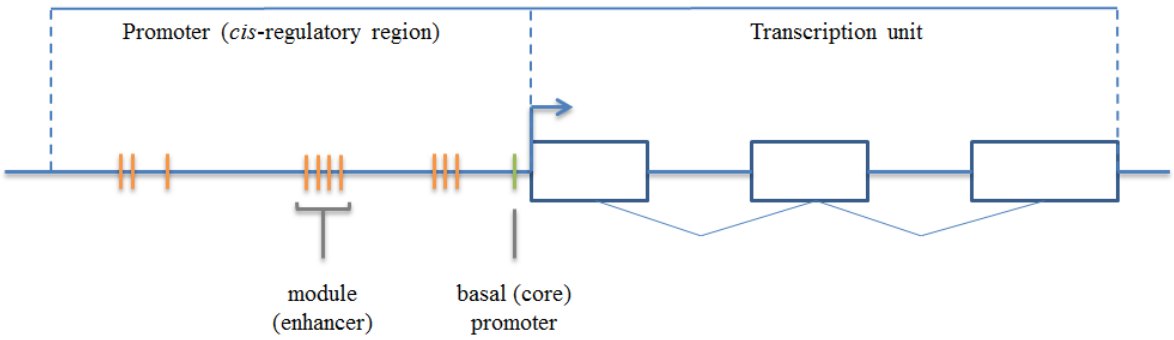


Figure 4 | Organization of a generalized eukaryotic gene. The transcription units, basal promoter and specific transcription factor binding sites are shown above. Figure modified from (Wray, Hahn et al. 2003).

1.2.5.2 Transcription factors

Transcription factors are proteins that bind to the specific DNA sequences and regulate the transcription rate of target genes (Figure 5). Basal factors, activators and coactivators are different types of transcription factors. The basal factors bind the start point in the promoter to the RNA polymerase. Activators work by making protein-protein interactions with the basal apparatus at the promoters or enhancers. Coactivators are needed by some activators to mediate the interaction. Activators control the frequency of transcription to make sure that genes are expressed correctly depending on the requirements of cellular environments. In human, Vaquerizas et al. estimated that roughly 6% of the expressed genes are TFs and the number of TFs expressed in each tissue is between 150 and 300 (Vaquerizas, Kummerfeld et al. 2009). Although only an average 6% of protein-coding genes in a tissue are TFs, they are still able to play different roles in regulating different genes as they are able to affect the transcription of multiple genes at the same time.

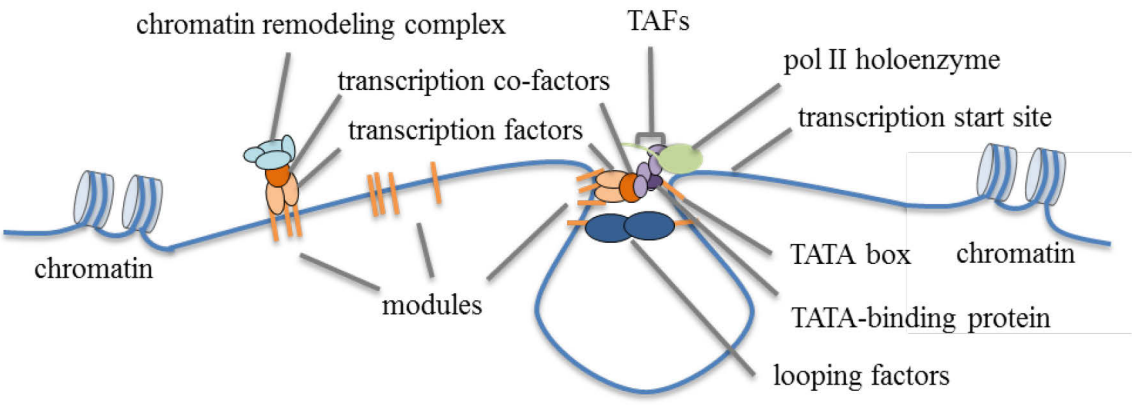


Figure 5 | Eukaryotic transcription factors. Several TATA-binding protein (TBP)-associated factors (TAFs, also known as general transcription factors) bind the basal promoter with RNA polymerase II holoenzyme complex and TBP. Transcription factors interact with transcription cofactors and chromatin remodeling complexes at the regulatory regions. Figure modified from (Wray, Hahn et al. 2003).

1.2.5.3 Chromatin structure

The DNA of all eukaryotic cells is compacted with the help of chromatin proteins such as histones H2A, H2B, H3, and H4. These positively charged proteins form complexes called nucleosomes by binding to negatively charged DNA. Nucleosomes can pack into a 30-nm chromatin fiber, which forms loops with a length of about 300 nm. The 300 nm fibers are further compressed and coiled into the chromatid of a chromosome (Pierce 2012). This packing of eukaryotic DNA has an important impact on gene expression. The general process of the dynamic modification of chromatin architecture is called chromatin remodeling. There are two major chromatin-remodeling complexes: covalent histone-modifying complexes and ATP-dependent chromatin remodeling complexes (Teif and Rippe 2009). When remodeling is carried out by covalent histone-modifying complexes, the modifications of histones match well with gene expression states. Histone acetylation is linked to transcription activation and deacetylation is linked to the repression of gene activity. However, the methylation of histones correlates with both active and inactive regions. The methylation of DNA is another feature linked to the chromatin structure and is a feature of inactive chromatin.

mRNAs are unstable molecules and are regulated in a number of different ways. Cells can respond to signals from their environment through changing gene expression and protein activity, both of which play crucial roles in regulating mRNAs. Understanding the interplay

between cell signaling and transcriptional regulation is an active area of research that remains to be explored and investigated.

1.3 The TGF-β and Nodal/Activin signaling pathways

1.3.1 The TGF-β superfamily

The transforming growth factor-β (TGF-β) superfamily plays a crucial role in controlling cell proliferation, migration, and apoptosis through the regulation of gene expressions (Wu and Hill 2009). In the early 1980s, the TGF-β family was isolated from many non-neoplastic tissues of the adult mouse and was discovered to induce a transformed phenotype in non-neoplastic cultured cells (Roberts, Anzano et al. 1981). Since then, several other TGF-β superfamily proteins have been identified, bringing the number up to more than 33 members, including TGF-βs themselves, bone morphogenetic proteins (BMPs), growth and differentiation factors (GDFs), Activins, and Nodal (Feng and Derynck 2005) (Figure 6). The TGF-β superfamily proteins are distinct, but they have similar structures and their regulated downstream components are well-conserved during evolution.

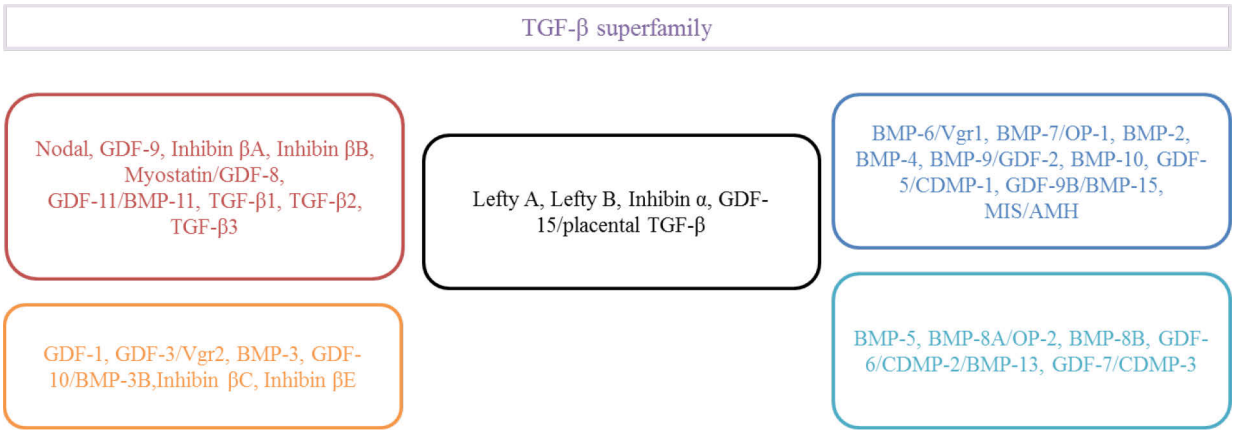


Figure 6 | The 33 TGF-β family polypeptides in human. Ligands that signal through R-Smads activated by Activin or TGF-β or R-Smads activated by BMP are shown in red or blue, respectively. Ligands that may signal through these two types of R-Smads, but whose receptors and pathways have not been fully identified, are shown in orange or light blue, respectively. BMP, Bone morphogenetic protein; OP, osteogenic protein; GDF, growth and differentiation factor; CDMP, cartilage-derived morphogenetic protein; MIS/AMH, Müllerian-inhibiting substance/anti-Müllerian hormone; TGF-β, transforming growth factor β. Figure modified from (Morikawa, Derynck et al. 2016).

The synthesis, secretion, and activation of polypeptides from the TGF-β superfamily consist of multiple complex processes and they are controlled by several proteins (Figure 7). TGF-β ligands are initially synthesized as pre-propeptides, which are called pre-pro-TGF-β. Then, the N-terminal signal peptide (SP) of the pre-pro-TGF-β is removed to produce a dimeric mature peptide (Derynck, Jarrett et al. 1985, Gentry and Nash 1990). After it is synthesized, the pro-TGF-β, which contains the C-terminal mature TGF-band and the N-terminal pro-domain TGF-β latency associated protein (LAP), is cleaved by the cleaving enzyme (furin, PACE) from its propeptide (Dubois, Blanchette et al. 2001, Kusakabe, Cheong et al. 2008). That creates a small latent TGF-β complex (SLC) which is connected with the latent TGF-β-binding proteins (LTBPs) to form the large latent complex (LLC) (Saharinen, Taipale et al. 1996). In the next step, the LLC is secreted from a cell and then processed to release active TGF-β (Annes, Munger et al. 2003). The key difference for distinguishing TGF-β superfamily members is the number and location of the cysteines. For example, three TGF-β isoforms contain nine cysteines while all other TGF-β superfamily members contain either seven or five cysteines (Morikawa, Derynck et al. 2016).

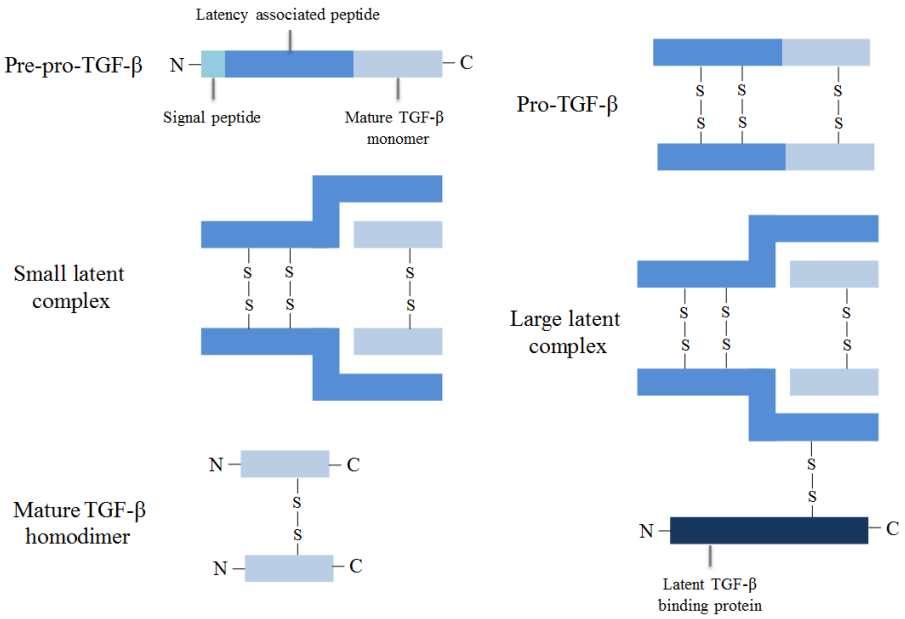


Figure 7 | A schematic representation of the different forms of TGF-β which occur during synthesis, secretion, and activation. Figure modified from (Poniatowski, Wojdasiewicz et al. 2015).

The biological effects of TGF-β superfamily are highly cellular-context dependent. For example, TGF-β1 enhances the endometrial decidualization (Kim, Park et al. 2005) and is related to embryo implantation (Manova, Paynton et al. 1992). BMPs are key players in the regulation of cell fate choices in developmental contexts (Bier and De Robertis 2015). These lead to a diversity cellular response during embryonic development.

1.3.2 TGF-β receptors

TGF-βs transduce their signals through a serine/threonine transmembrane receptor complex, forming a heterotetrameric combination. The receptors, which are known as TGF-β type I (TβRI, TGFBR1) and type II (TβRII, TGFBR2) receptors, have similar structures (Greenwald, Fischer et al. 1999). They can be distinguished from one another by peptide mapping. These receptors can be further categorized as seven TβRI, termed the Activin-like receptors (ALK1–7) and five TβRII receptors, termed TGFβR2, BMPR2, ACVR2A, ACVR2B and AMHR2 (Wrana, Attisano et al. 1992). Other than TβRI and TGFBR2, a type III (TβRIII, TGFBR3, betaglycan) receptor was also determined. It acts as a reservoir of ligand for TGF-beta receptors but lacks the domain for serine/threonine activity (Andres, Stanley et al. 1989).

Upon activation, TβRII is autophosphorylated in serine and threonine residues after binding with TGF-β (Wrana, Attisano et al. 1992). Then, it recruits TβRI and phosphorylates the serine and glycine rich domain (GS domain) of the TβRI. Two TβRI subunits and two TβRII subunits form a complex consisting of a TGF-β ligand and a receptor heterotetramer. Phosphorylation allows TβRI to propagate the signal to downstream substrates (Wrana, Attisano et al. 1994). Activated TβRI regulates the intracellular signaling by phosphorylating receptor-regulated Smads (R-Smads).

1.3.3 Smad proteins and canonical TGF-β signaling

The diverse effects of the TGF-β superfamily are mainly mediated through the so-called "mothers against decapentaplegic" or Smad proteins. The Smad family is conserved and has eight Smads members that can be classified into three categories (Huminiecki, Goldovsky et al. 2009) (Figure 8):

1) Receptor-activated Smads (R-Smads, Smad1, 2, 3, 5 and 8), which are regulated by TβRI and can be further separated into two subcategories:
    a. Smad2 and 3, which are phosphorylated by ALK4, 5, and 7. They are downstream of TGF-β and Activin signals.
    b. Smad1, 5 and 8, which are phosphorylated by ALK1, 2, 3, and 6. They are transduced by BMPs.
2) Common mediator Smads (Co-Smads, Smad4), which form heteromeric complexes with R-Smads to recruit co-regulators.
3) Inhibitory Smads (I-Smads, Smad6 and 7), which inhibit the activation of R-Smads.

R-Smads and Co-Smads have conserved Mad-homology 1 (MH1) and MH2 domains at their N-termini and C-termini, which are connected by a linker (Massague 1998). On the other hand, I-Smads only have conserved MH2 domains but lack MH1 domains, which are fundamental for DNA binding (Huminiecki, Goldovsky et al. 2009). MH1 region contains β-hairpin (βH) domain to mediate specific DNA binding of Smad3 and 4 (Shi, Wang et al. 1998). MH2 regions play crucial roles in the Smad–Smad protein interactions and transcriptional activations. The C-terminus SxS motif in MH2 of R-Smads is the target of TβRI (Abdollah, Macias-Silva et al. 1997, Feng and Derynck 2005). The L3 loop within MH2 domain indicates specificity in TβRI interaction (Lo, Chen et al. 1998). The Smad4 activation domain (SAD),

instead of PPXY motif in Smad4, is essential for mediating activation of Smads complex (De Caestecker, Yahata et al. 2000).

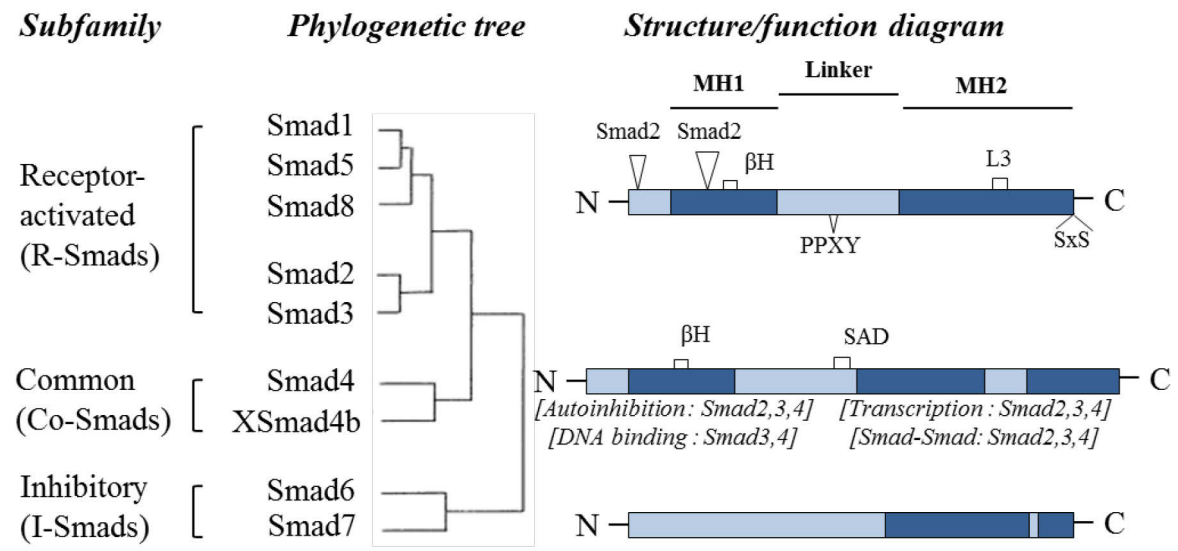


Figure 8 | The Smad family. Left-hand shows the phylogenetic tree of vertebrate Smads (except for Xenopus Smad4b [XSmad4b]). The schematic structure/function diagram illustrates receptor-activated Smads (R-Smad) (here is an example of Smad2, 3), the common-partner Smads (Co-Smad) (example shown here Smad4) and inhibitory Smads (I- Smads) (example here Smad7). βH, the β-hairpin domain. PPXY, the proline-tyrosine motif. L3, the L3 loop. SxS, the C-terminus SxS motif. SAD, the Smad4 activation domain. Triangles (Smad2) indicate extra exons in Smad2 compared with Smad3. This figure is modified from (Schiffer, von Gersdorff et al. 2000).

After phosphorylation of the C-terminal serines by TβRI, R-Smads are activated, then form complexes with a Co-Smad and translocate into the nucleus where they recruit sequence-specific TFs to regulate gene transcriptions (Miyazawa, Shinozaki et al. 2002) (Figure 9). In the cytoplasm, the expression of I-Smads and E3 ubiquitin ligases (Smurf1 and Smurf2) further regulates the activity of R-Smads. Because the induction of I-Smads is controlled by members of the TGF-β superfamily, an auto-inhibitory feedback mechanism is embedded in TGF-β signaling (Itoh, Itoh et al. 2000, Moustakas, Souchelnytskyi et al. 2001).

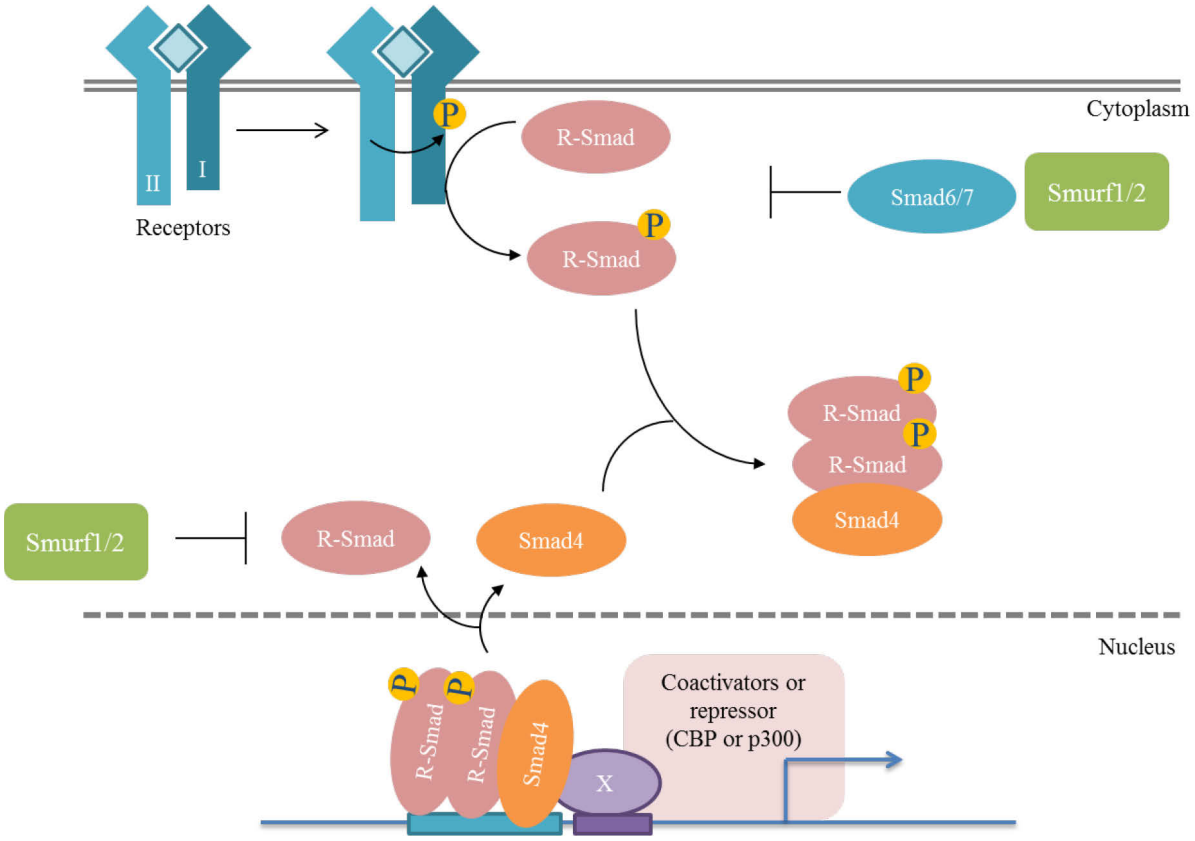


Figure 9 | A general mechanism of TGF-β receptor and Smad activation. At the cell surface, the ligand binds to the extracellular domains of TβRII, which recruits and phosphorylates TβRI. Activated TβRI then phosphorylates the R-Smads. The R-Smads form a complex with a Co-Smad and translocate into the nucleus, where they regulate transcription of target genes by interacting with DNA-binding factors. Smad6 or Smad7 can inhibit the activation of R-Smads. The E3 ubiquitin ligases (Smurf1 and Smurf2) mediate ubiquitination and degradation of R-Smads. Figure modified from (Derynck and Zhang 2003).

The Smad3 and Smad4 complexes bind at a special DNA sequence which is called Smad-binding element (SBE). It is a palindromic sequence of 8 bp (GTCTAGAC) (Zawel, Dai et al. 1998). In contrast to Smad3 and Smad4, Smad2 complexes do not bind directly to DNA because of an extra exon (exon 3) in the MH1 domain (Figure 8) (Yagi, Goto et al. 1999) and instead, require interaction with other TFs (Gaarenstroom and Hill 2014). It is still unclear how Smad2 complexes find their DNA binding site. Other than Smad-dependent TGF-β signaling (known as canonical TGF-β signaling), the TGF-β superfamily can modulate other signaling pathways without intermediate Smads.

1.3.4 Non-canonical TGF-β signaling

The TGF-β superfamily signaling not only trigger cellular signaling that are converted by the five R-Smads, but also induce non-canonical TGF-β signaling in a Smad-independent manner. Non-canonical TGF-β signaling pathways are activated directly by ligand-occupied receptors to modulate downstream cellular responses. For example, the mitogen-activated protein kinase (MAPK) pathways, including the extracellular signal-regulated kinases (ERK), c-Jun amino terminal kinase (JNK), p38 MAPK, as well as the IκB kinase (IKK), phosphatidylinositol-3 kinase (PI3K) and Akt, and Rho family GTPases, are well-known non-Smad signaling pathways (Zhang 2017). The TβRII interacts with the Daxx, a protein that actives JNK as well as programs cell death in epithelial cells and hepatocytes (Perlman, Schiemann et al. 2001). This Daxx-JNK pathway also involves homeodomain-interacting protein kinase 2 (HIPK2) and induces apoptosis in human p53-deficient hepatocellular carcinoma cells (Hofmann, Stollberg et al. 2003). In mouse mammary epithelial (NMuMG) cells, a mutant TβRI can induce the activation of p38 MAPK, which is required for TGF-β-induced apoptosis and epithelial-to-mesenchymal transition (EMT) (Yu, Hebert et al. 2002). The ERK was also found to be one of the key pathways of EMT induction (Zavadil, Bitzer et al. 2001). Another non-canonical pathway activated by TGF-β is PI3K/Akt signaling pathway, which also contributes to TGF-β induced EMT in epithelial cells (Bakin, Tomlinson et al. 2000) and induction of proliferation in mesenchymal cells (Wilkes, Mitchell et al. 2005). Finally, Ras homolog family member A (RhoA) and Cell division cycle 42 (Cdc42) of the Rho family of GTPases are involved in the TGF-β induced membrane ruffles and stress fibers formation (Edlund, Landstrom et al. 2002). In the past few years, many studies have so far focused on TGF-β signaling networks. However, it is not yet understood how to establish the balance between Smad and non-Smad signaling pathways.

1.3.5 The Nodal/Activin signaling pathway

1.3.5.1 Nodal and Activin

Nodal and Activin are two members of the TGF-β superfamily that are able to send signals via the heterotetrameric receptor complex. Nodal has been first reported in retroviral mutation mice (Robertson, Bradley et al. 1986), and encodes a basement-signaling molecule to induce mesoderm formation and axial structures in the early development of mice (Zhou, Sasaki et al.

1993). Activin has been discovered to be a gonadal protein of several species (Vale, Rivier et al. 1986). It enhances the secretion of follicle-stimulating hormone (FSH) and expresses in various cell types at almost all development stages.

In mice, human beings and birds, there is only one member in Nodal (Zhou, Sasaki et al. 1993). Other than that, there are three members in the zebrafish (Feldman, Gates et al. 1998, Rebagliati, Toyama et al. 1998, Sampath, Rubinstein et al. 1998) and five in *Xenopus* (Jones, Kuehn et al. 1995, Joseph and Melton 1997). Activin is a dimer composed of homodimers or heterodimers of Inhibin subunits (βa, βb, βc, βe). The different Inhibin subunits lead to the diversity of Activins. For example, two βa subunits form Activin A, βa and βb form Activin AB and two βb subunits form Activin B.

Nodal and Activin can bind TβRI and TβRII with other co-receptors (Table 1, Figure 10). The downstream effectors, Smad2/3 and Smad4, mediate signaling to the promoter region of target genes. The Smad proteins act as enhancers or repressors, leading to various cell type-dependent effects. Many mechanisms can regulate the activation of Nodal and Activin. For instance, Lefty1/2 act as competitive inhibitors of Nodal in zebrafish (Thisse and Thisse 1999), Smad7 can inhibit the activation of Smad2/3, while BMPs (BMP3, BMP7) interact with Nodal at the level of dimeric ligand production (Yeo and Whitman 2001). Therefore, Nodal/Activin signaling is regulated by many molecules in both extracellular and intracellular compartments.

Table 1 | Main components of the Nodal/Activin signaling pathway. Table modified from (Pauklin and Vallier 2015).

Signaling pathway	Pathway component	Gene name/symbol	Binding partners
Nodal	Ligands	Nodal (human, mouse, bird), cyclops,squint, southpaw (fish), xnr1, xnr2, xnr4, xnr5, xnr6 (frog)	Nodal pathway inhibitors
		Gdf1 (mouse)	
		Gdf3 (mouse)	
		Vg1 (frog, fish, bird)	
	Receptors	ALK4, ALK7	ActRII, ActIIB
			Co-receptors
	Co-receptors	Cripto (human), cryptic (mouse) (Cfc1-Mouse Genome Informatics), one-eyed oinhead (zebrafish), FRL-1/xCr1,xCR2,xCR3 (frog)	ALK4
	Inhibitors	Lefty1, Lefty2	ActRII
	Intracellular transduction proteins	Smad2	Smad3, Smad4
Activin	Ligands	Activin βa, βb, βc, βe (human)	Follistatin
	Receptors	ActRII,ActIIB	ALK4, ALK7
		ALK4	ActRII
		ActRII	ALK4
	Inhibitors	Cer1, Cer2, Gremlin	Nodal
		Follistatin	Activin
	Intracellular transduction proteins	Smad3	Smad2, Smad4
		Smad4	Smad2, Smad3
		Smad7	Smad2, Smad3

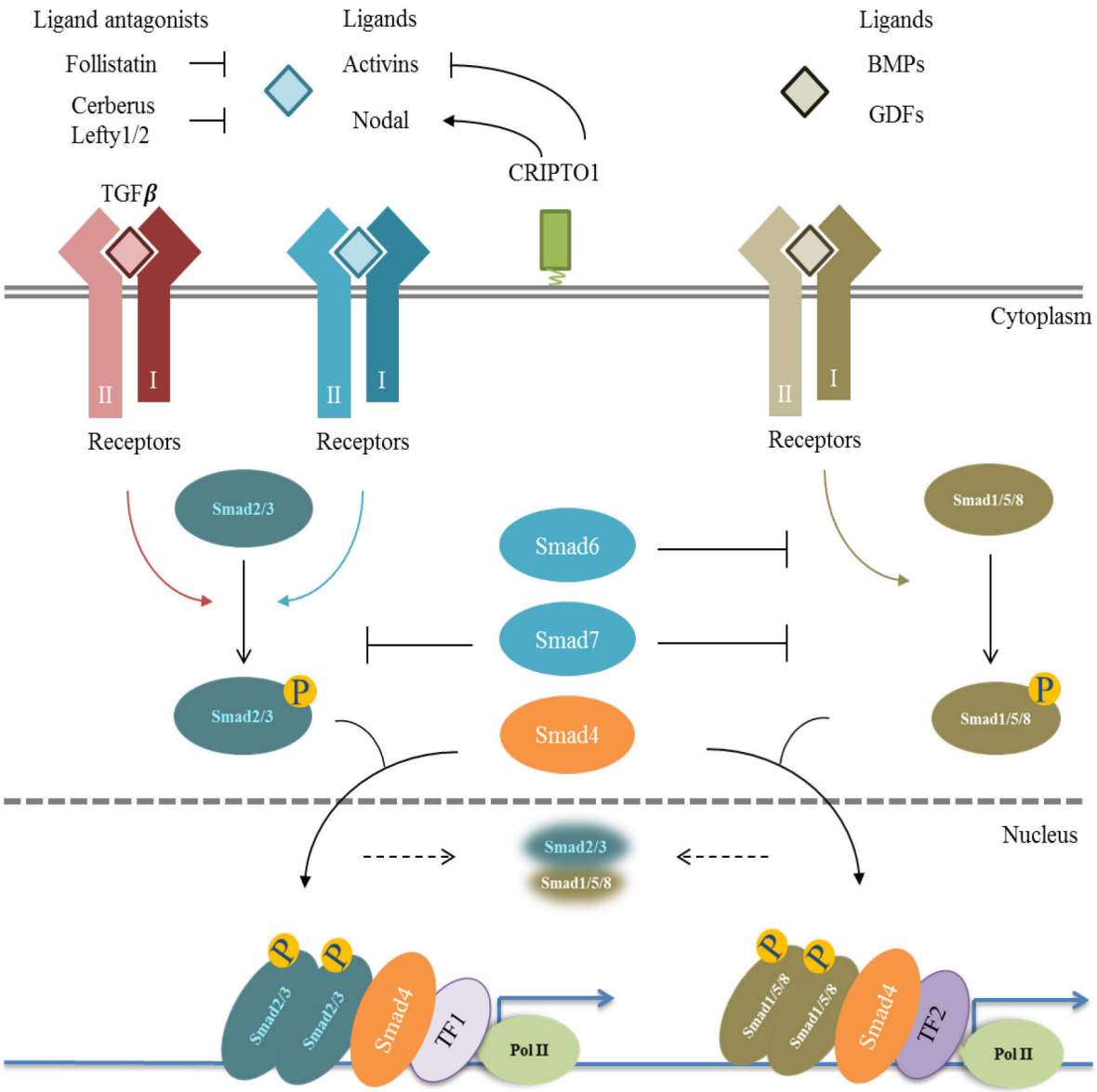


Figure 10 | Components of Nodal/Activin pathway. Extracellular ligand Nodal/Activin binds to TβRI (ACVRIIA/IIB) and TβRII (ALK4/7). Nodal requires additional binding of co-receptor CRIPTO1 to form an activated receptor complex with TβRI and TβRII. These activated receptor complexes phosphorylate Smad2/3. Smad2/3 forms a complex with Smad4 and enter the nucleus. Smad proteins target different genes by sequence-specific transcription factors, which are cell type dependent. Smad proteins can induce or inhibit the transcription of target genes and crosstalk with other signaling pathways in some cell types. Figure modified from (Pauklin and Vallier 2015).

1.3.5.2 Roles of Nodal/Activin in development and cancer

Nodal/Activin plays a crucial role in many development stages and cancers. For examples:

- ***Nodal/Activin signaling in mesoendoderm induction***: Nodal signaling regulates endoderm differentiation by interacting with WNT and BMP signaling (Tam and Loebel 2007). It has been shown that differential activation of the Nodal signaling pathway acts essentially in establishing the anteroposterior pattern in the organizer (Gritsman, Talbot et al. 2000). Activin, working through its transient precursors, also delay the induction of endoderm in mES cells (Parashurama, Nahmias et al. 2008).
- ***Nodal/Activin signaling in left-Right patterning*:** Nodal is the key morphogen in the regulation of the left-right axis specification (Brennan, Norris et al. 2002). Nodal produced in the node is required for the expression of left side-specific genes in the lateral plate mesoderm (Saijoh, Oki et al. 2003). Interestingly, the studies of Nodal signaling pathway and its downstream genes, such as Pitx, suggest that the regulation of Nodal in the development of the LR axis is conserved (Namigai, Kenny et al. 2014).
- ***Nodal/Activin signaling in neural patterning:*** In zebrafish, the fate of cells in the organizer is transformed from dorsal mesoderm to neural ectoderm without Nodal signaling (Schier and Talbot 2001). In the absence of Nodal signaling, neural differentiation occurs concurrently in mouse epiblast cells and wild-type embryo, which suggests the role of Nodal in the inhibition of neural fate determination (Camus, Perea-Gomez et al. 2006). Moreover, Activin provides telencephalic neural precursors with a caudal ganglionic eminence (CGE) identity. This function is conserved between the mouse and humans (Cambray, Arber et al. 2012).
- ***Nodal/Activin signaling and cancer:*** The mutation of multiple Nodal/Activin components leads to a cancerous effect, including TβRI, TβRII, Smad2 and Smad4 (Massague 2008). Nodal is expressed in different tumors with overexpressed co-receptor Cripto (Friess, Yamanaka et al. 1994, Kleeff, Ishiwata et al. 1998). It also increases the plasticity of tumor cells, which is important for tumor progression (Bodenstine, Chandler et al. 2016). Furthermore, an Activin α-inhibin acts as a tumor-suppressor gene of gonadal stromal cell proliferation (Matzuk, Finegold et al. 1992), which means that Activin can be tumorigenic if it gets out of control. Both Nodal and Activin positively regulate self-renewal of pancreatic cancer stem cells (Lonardo, Hermann et al. 2011).

1. 4 Next-generation sequencing technologies

1.4.1 Next-generation sequencing

Next-generation sequencing (NGS), also known as high-throughput sequencing (HTS), is a method that follows the first-generation sequencing technology Sanger method (Sanger, Nicklen et al. 1977). Compared with traditional methods such as PCR or Northern blot, NGS can output millions of reads in massively parallel sequencing. It also has the ability to produce an enormous volume of data cheaply. For example, it took 13 years and 3 billion dollars to accomplish the Human Genome Project by Sanger in 2003. Illumina, one of the NGS platforms, estimated the completion of 228,000 human genomes sequencing in 2014 and the price was as little as 1,000 dollars. Now, the NGS technologies apply to genome, transcriptome, DNA-protein interactions and epigenome characterization.

1.4.2 Overview of the Illumina sequencing method

There are many NGS platforms these days, for instance Roche/454, Illumina/Solexa, Life/APG and Helicos BioSciences. Among them, Illumina takes up around 75% of the sequencing applications. Although the different NGS platforms have unique aspects in each step, they adhere to a similar fundamental methodology. The work of the NGS platforms can be divided into three basic processes: library preparation, amplification and sequencing. Here, we go through the NGS method based on the Illumina platform (Figure 11).

i. **Template preparation.** The DNA sample is broken randomly into small fragments before the preparation of the template. Then, specialized adapters are added to both ends of the fragments.
ii. **Immobilization of strands on flow cell.** The single-stranded fragments with added adapters are randomly attached to the inside surface of the flow cell channels.
iii. **Bridge amplification.** After the loading of template fragments into the flow cell and the hybridization of the flow cell to the surface, the DNA fragment looks like a bridge. This is because both ends of the fragment are bound to the surface of the flow cell. Then, the unlabeled nucleotides and enzyme are added to the flow cell to initiate bridge amplification. Polymerases move along the DNA fragment and create its reverse strand.

iv. **Production and denaturing of double strands.** After bridge amplification, the double stranded DNA is denatured to leave single-stranded DNA separately anchored to the flow cell.

v. **Complete amplification.** The bridge amplification is repeated until the fragments are amplified into clonal clusters. Finally, several million clusters of DNA fragment are generated on the flow cell.

vi. **Laser excitation.** Before laser excitation, all of the reverse strands are washed out from the flow cell, leaving only the forward strands. When sequencing, four differently labeled fluorescent nucleotides are added into the flow cell one by one. The nucleotides are also labeled reversible terminators to avoid multiple additions. After adding, the unincorporated nucleotides are washed.

vii. **Signal image.** The emitted fluorescence is imaged and each point on the image refers to a cluster. Next, the fluorescent labels are cleaved and the 3'-OH group is regenerated for the next round. For paired-end sequencing, the clusters will be regenerated and the process of sequence will be repeated for the reverse strand.

viii. **Sequencing.** The laser excitation and signal imaging cycle is repeated to identify the sequence of bases in a fragment. The four colored images are translated into nucleotide, which is then exported into an output file. Besides the sequence of bases, the quality scores are also generated during a sequencing run. The quality score is a prediction of the probability of an error in case calling. It is used to assess the accuracy of the sequence and evaluate sequencing reads qualities for downstream analysis.

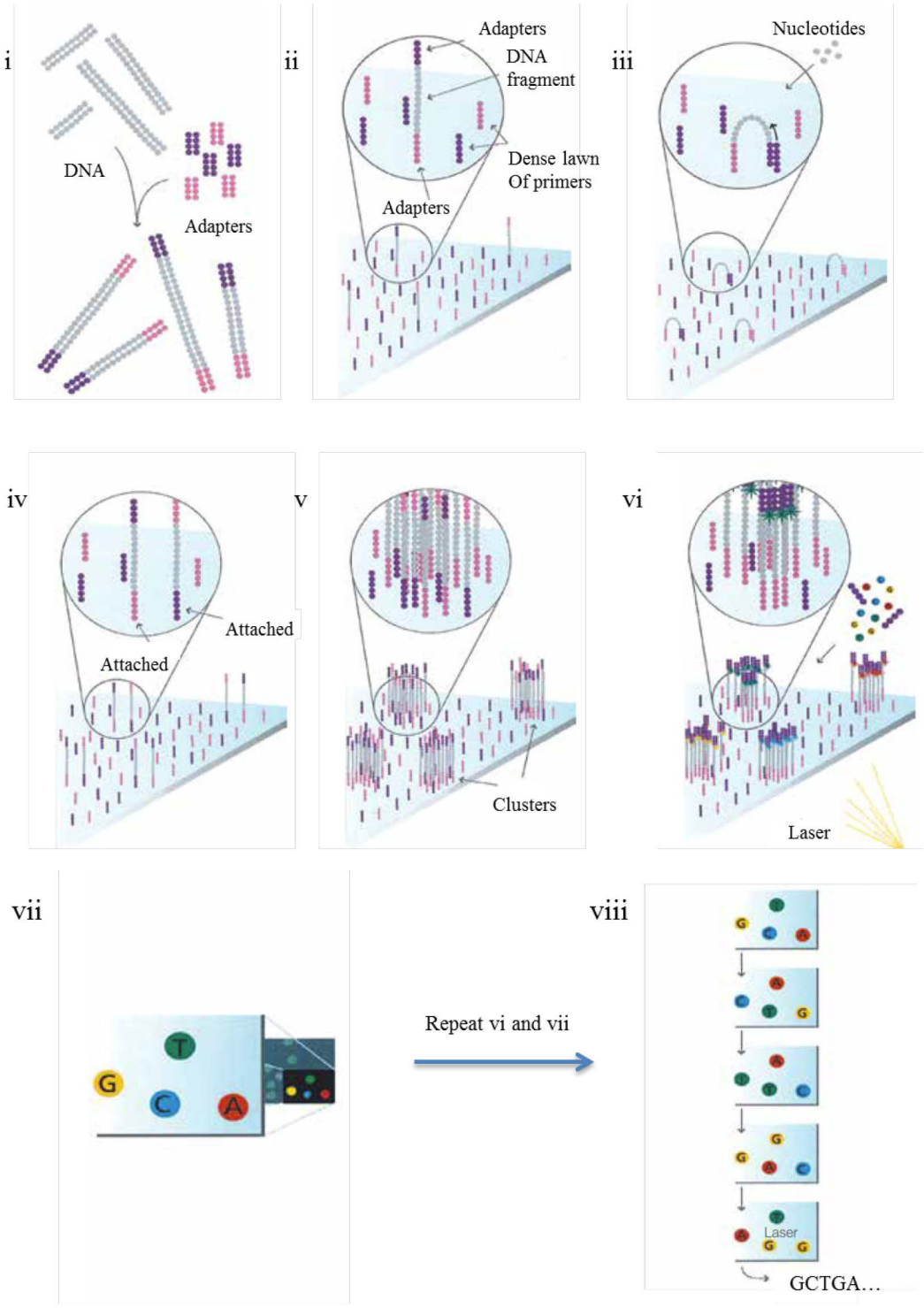


Figure 11 | Illumina sequencing workflow. i) Prepare genomic DNA sample; ii) Attach DNA to surface; iii) Bridge amplification; iv) Denature the double-stranded molecules; v) Complete amplification; vi) Determine base; vii) Signal image; vii) Sequencing. Figure modified from https://www.illumina.com.

The Life/APG sequencing platform, which is also known as SOLID, sequences by oligonucleotide ligation (Valouev, Ichikawa et al. 2008). The Roche/454 sequencing platform amplifies the DNA fragments using emulsion PCR on the surface of beads (Rothberg and Leamon 2008). Then, the beads are loaded onto the picotiter plate (PTP). Like Illumina, the Roche/454 sequences are also based on sequencing-by-synthesis (SBS) and PCR bridge amplification. The Illumina clusters, different from the Roche/454 beads that contain millions of copies, contain up to 1000 copies which are amplified from a single DNA fragment (Goodnow Jr 2014). All 1000 copies are sequenced together. The more laser excitation and signal imaging cycles we have, the more wrong signals we are likely to receive from the molecules. That is the main limitation of the Illumina sequencing length. The shorter read lengths of Illumina sequencing platform can be balanced by a higher throughput; 10 Gb - 1 Tb data can be produced by HiSeq 2500 system.

Now, there are many different NGS methods for different applications. For example, methylation sequencing can provide insight into methylation patterns; whole-genome sequencing (WGS) is used to analysis entire genomes; Hi-C is a method for analyzing chromatin interactions. The key NGS methods for the datasets used within this thesis will be described in detail.

1.4.3 RNA sequencing

RNA sequencing (RNA-seq), also known as whole transcriptome shotgun sequencing, generates relative measures of mRNA and individual exon abundance in a sample at a given time point by NGS (Morin, Bainbridge et al. 2008). When using RNA-seq, there are some additional steps in the template preparation. After isolating the RNA of the sample by deoxyribonuclease (DNase), the RNA would be selected by poly(A) tails to catch only mRNA, deplete rRNA, capture RNA targets or keep total RNA of the sample. Once the RNA is obtained, it will be reverse-transcribed into cDNA for the following steps (Chu and Corey 2012). Compared with previous technologies such as microarrays, RNA-seq has significant advantages. It can better estimate the absolute transcript levels (Fu, Fu et al. 2009). RNA-seq currently supports a wide range of applications, which include:

- Discovery of novel genes or splice junctions of expressed genes
- Transcript quantification and expression comparison across cell types

- Gene annotation and functional analysis

RNA-seq is not limited by prior knowledge as it can capture both known and novel features. In addition, RNA-seq can be applied to the species even if the reference sequencing is not available. However, RNA-seq still faces some challenges. In current sequencing platforms, the homopolymers, which play crucial roles in transcript metabolism, are hard to handle now (Hrdlickova, Toloue et al. 2017). The size limitations and sequencing sensitivity also need to be improved in the future.

1.4.4 ChIP-sequencing

Chromatin immunoprecipitation followed by sequencing (ChIP-seq) is used to analyze the interactions between protein and DNA (Figure 12). In ChIP experiment, DNA sequences bound by a particular protein in cells are enriched. The cells are treated with formaldehyde to crosslink DNA-binding proteins to DNA. Then, the chromatin is beaked into 200–600 bp fragments by sonication or endonuclease for DNA-binding proteins and micrococcal nuclease (MNase) for histone modifications (Furey 2012). Next, a specific antibody against the DNA-binding protein is added to immunoprecipitate DNA-protein fragments. Finally, the crosslinks are reversed and the DNA is purified by phenol-chloroform extraction and ethanol precipitation. After this, the DNA can be used in the template-preparation step of NGS (Schmidt, Wilson et al. 2009).

ChIP-seq can narrow the locations of the protein binding site to a range of a few tens of bp, and this unlimited dynamic range is one of its advantages. Because of its higher genome coverage, higher resolution and less noise, ChIP-seq is widely used in discovery of new regulatory elements, gene structure and the interactions between protein and DNA. ChIP-seq can capture DNA targets of specific TFs or histone modifications across the whole genome. Integrative analysis of ChIP–seq data, along with other data types like RNA–seq data, may clarify gene regulatory networks and the relationship between the transcriptome and TF binding. ChIP-seq is mainly limited by the alignability of reads to the genome. For this reason, the increasing read length can improve ChIP-seq coverage (Rozowsky, Euskirchen et al. 2009). Like any technology, ChIP-Seq has its artifacts. One of the biases is towards high GC content in fragment selection, which leads to false-positive peak calls (Teng and Irizarry 2017). In addition, the quality of the antibody and sample determine the value of ChIP-Seq. A sensitive

antibody will provide a high-level enrichment as compared to the background. For the quality of the sample, samples that are too small will bring about too few labels; while too many samples will cause the fluorescent labels to be too close to each other, resulting in lower data quality (Park 2009). In the future, not only experimental challenges that include antibodies selection need to be improved on, but also methods that can work with a small number of cells or signal cells are required.

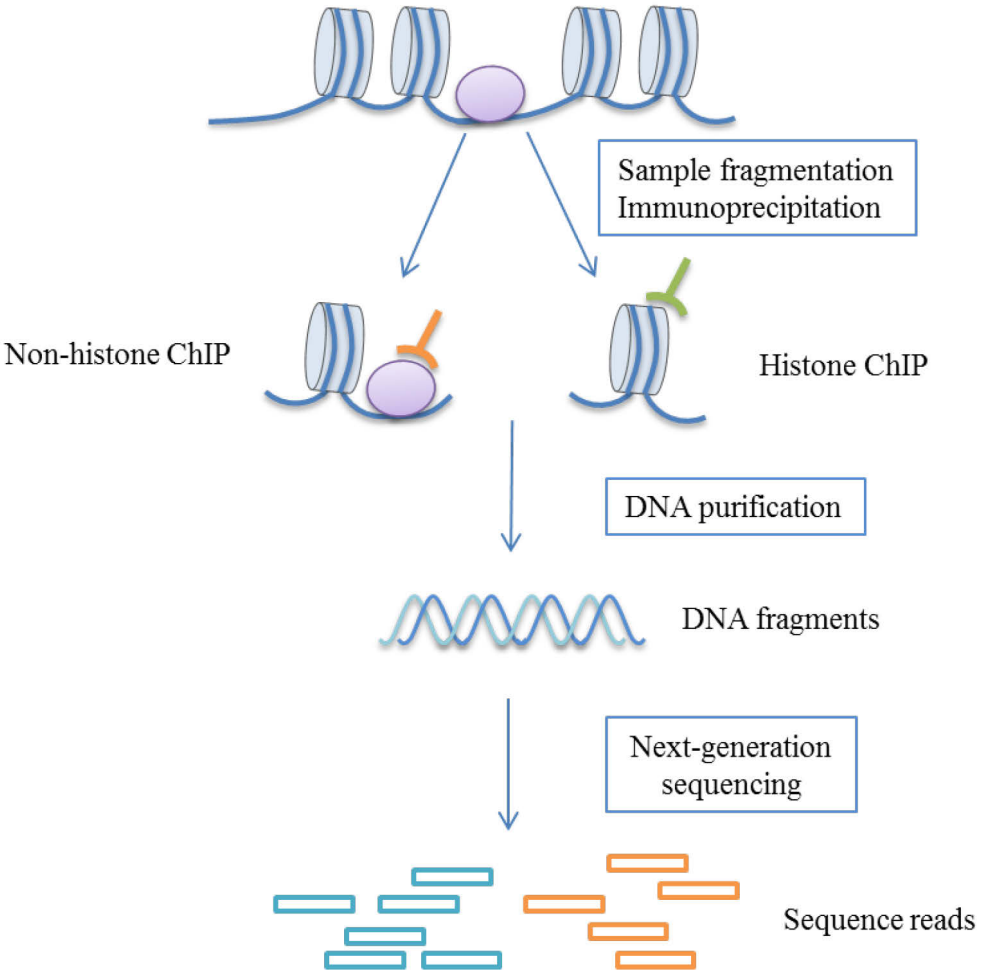


Figure 12 | Overview of a ChIP-seq experiment. The DNA targets for transcription factors or histone modifications are captured by chromatin immunoprecipitation (ChIP). After DNA purification, the ChIP-isolated DNA fragments are sequenced by next-generation sequencing to identify and quantify the sites bound by a protein of interest. Figure modified from (Park 2009).

1.4.5 Next-generation sequencing data

The output of NGS is stored as FASTQ files. FASTQ format usually has four different information lines (Figure 13). The first line which begins with '@' contains the sequence name. The second line is formed by a nucleotide sequence which contains A, T, C, G, N (low quality). The third line begins with '+' and is used to break the sequence and quality scores. The fourth line contains quality scores of the sequence as ASCII characters.

```
1  @SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
2  GGGTGATGGCCGCTGCCGATGGCGTCAAATCCCACC
3  +SRR001666.1 071112_SLXA-EAS1_s_7:5:1:817:345 length=36
4  IIIIIIIIIIIIIIIIIIIIIIIIIIIII9IG9IC
5  @SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
6  GTTCAGGGATACGACGTTTGTATTTTAAGAATCTGA
7  +SRR001666.2 071112_SLXA-EAS1_s_7:5:1:801:338 length=36
8  IIIIIIIIIIIIIIIIIIIIIIIIIIIII6IBI
```

Figure 13 | FASTQ file format example.

There are multiple FASTQ sequences per file, maybe millions. The length of the sequences can be different in each file. Because the output data is complex and huge, it is necessary to be careful when analyzing sequencing data.

1.4.6 Bioinformatics for next-generation sequencing

As shown above, NGS can be applied under various conditions to answer a variety of questions in different research fields. Because each read contains very little information, it is necessary to process NGS data using bioinformatics methods. The short reads of Illumina prove challenging for assembly software that are designed for Life/APG reads. New algorithms, programs and workflows that are specifically matched to short read sequence data are required. Open-source software accelerated the development of NGS data analysis (Pop and Salzberg 2008). Currently, there are many data analysis software for different purposes. For mapping short DNA sequencing data to an existing reference genome, we can use BWA (Li and Durbin

2009), Bowtie (Langmead, Trapnell et al. 2009), MAQ (Li, Ruan et al. 2008) and GSNAP (Wu and Nacu 2010). For mapping RNA sequencing data, TopHat2 (Kim, Pertea et al. 2013), STAR (Dobin, Davis et al. 2013), Subjunc (Liao, Smyth et al. 2013) and kallisto (Bray, Pimentel et al. 2016) are good choices. The sequencing data also can be de novo assembled using Velvet (Zerbino and Birney 2008), IDBA (Peng, Leung et al. 2012), Trinity (Grabherr, Haas et al. 2011) and ABySS (Simpson, Wong et al. 2009). These programs make it possible to quickly analyze large amounts of sequencing data in parallel. Software for identifying differentially expressed transcripts, such as RSEM (Li and Dewey 2011), edgeR (Robinson, McCarthy et al. 2010) and Cufflinks (Trapnell, Roberts et al. 2012), helps researchers to compare differences between samples.

Due to its advantages in quality, robustness and low noise, NGS is currently evolving into virtually every field of biological research, such as evaluation of genetic variations, RNA species distinction, epigenetic changes and so on. Through NGS, the gene expression dynamics in a signaling pathway can be quantified, which reflect the gene regulation results directly.

1.5 Computational modeling of transcription dynamics

The abundance of cellular RNA is determined by the regulation of RNA production, processing, and degradation. Additionally, it varies over time due to the response to external or internal stimuli. A dynamic model that can describe the change rate of continuous variables is needed to study the transcription dynamics. Non-linear ordinary differential equations (ODEs) have been widely used in this area. For example, Ciira et al. fitted transcription dynamics to Pol II occupancy time course data using a probabilistic model (wa Maina, Honkela et al. 2014). Antti et al. showed that a splicing-associated delay can play a key role in RNA production. Their models joined the data of transcriptional activation and mRNA production (Honkela, Peltonen et al. 2015). Michal et al. used newly transcribed RNA data to estimate temporally observed RNA processing and degradation rates. They proposed that the variable degradation rates between genes contribute to the observed differences in the dynamic response (Rabani, Levin et al. 2011). However, these models may not be suitable for the modeling of TGF-β superfamily induced gene expression due to the lack of TF dynamics.

Several researchers attempted to develop mathematical models to simulate the transcription process through TFs dynamics. Some studies prove that the binding of specific TFs can be used to predict gene expression *in vitro* (Kim and O'Shea 2008, Das, Dey et al. 2017, Choubey 2018). The TF titration effect has an important place in the expressions of TF-regulated gene pairs. Masayo Inoue and Mattias Rydenfelt et al. used it to predict the expression level of its regulated genes and relationship between TF–regulated gene pairs (Rydenfelt, Cox et al. 2014, Inoue and Horimoto 2017). All these models, which are based on predicted TF binding levels, may not reflect the TF activities of certain genes.

Some research works found TF occupancy, as detected by ChIP-seq, correlated with changes in Pol II occupancy. For example, compared with RNA-seq data, correlation between TF occupancy and Pol II occupancy is stronger (Mokry, Hatzis et al. 2012). Forkhead box O3 (FOXO3) acts as a transcriptional activator through Pol II recruitment (Eijkelenboom, Mokry et al. 2013). A straightforward mechanical relationship that explains how TF concentrations define target genes expression is needed.

CHAPTER 2

Aims & Objectives

Activin/Nodal signaling is required for the maintenance of pluripotency in human ESCs, while high level Activin/Nodal regulates mesendoderm induction and left-right patterning (James, Levine et al. 2005, Fei, Zhu et al. 2010, Pauklin and Vallier 2015). How Activin/Nodal induces different responses to different cell fates remains unknown. Mathematics models that combine experimental and theoretical data are effective methods to investigate this. For example, some signaling pathways not only encode information into proteins concentration or position but also through changes in the dynamic of these concentrations (Nelson, Ihekwaba et al. 2004, Kell 2005). In these cases, computer models can provide additional insights into the dynamics of the network. In my thesis, based on the time course data of Pol II and Smad2 chromatin bindings, we developed kinetic models to understand how different gene expression profiles are triggered by Activin signal.

The main objectives of the proposed project are the following:

- **To develop models for simulating Activin induced gene expression profiles based on Smad2 and Pol II binding dynamics.** Smad2 and Pol II time-course binding data was obtained by Coda et al. using ChIP-seq approach (Coda, Gaarenstroom et al. 2017). Different mathematical models were developed according to different hypotheses that describing the regulation of gene expression by Pol II, and Smad2.
- **To evaluate different models with the Akaike information criterion.** Model selection methods were employed to evaluate different models and rank the fitting results. These results uncovered the role of Smad2 in the transcriptional regulation of Activin signaling for each gene separately.
- **To identify the main features that can predict the types of transcriptional responses trigged by Activin.** Transcription dynamics can be associated with a variety of genetic

and epigenetic features. In this part, we set out to identify predictive features in transcriptional responses following Activin stimulation using logistic regression. This method can help us to develop a framework for linking gene features to Activin induced gene expression.

CHAPTER 3

Materials & Methods

3.1 Datasets

In this thesis, we used two datasets that were downloaded from the NCBI Gene Expression Omnibus (GEO) data repository (GSE77262, GSE77488) (Shum, Jones et al. 2016, Coda, Gaarenstroom et al. 2017). An overview of these datasets is shown in Table 2. The first dataset (GSE77262) is a collection of RNA sequencing data that provide a genome-wide half-life analysis of mRNAs in the P19 cell line. After transfected first with siControl and then with actinomycin D, the RNA was isolated and sequenced at 0, 15, 105, and 225 minutes, post-actinomycin D treatment. Another dataset used in the thesis are derived from the GSE77488 dataset provided by Coda et al. In this dataset, the samples were first pre-treated with Activin receptor inhibitor (SB-431542) overnight. and then treated either the SB-431542 inhibitor (SB treated sample) or Activin. The RNA-seq, Smad2 ChIP-seq and Input ChIP-seq samples were isolated at 1 and 8 hours, following the Activin stimulation. The Pol II Ser2, H3K27ac and H3K9ac ChIP-seq samples were isolated at 1, 4 and 8 hours, following the Activin stimulation. In this dataset, specially, the Smad3 seems undetectable from western blot. Hence, the ChIP-seq data focused on Smad2.

Table 2 | Overview of the datasets used in this study

Dataset	Strategy	Layout	Sequence length	Instrument
GSE77262 Half-life analysis of mRNAs in mouse P19 cell line	RNA-seq	SINGLE	100 bp	Illumina HiSeq 2500
GSE77488 Time-course data following Activin stimulation or inhibition in mouse P19 cell line	RNA-seq	SINGLE	101 bp	
	Pol II Ser2 ChIP-seq	SINGLE	51bp	
	Smad2 ChIP-seq	SINGLE	51bp	
	Smad2 Input ChIP-seq	SINGLE	51bp	
	H3K27ac ChIP-seq	SINGLE	51bp	
	H3K9ac ChIP-seq	SINGLE	51bp	

3.2 Sequencing data analysis

3.2.1 Data processing overview

The main procedure of sequencing data processing include the following steps: (i) quality control to remove adapter sequences and low-quality tails, (ii) sequence alignment to the genome, (iii) transcript abundance quantification, in which the number of aligned reads that overlap each gene or special region in the annotations are counted, (iv) data normalization for further analysis. Detailed explanations of these steps are provided in the following sections.

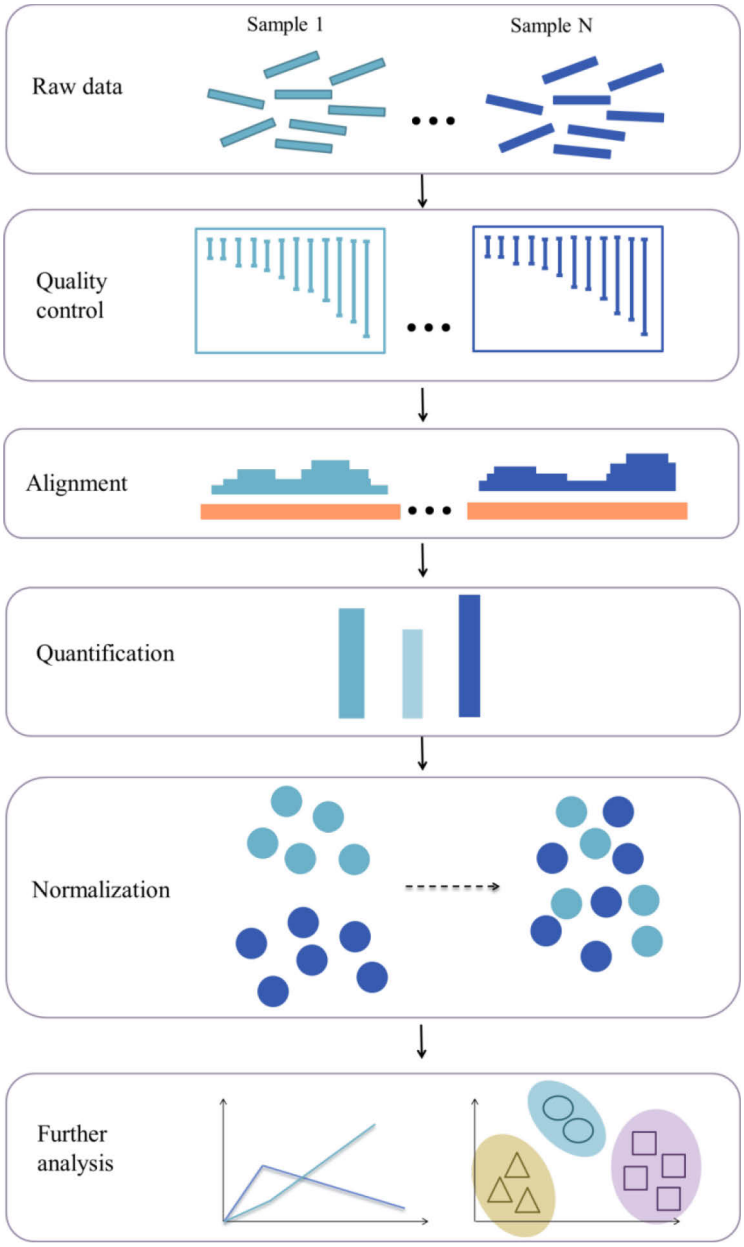


Figure 14 | Workflow for the processing of RNA sequencing data. The workflow begins with the quality control for the raw data. Then, the reads are mapped to the genome in order to obtain the number of aligned reads. Finally, the reads number per sample is normalized before downstream analysis.

3.2.2 RNA-seq data processing

3.2.2.1 Quality control

NGS can perform massively parallel sequencing. However, because of its complex structure and multiple processing steps, a careful analysis of NGS data is needed. The results can suffer

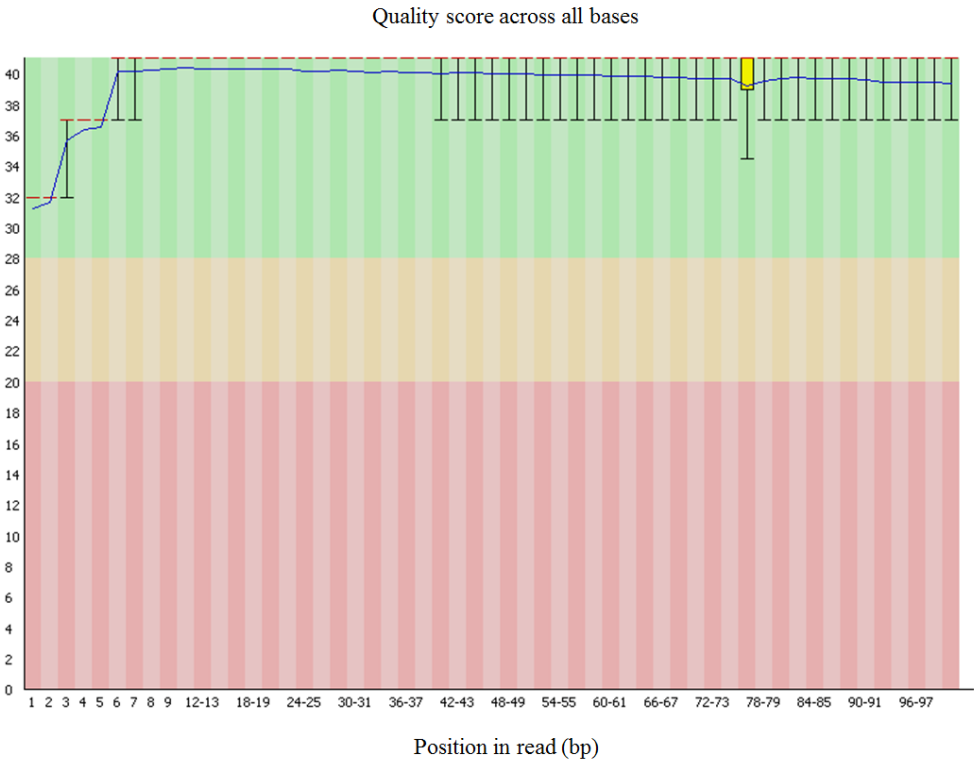
from a number of issues, such as poor sample quality, specific sequencing biases and inconsistent library preparation. In addition, the conditions of library construction could influence the sequencing bias (Ross, Russ et al. 2013). Therefore, it is necessary to control the quality of the NGS data from the beginning of the data analysis. There are many tools available for the quality control of NGS data. In this study, we used FastQC, which provides various statistics, including per base quality, GC content, sequence length, sequence duplication level and adapter types (http://www.bioinformatics.babraham.ac.uk/projects/fastqc). The users can get a quick impression of the data conditions and the warnings that are reported based on its default values. One of the most important statistics is per base quality, which provides the Phred quality scores Q for each read position. Phred scores Q are defined as a property that is logarithmically related to the base calling error probabilities P (Ewing and Green 1998).

$$Q = -10\, log_{10}\, P$$

For example, if Phred assigns a Q score of 20 (Q20) to a base, this is equal to the probability of an incorrect base call 1 in 100 times. This means that the base call accuracy is 99% and every 100 bp sequencing read may contain an error.

All of the RNA-seq data we used were first analyzed by FastQC (Figure 15). The half-life RNA-seq data have good qualities, while the time-course RNA-seq data from GSE77488 have many low-quality reads. The reads that have Phred scores lower than 20 were removed before further analysis.

A. The half-life RNA-seq data



B. The time-course RNA-seq data

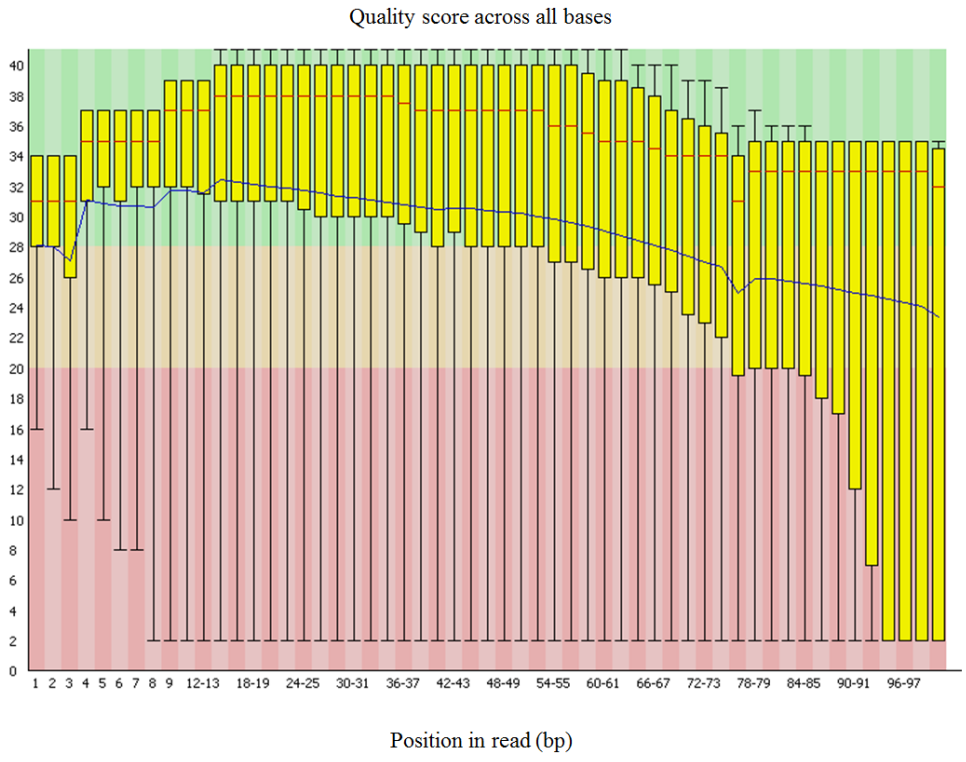


Figure 15 | Per base quality plot of FastQC. The median (red) and mean (blue) Phred scores are shown on the plots. A. The half-life RNA-seq data (SRR3126197). B. The time-course RNA-seq data (SRR3138964).

3.2.2.2 Alignment and quantification

After quality control, all RNA-seq reads need to be aligned to the genome. This alignment arranges the sequences to identify the location from which the reads originated (Garber, Grabherr et al. 2011). Because of the introns, the alignment of the RNA-seq reads needs to be distinguished from the intron-spanning reads, which spams at least one intron. The alignment is a time-consuming step during mapping. Here, we aligned RNA-seq reads and the mouse reference genome (GRCM38) using the HISAT2 (Kim, Langmead et al. 2015) aligner. The HISAT2 divides these exon-spanning reads into four categories: "(1) long-anchored reads, which exhibit at least 16 bp in each of the two exons; (2) intermediate-anchored reads, which exhibit 8–15 bp in one exon; (3) short-anchored reads, with just 1–7 bp aligned to one of the exons; and (4) the reads spanning more than two exons" (Figure 16A). In a simulated human RNA-seq data set, about 25.1% of the reads span two exons with more than 15bp anchors in both exons and about 12.4% of the reads span at least two exons with intermediate or short anchors, which are hard to be mapped to a unique location in the genome (Figure 16B).

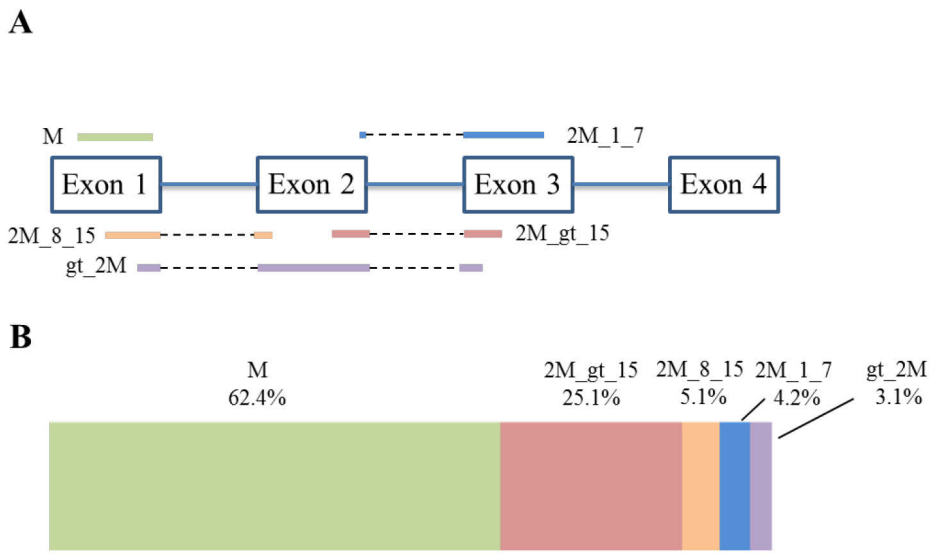


Figure 16 | HISAT RNA-seq reads types and their relative proportions. A. Five types of RNA-seq reads: i) M, exonic read; ii) 2M_gt_15, junction reads with long; iii) 2M_8_15, junction reads with intermediate; iv) 2M_1_7, junction reads with short; and v) gt_2M, junction reads spanning more than two exons. B. Relative proportions of different types of reads in the 20 million 100-bp simulated read data. Figure modified from (Kim, Langmead et al. 2015).

HISAT2 solves the challenging spliced-alignment problems using hierarchical indexing, which is based on the Burrows-Wheeler transform (BWT) (Burrows and Wheeler 1994) and the Ferragina-Manzini (FM) index (Ferragina and Manzini 2000). BWT is a reversible permutation of the characters in a text (Figure 17). It uses a matrix to re-order the reference sequence and the matrix has a property called 'last first (LF) mapping', which means the ith occurrence of character X in the last column corresponds to the same text character as the ith occurrence of X in the first column (Langmead, Trapnell et al. 2009). HISAT2 applies a whole-genome FM index to anchor each alignment and a number of local FM indexes in order to rapidly extend these alignments. As a result, HISAT2 is a highly accurate and efficient system for sequencing data alignment.

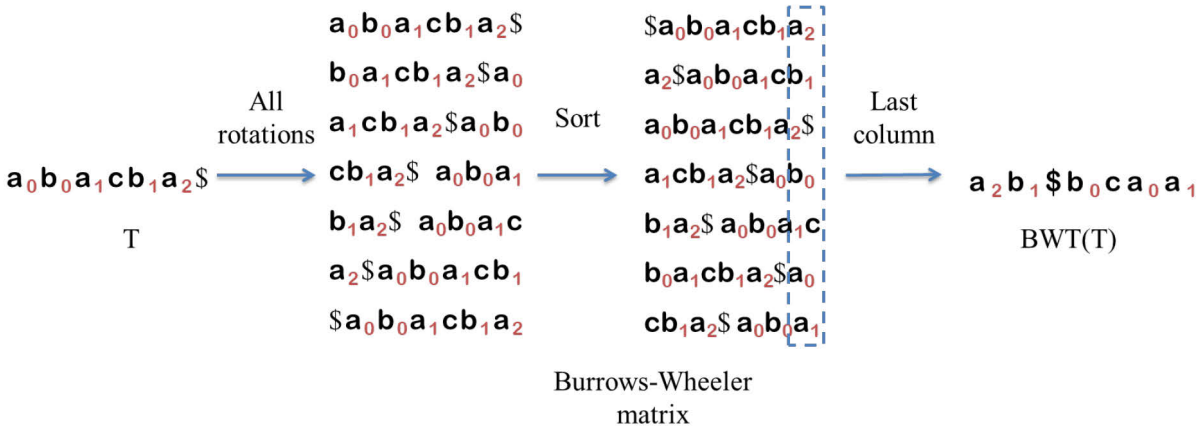


Figure 17 | The Burrows-Wheeler transform. String T is looped to generate seven strings, which are then sorted in lexicographic order. After sorting, the last column gives the BWT string.

All of the 101 bp RNA-seq reads were aligned to the HISAT2 GRCm38 index that was downloaded from the HISAT website. This index contains the Hierarchical Graph FM index for GRCm38. After alignment, the mapped reads were counted by ReadCounter (http://www.genefriends.org/ReadCounter) (Table 3). Only the reads mapped uniquely to exonic regions of one gene were counted, the ambiguous reads were not used in further analysis. The mean RNA-seq gene alignment rate was 25.0 million reads for half-life of mRNA data (dataset 1) and 34.1 million reads for time-course RNA-Seq data in dataset 2.

Table 3 | Mapping results of RNA-seq samples.

Dataset	SRR number	Time point	Total number of reads	Number of reads uniquely mapping to exonic regions of one gene
GSE77262 half-life analysis of mRNAs in mouse P19 cell line	SRR3126196	0min_1	19,296,773	16,158,126 (83.73%)
	SRR3126197	0min_2	26,666,632	23,070,791 (86.52%)
	SRR3126200	15min_1	27,258,464	22,749,449 (83.46%)
	SRR3126201	15min_2	20,580,569	17,750,958 (86.25%)
	SRR3126204	105min_1	19,437,322	17,325,594 (89.14%)
	SRR3126205	105min_2	27,370,733	24,931,443 (91.09%)
	SRR3126208	225min_1	24,341,837	22,221,103 (91.29%)
	SRR3126209	225min_2	22,785,314	20,486,015 (89.91%)
GSE77488 Time-course data following Activin stimulation or inhibition in mouse P19 cell line	SRR3138964	SB-431542_1	28,894,091	22,506,012 (77.89%)
	SRR3138965	SB-431542_2	33,754,546	26,236,341 (77.73%)
	SRR3138966	1h Activin_1	52,349,182	40,329,813 (77.04%)
	SRR3138967	1h Activin_2	55,390,050	42,891,396 (77.44%)
	SRR3138968	8h Activin_1	52,536,698	40,980,998 (78.00%)
	SRR3138969	8h Activin_2	40,600,328	31,681,450 (78.03%)

3.2.2.3 Data normalization

After the alignment of reads, it is possible to detect the expression of a certain RNA based on the count of reads. In RNA-seq, it has been found that the reads count is linearly related to the abundance of the target transcript (Mortazavi, Williams et al. 2008). However, due to different sample sizes or sequencing libraries, we need to normalize the number of aligned reads before analysis. Mapped reads of mRNA half-life samples were normalized by the reads count of glyceraldehyde-3-phosphate dehydrogenase (GAPDH) gene, which is a stable transcript, and used as a control in this analysis. Mapped reads of time-course RNA-seq samples were normalized by the DESeq package from Bioconductor (Anders and Huber 2010). The DESeq package is based on the negative binomial distribution assumption, which assumes that most genes are not differentially expressed and these genes should have similar read counts among different samples. The DESeq package compares and scales all reads count of the samples to validate the hypothesis (Dillies, Rau et al. 2013). This approach helps to identify a small number of differentially expressed genes which have a strong effect on the total reads count.

To determine the differentially-regulated genes, we converted the Activin stimulation data into fold change (FC) by normalizing it against the SB-431542 treated sample. To identify the differentially-regulated genes, we set the false discovery rate (FDR) threshold with a value of 0.05 and a fold change cutoff of 2 Furthermore, genes with lower than 30 reads across all samples were discarded, as they appear as background noise. In total, 198 differentially-regulated genes were identified. The corresponding mRNA half-life data were normalized to the value of the 0min sample and converted into FC values.

3.2.3 ChIP-seq data processing

3.2.3.1 Quality control

All of the ChIP-seq reads were quantified by FastQC, and none of them has a low per-sequence quality. Therefore, the ChIP-seq data have a high quality for downstream analysis.

3.2.3.2 Alignment

Because of the short sequencing length, the ChIP-seq reads and the mouse reference genome (GRCM38) were aligned using Bowtie (Langmead, Trapnell et al. 2009) (Table 4). Bowtie

applies the BWT and the FM index to map the NGS data. During mapping, the maximum mismatch was set to 2 for a seed length of 51 bp. Uniquely mapped reads were used for further analysis.

Table 4 | Mapping results of ChIP-seq samples.

Strategy	SRR number	Time point	Total number of reads	Deduplication unique mapped reads
Pol II Ser2 ChIP-seq	SRR3096935	SB-431542_1	36,330,127	22,351,909 (64.19%)
	SRR3096936	SB-431542_2	68,651,414	42,259,967 (64.23%)
	SRR3096937	1h Activin_1	61,449,305	39,808,852 (66.54%)
	SRR3096938	1h Activin_2	43,241,380	28,063,594 (66.60%)
	SRR3096939	4h Activin_1	57,786,670	36,941,071 (66.31%)
	SRR3096940	4h Activin_2	52,770,475	33,416,122 (65.12%)
	SRR3096941	8h Activin_1	63,973,606	40,229,252 (65.44%)
	SRR3096942	8h Activin_2	45,714,835	28,741,969 (65.29%)
Smad2 Input ChIP-seq	SRR3138984	SB-431542_1	79,719,094	53,319,157 (66.88%)
	SRR3138985	1h Activin_1	64,969,548	44,115,539 (67.90%)
	SRR3138986	8h Activin_1	61,611,461	41,516,897 (67.39%)
Smad2 ChIP-seq	SRR3138989	SB-431542_1	61,905,267	42,508,561 (68.67%)
	SRR3138990	1h Activin_1	78,813,534	41,913,149 (53.18%)
	SRR3138991	8h Activin_1	64,295,840	36,933,125 (57.44%)

Strategy	SRR number	Time point	Total number of reads	Deduplication unique mapped reads
H3K27ac ChIP-seq	SRR3096955	SB-431542_1	62,580,321	43,653,788 (69.76%)
	SRR3096956	SB-431542_2	70,930,705	49,778,672 (70.18%)
	SRR3096957	1h Activin_1	59,096,422	41,475,768 (70.18%)
	SRR3096958	1h Activin_2	46,577,906	33,978,972 (72.95%)
	SRR3096959	4h Activin_1	83,127,468	57,180,028 (68.79%)
	SRR3096960	4h Activin_2	64,634,339	45,690,199 (70.69%)
	SRR3096961	8h Activin_1	48,272,522	34,050,852 (70.54%)
	SRR3096962	8h Activin_2	96,611,571	65,655,485 (67.96%)
H3K9ac ChIP-seq	SRR3096965	SB-431542_1	59,781,351	40,229,769 (67.29%)
	SRR3096966	SB-431542_2	61,309,061	40,820,014 (66.58%)
	SRR3096967	1h Activin_1	67,570,064	45,646,158 (67.55%)
	SRR3096968	1h Activin_2	72,466,133	47,932,200 (66.14%)
	SRR3096969	4h Activin_1	62,747,501	41,969,430 (66.89%)
	SRR3096970	4h Activin_2	57,600,750	37,943,964 (65.87%)
	SRR3096971	8h Activin_1	80,545,904	54,843,433 (68.09%)
	SRR3096972	8h Activin_2	81,769,081	53,229,993 (65.10%)

3.2.3.3 Peak calling

After sequencing reads are mapped to the genome, the next step for Smad2 ChIP-seq samples is to identify enriched regions (Figure 18). The enriched regions (so-called ‘peaks’)

are enriched in the ChIP sample relative to the control and with statistical significance. They are candidates of the binding locations of the protein of interest. These regions are estimated by the number of reads in a given size window and assessed by criteria such as enrichment over the control and minimum read density. The Poisson model was used previously to access the significance of peaks. The corresponding hypothesis is that if the reads were randomly distributed among the genome, then the probability of observing a peak with a coverage depth of at least H reads can be given by a sum of Poisson probabilities (Robertson, Hirst et al. 2007, Visel, Blow et al. 2009)

$$1 - \sum_{k=0}^{\mathrm{H}-1} \frac{e^{-\lambda}\lambda^{k}}{k!}$$

whereby λ is the global coverage level which is given by: (read length * number of aligned reads)/mappable genome length.

Unfortunately, the genomic background is not uniform in the ChIP-seq data (Zhang, Rozowsky et al. 2008). This bias may be explained by the numerous variations of genome copies, chromatin structure and sequencing or mapping biases (Zhang, Liu et al. 2008). Because of this, we used MACS (Zhang, Liu et al. 2008) for peak calling. The MACS program implements a two-step approach. First, it determines a fixed size of windows and locates enrichment regions which have more reads than the fold-enrichment of the windows as relative to a random distribution of genome-wide reads. The MACS program selects 1000 of these regions randomly and aligns them using the center of their Watson and Crick peaks. The distance between the summits of the Watson and Crick peaks are defined as '*d*'. All of the reads are shifted by *d*/2 toward the 3' ends to the most probable site of protein-DNA interaction. In the second step, the MACS program slides 2*d* windows across the genome and calculates the enrichment (Poisson distribution *p-value* based on local λ) to find candidate peaks. In an evaluation of peak-calling algorithms, MACS shows great estimates of precise binding location with respect to the qPCR data (Wilbanks and Facciotti 2010).

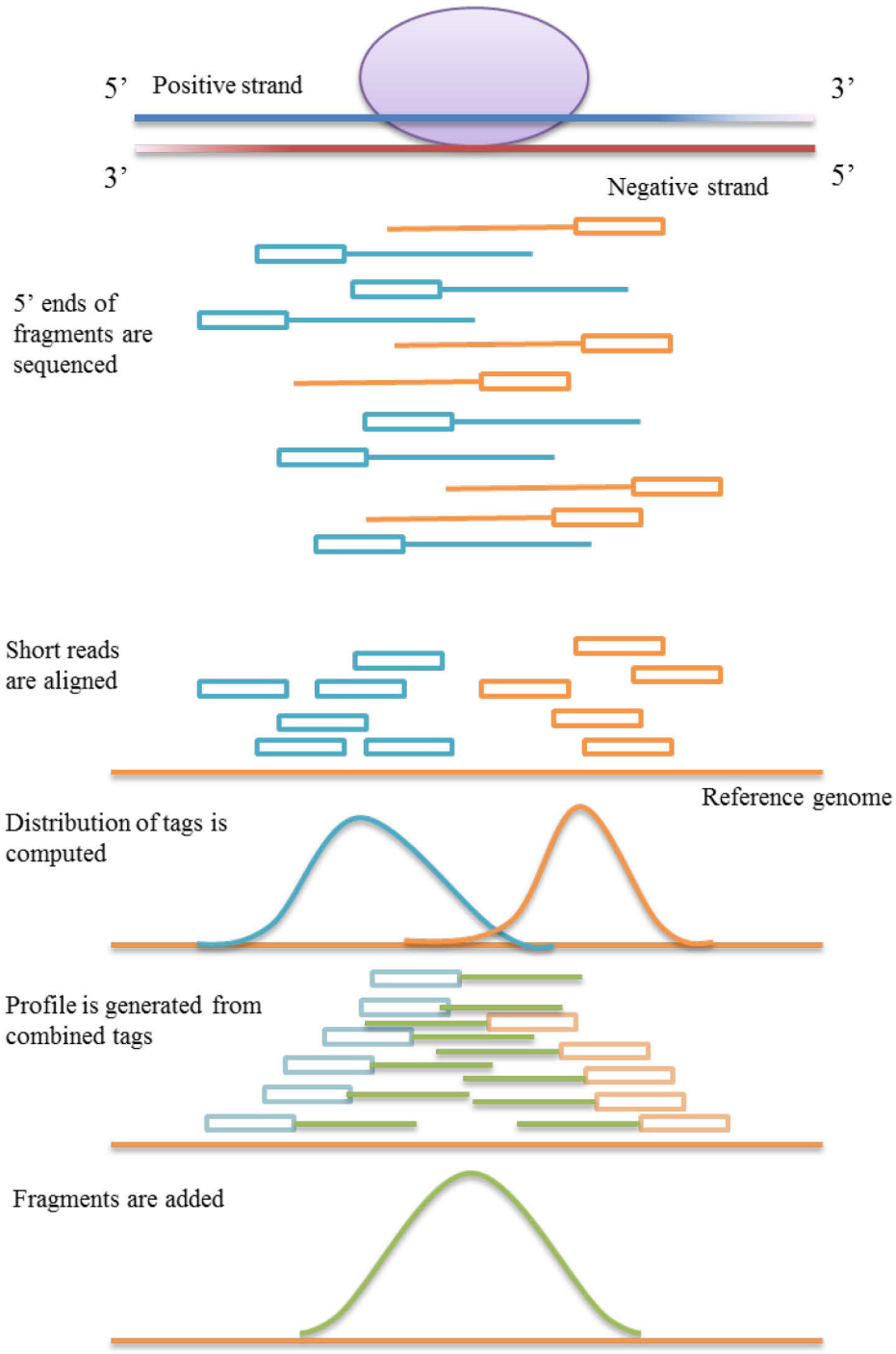


Figure 18 | Strand-specific profiles at enriched sites. The reads aligned to the reference genome result in either sense (blue) or antisense (red) peaks. These two peaks show the target protein binding regions. Each mapped read is extended by an estimated fragment size to create an approximate distribution of all reads. Figure modified from (Park 2009).

We ran the MACS program with a q-value filter of 0.05 and other parameters were set as default. After peak calling, no peaks were detected in the SB-431542 treated sample. By

comparing the 1h and 8h Activin stimulation samples to the input samples, 859 and 4410 peaks of significant Smad2 enrichment were defined, respectively. The peaks located on mitochondrial DNA were removed before further analysis.

3.2.3.4 Peak annotation

To define the activity loci that are applicable across the different time points, all of the peaks were merged into one peak if there was at least 1 bp overlap between two peaks. 1,298 peaks overlapped and the mean length of the overlap was 344 bp (Figure 19).

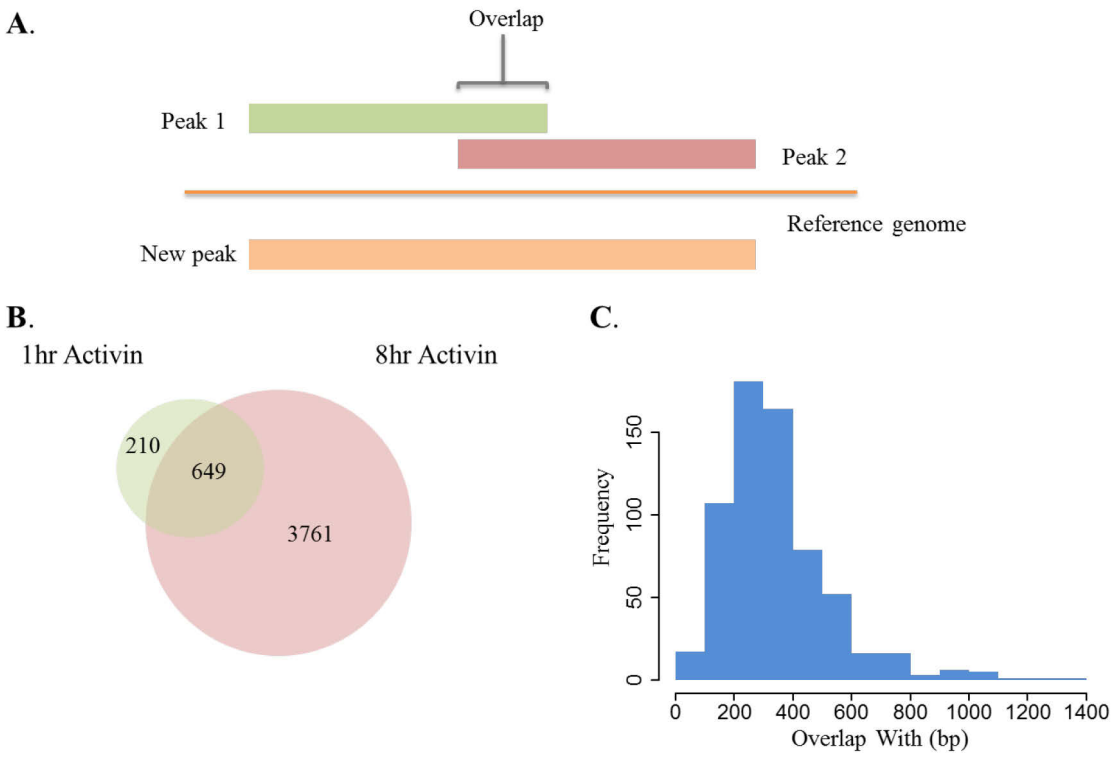


Figure 19 | Merging Peaks from peak calling results. A. If the two peaks have more than 1 bp overlap, they will be merged into a new peak. B. The number of peaks in each condition is shown in the Venn diagram. C. The distribution of the lengths of peak overlaps.

The merged peaks were annotated to the nearest gene within 100 kb from the gene's transcription start site/transcription termination site (TSS/TTS) using PAVIS (Huang, Loganantharaj et al. 2013) (Figure 20). 4,339 of 4,615 (94.02%) merged peaks were successfully associated with genes. The majority of binding events occur at regions distance from TSS. In this way, we identified 70 Activin-Smad2 target genes related to 133 Smad2 peaks.

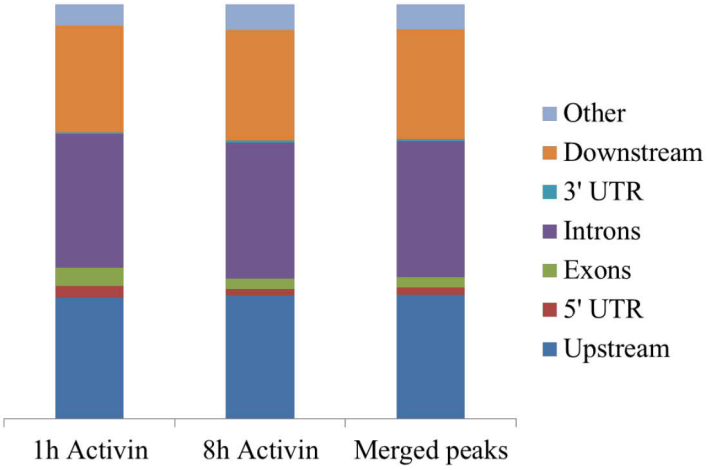


Figure 20 | Distribution of peaks in relation to genes. The plot shows the genomic distribution of peaks based on annotation results. The mouse genome was subdivided into upstream, 5' UTR, exons, introns, 3' UTR, downstream and other regions. Subsequently, the peaks were counted for each region. The distribution of peaks is in a similar fashion over the three conditions.

3.2.3.5 Quantification

The ChIP-seq data were counted differently using bedtools (Quinlan and Hall 2010) (Figure 21). We quantified gene Pol II activity by summing up the read counts in the last 20% of the gene body to the 3' end. The gene body was defined as the gene range in the reference genome. Smad2 activity is quantified by summing up the reads across the merged peaks that were annotated to the same gene. The H3K27ac and H3K9ac counts are in the region of +/- 2,000 bp around the gene TSSs.

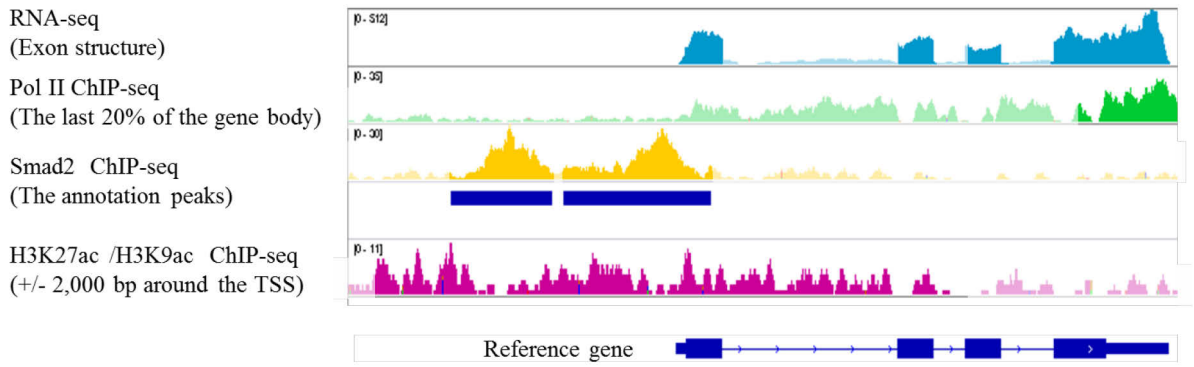


Figure 21 | Quantification regions for different data types. The coverage profiles and quantification regions for RNA-seq reads (blue), Pol II ChIP-seq reads (green), Smad2 ChIP-seq reads (yellow) and histone modification ChIP-seq reads (rose) were visualized using the Integrative Genomics Viewer (IGV) (Robinson, Thorvaldsdottir et al. 2011).

3.2.3.6 Data normalization

For Smad2, H3K9ac and H3K27ac ChIP-seq data, the raw read counts were normalized to reads per million (RPM). For Pol II ChIP-seq data, the raw read counts were normalized by DESeq2. The ChIP-seq data were compared against the SB-431542 sample and converted into FC values.

3.3 Summary

In this chapter, we analyzed the raw sequencing data and selected different analysis methods depending on the types sequencing data. In the workflow, we mapped the sequencing data to the genome and converted the sequencing data into FC value that are related to the baseline of the sample after we screened the data quality and homogenized the alignment results. Based on the FC, we identified 198 differentially-regulated genes after Activin stimulation and annotated 70 target genes with Smad2 binding peaks (Figure 22). This provides valuable data for the subsequent modeling analysis (Chapter 4). In addition, we annotated the Smad2 peaks and discovered that the percentage of annotation of Smad2 binding sites did not change much across different time points. Furthermore, most of the Smad2 peaks were annotated far apart from the TSS of genes. These loci indicate the cis-regulatory regions of the target genes.

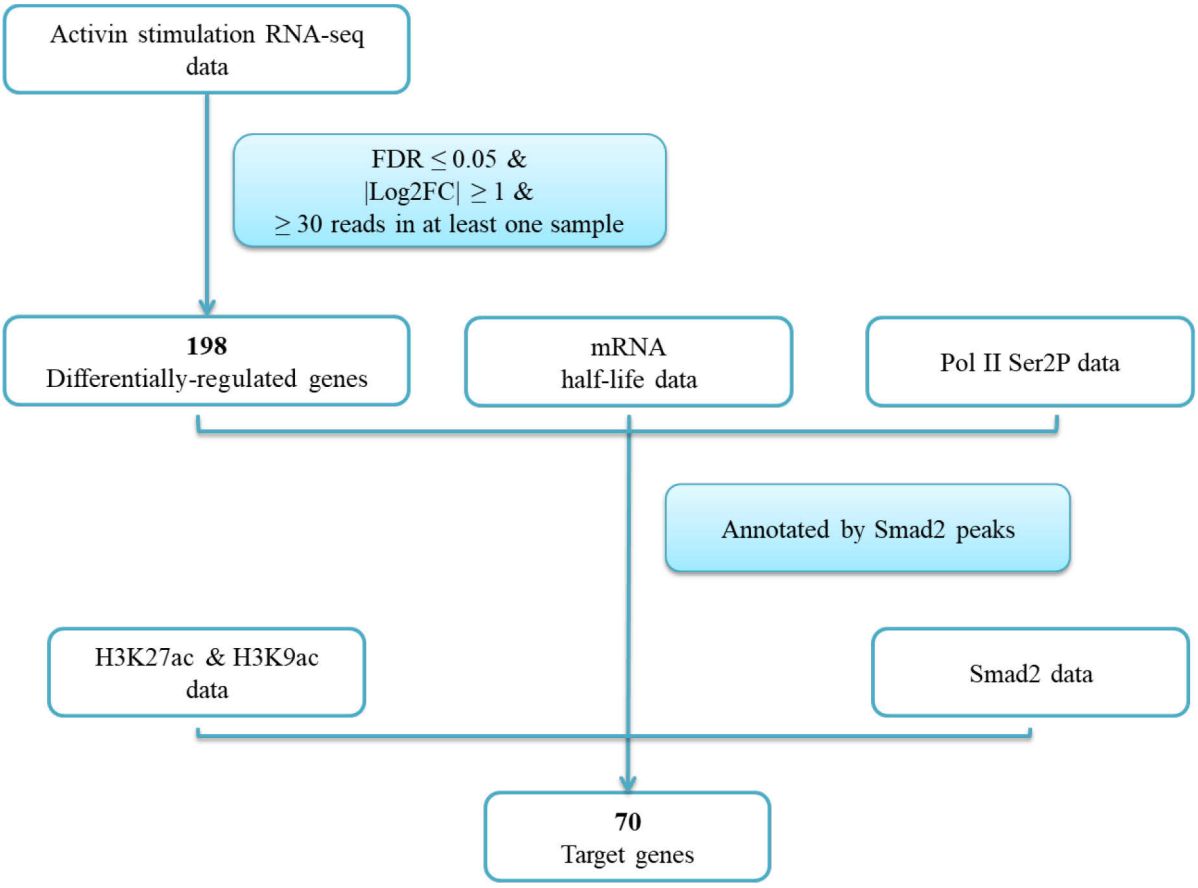


Figure 22 | Differentially-regulated genes and target genes selection.

CHAPTER 4

Kinetic modeling of the transcriptional responses to the Activin signal

4.1 Introduction

A major objective of this chapter is to understand the transcriptional responses of P19 mouse cells with Activin stimulation. After selecting differentially-regulated genes (Chapter3), we started to develop mathematical models for explaining different dynamics of gene expressions induced by Activin. In the absence of Smad2 activity, the transcriptional dynamics of most genes cannot be explained just by the differences of Pol II occupancy and/or mRNA half-life. After fitting different models to the transcriptional data of differential-regulated genes, we selected the best model for each gene based on Akaike information criterion (AIC). The model selection results suggested that: (i) the mechanism by which Activin regulates gene expression is diverse and has different regulatory approaches for different genes, (ii) the relationship between Pol II binding activity and mRNA expression level is non-linear, and (iii) some of target genes do not require local Smad2 chromatin binding.

4.2 Activin induces multiple temporal patterns of gene expression

Through analysis of the RNA-Seq data, we identified 198 differentially-regulated genes relative to the SB-431542-treated state. Due to the use of different sequence alignment tools and strict filters for gene identification, we had fewer differentially-regulated genes than the original paper. These expressions were converted into log2FC values (relative to SB-431542)

for each time point. Gene ontology analysis showed the enrichment of these differentially-regulated genes for TGF-β signaling pathway and development processes (Ashburner, Ball et al. 2000, The Gene Ontology 2019) (Table 5). We related how the Pol II occupancy changed over time relative to the transcription as measured by RNA-seq (Figure 23). The differentially-regulated genes showed various responses after Activin simulation.

Table 5 | Gene ontology enrichment analysis of differentially-regulated genes. The data was analyzed by GO Enrichment Analysis (Mi, Muruganujan et al. 2019).

	PANTHER Pathways	ID	In Data	Fold Enrichment	P-value	FDR
1	TGF-beta signaling pathway	P00052	7	9.91	1.07E-05	8.79E-04
2	Gonadotropin-releasing hormone receptor pathway	P06664	8	4.63	4.29E-04	2.35E-02
	Reactome pathways	ID	In Data	Fold Enrichment	P-value	FDR
1	MAPK family signaling cascades	R-MMU-5683057	11	5.52	7.57E-06	1.22E-02
2	RAF/MAP kinase cascade	R-MMU-5673001	10	5.64	1.65E-05	1.34E-02
3	MAPK1/MAPK3 signaling	R-MMU-5684996	10	5.5	2.03E-05	1.09E-02
4	Signaling by TGF-beta family members	R-MMU-9006936	6	8.87	8.36E-05	3.38E-02
5	Signaling by BMP	R-MMU-201451	4	19.42	8.40E-05	2.72E-02
	GO biological process	ID	In Data	Fold Enrichment	P-value	FDR
1	Regionalization	GO:0003002	29	11.3	2.17E-21	3.42E-17
2	Pattern specification process	GO:0007389	31	9.41	1.19E-20	9.35E-17
3	Tissue development	GO:0009888	53	4.35	2.59E-20	1.36E-16
4	Animal organ morphogenesis	GO:0009887	42	5.68	5.41E-20	2.13E-16
5	Epithelium development	GO:0060429	42	5.55	1.25E-19	3.94E-16

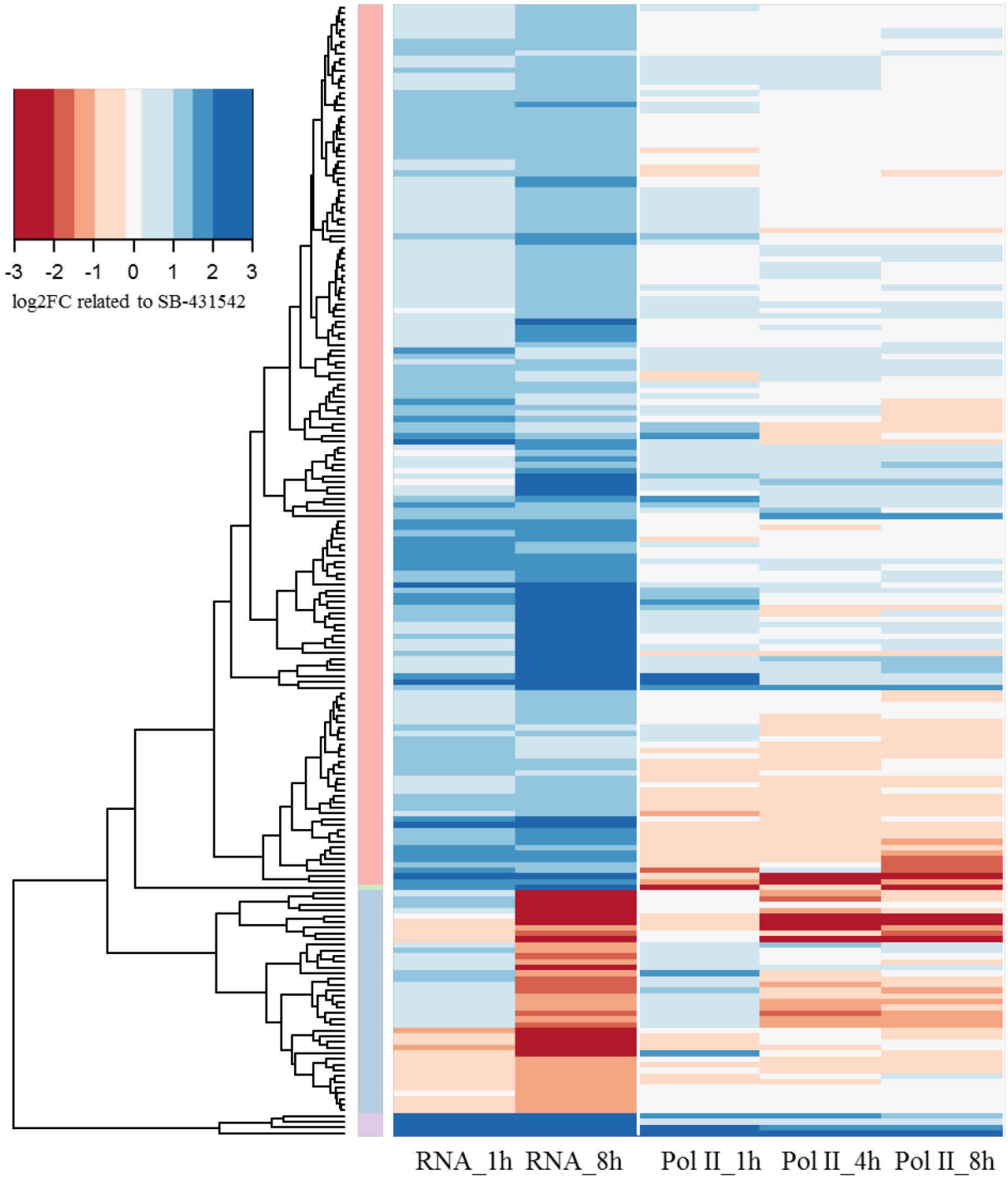


Figure 23 | Hierarchically-clustered heatmap for each differentially-regulated gene showing log2FC values, relative to SB-431542 for gene expression, as determined by RNA-seq (left), Pol II Ser2P binding level (right).

4.3 Estimation of transcriptional activity using Gaussian process regression

Due to only limited mRNA data points are available, fitting models to these data may cause overfitting, which is "the use of models or procedures that violate parsimony-that is, that include more terms than are necessary or use more complicated approaches than are necessary" (Hawkins 2004).To overcome this issue, we modeled mRNA half-life and Activin time series data as a Gaussian process (GP) prior distribution that is nonparametric and avoids assuming a specific shape (Rasmussen and Williams 2006). A GP distribution on *f* is written as:

$$p(f) = GP(f;\ \mu, K) \tag{1}$$

It is completely specified by its first two moments: the mean function μ and covariance function *K*.GP is a collection of random variables. This means that a GP is a Gaussian distribution over functions. In recent studies, GPs have been applied in biological dynamical systems (Gao, Honkela et al. 2008, Honkela, Girardot et al. 2010, Liu and Niranjan 2012).

In this study, we used a GP to model the temporal dynamics of the mRNA data. The used GP kernel function is the squared exponential covariance function with the length scale *l*, as shown below:

$$K(x, x') = \exp(-\frac{(x - x')^2}{l^2}) \tag{2}$$

Additionally, the mean function was set to zero. The parameters of a prior distribution such as the length scale *l*, the output scale λ and the noise variance σ^2, are called hyperparameters θ, which denotes the vector of the hyperparameters and is chosen by maximizing the Gaussian likelihood as shown below:

$$Gaussian\ likelihood(t) = \frac{\exp(-\frac{(t-y)^2}{2} * sn^2)}{sqrt(2 * \pi * sn^2)} \tag{3}$$

Where *y* is the mean and *sn* is the standard deviation.

Here, we used the *GPML* function of MATLAB for Gaussian process regression models on mRNA-seq data (Rasmussen and Nickisch 2010) (Figure 24). With the Gaussian process regression models, we sampled 16 mRNA half-life data points (from 0h to 3.75h, the data was taken every 15min) and 17 Activin induced gene expression data points (from 0h to 8h, the data was taken every 30min).

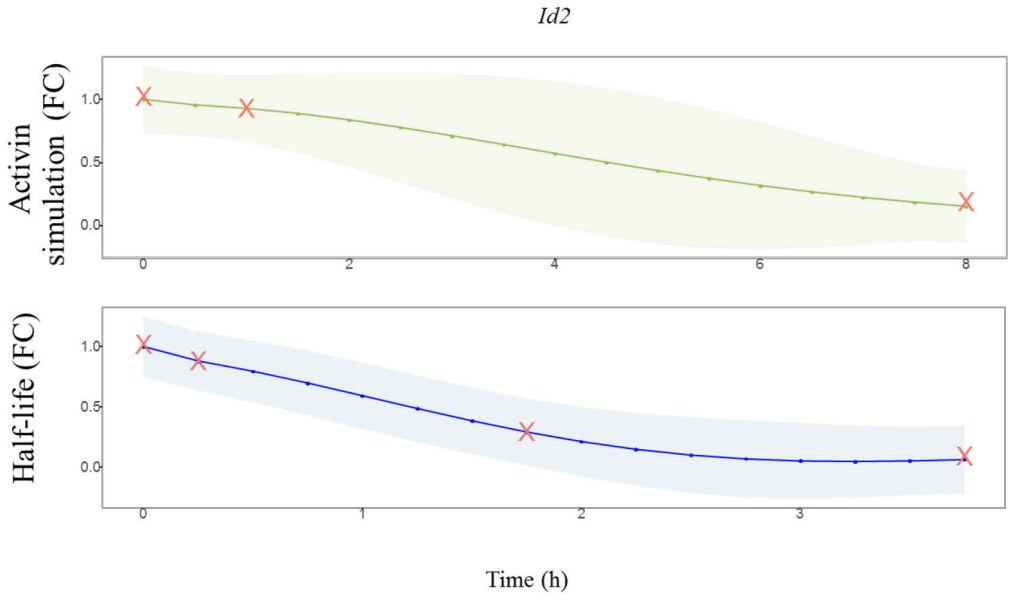


Figure 24 | An example of the mRNA half-life data sampled from the GPR model. For each gene, we used the GPML function to simulate the data. Solid green and blue lines show the data for mRNA data after Activin stimulation and mRNA half-life data, respectively. The shadow zones indicate associated 95 confidence interval regions.

4.4 A simple kinetic model for transcriptional responses to Activin

To understand how TGF-β signaling regulates transcripts dynamics, A system of ordinary differential equations (ODE) model is widely used for simulating the kinetics of gene expression (Ropers, de Jong et al. 2006). We first developed a simple model to fit gene transcriptional data sets. This model assumes that mature mRNA levels are determined by a first-order integration of production and degradation rates over time (Rabani, Levin et al. 2011).

In this model (model 1), the changes of mRNA (x) amount is determined by mRNA production rate and degradation rate. The mRNA production rate is assumed to be proportional to the activity of Pol II (P).

Model1
$$\frac{dX}{dt} = a * P - \gamma * X \tag{4}$$

The activity of Pol II for a certain gene is measured by the binding signal of Pol II Ser2. The mRNA abundance is measured by normalized RNA-seq reads mapped onto the gene exons. As the data were converted to fold change (FC) values, we transform equation (4) to a dimensionless equation by rescaling the variables to their initial values according to equation (5) and (6).

$$d\frac{\frac{X}{X_0}}{dt} = \frac{PII_0}{X_0} * a * \frac{PII}{PII_0} - \gamma * \frac{X}{X_0} \tag{5}$$

$$\frac{dx}{dt} = \alpha * P - \gamma * x \tag{6}$$

Here, α is a scaling factor for the transformation. The variables x and P denote fold change of mRNA and Pol II activity for a gene, respectively. The parameter γ determines the mRNA half-life. For mRNA, an exponential decay can be described by

$$N(t) = N_0 * e^{-\gamma t} \tag{7}$$

where N_0 is the initial mRNA abundance, $N(t)$ is the remaining mRNA at time t after degradation. For each gene, the half-time ($t_{1/2}$) is defined as the time point when $N(t)$ is a half of N_0. The mRNA half-life can be calculated from γ:

$$t_{1/2} = \frac{ln(2)}{\gamma} \tag{8}$$

γ can be estimated from mRNA half-life dataset by fitting the following model to the mRNA half-life data:

$$\frac{dx}{dt} = -\gamma * x \tag{9}$$

We assume that gene transcriptions are in a steady state in the baseline sample (0h), which means that the mRNA production rate is equal to the degradation rate for each gene ($\alpha * P - \gamma * x = 0$). Because the variables are transformed to fold changes, P and x at 0h equal to 1. We can obtain $\alpha = \gamma$.

We applied the Data2Dynamics (D2D) tool to fit this simple ODE model (model 1) to corresponding data for differential regulated genes (Raue, Schilling et al. 2013, Raue, Steiert et al. 2015). The range of estimated parameter values were listed in Table 6.

Table 6 | The ranges of estimated parameter values used in D2D.

Parameter symbol	Meaning	Range	Ref.
γ	mRNA half-life	$10^{-1.6}$~$10^{-3.6}$ min^{-1}	(Schwanhausser, Busse et al. 2011)

After simulation, the simulated results would be checked by a statistical measure called chi-square (x^2) to estimate how well this model fits the observations (Press, Flannery et al. 1992). The chi-square is based on chi-square distribution (a special kind of gamma distribution) and it is defined as the sum of the squares of independent standard normal random variables (Bennett and Franklin 1954):

$$chi - square = \sum_{i=1}^{n} \frac{(x_i - y_i)^2}{{\sigma_i}^2} \tag{10}$$

Where in our model, x_i is the fold change of mRNA, y_i is the simulated data, and σ_i is the standard deviation for x_i. When the sample size is large, the chi-square goodness-of-fit statistic is approximately a chi-square random variable. It is commonly used to test whether the given data are well described by the hypotheses. After calculating the chi-square distribution, the *p-value,* which refers to the probability of observing a more extreme value than the simulated data in the chi-squared distribution, can be evaluated based on degrees of freedom (*df*). We

used a *p-value* filter of 0.05 to evaluate whether the model can fit the data for the corresponding gene.

As a result, most target genes cannot be explained by this model. Only 21 of 198 (10.6%) target genes can be fitted by this simple linear model. This result indicate more complicated models are required to explain the observed gene transcription data sets.

4.5 Delays cannot explain transcription kinetics

Recent studies have shown that RNA splicing can be the timing-limited step of mRNA production (Hao and Baltimore 2013). A splicing-associated delay can play a key role in RNA production. We next developed a delay model to account for the processing delay between the Pol II activity and final mRNA production (Honkela, Peltonen et al. 2015) (Eq. 9).

$$\frac{dx}{dt} = \alpha * P_{(t-\tau)} - \gamma * x \qquad (11)$$

The new parameter τ is defined as the delay time during mRNA processing.

It seems that the delay model could not fit more genes than the dynamic model. Only 21 of 198 (10.6%) target genes can be fitted by this delay model. In addition, the estimation of parameter τ shows that the majority of genes do not have a long delay time (Figure 25). The estimated delay time for most of genes (65.66%) is less than 10 min.

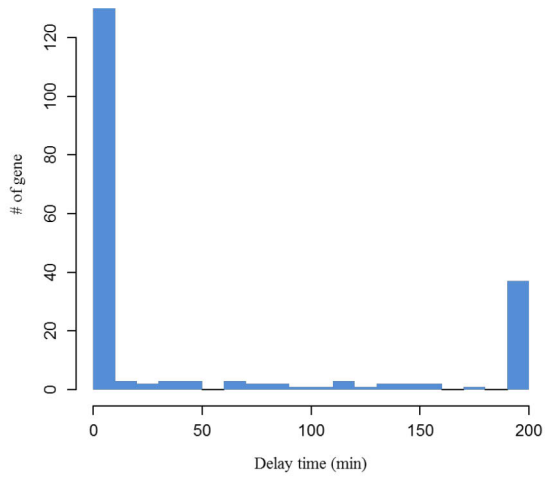


Figure 25 | Distribution of delay from target genes.

The results show that the delay model in equation (11) cannot explain the transcription dynamic of the target genes. We analyzed the FC of mRNA and Pol II at the 8h time point. If we assume that gene expression at 8h after Activin stimulation is approximately in a steady state, the FC of mRNA would approximately equal to FC of Pol II at 8h, when $\alpha = \gamma$. However, for most of genes, the FC of mRNA is much higher than the FC of Pol II (Figure 26). This may be explained by the variations of Pol II elongation rates among different genes. It has been reported that the Pol II density of the gene body has a striking correlation with the elongation rate ($\rho = 0.46$; $p = 4.5 \times 10{-}10$) (Danko, Hah et al. 2013).

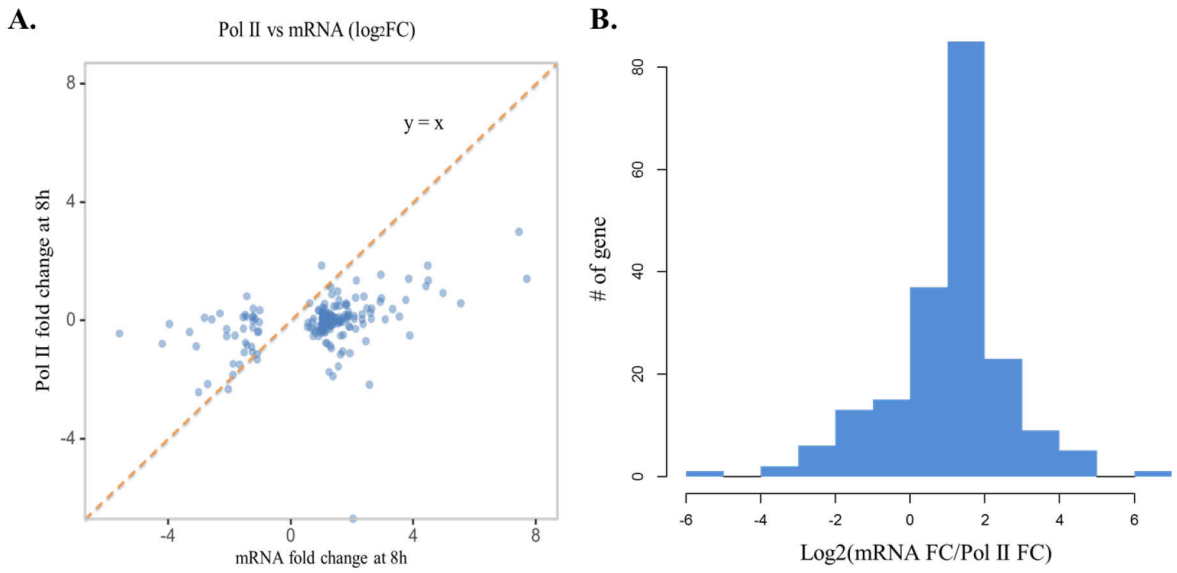


Figure 26 | The FC of mRNA is stronger than the FC of Pol II. A. Point plots showing normalized (log2) mRNA FC and Pol II Ser2P FC at 8h.The orange line represents the graph 'y = x'. B. Distribution of mRNA FC/Pol II FC.

4.6 A revised model with a non-linear relationship between RNA polymerase II density and mRNA expression

In order to estimate the elongation rate via Pol II occupancy, we used Hill functions to model the transcription rate:

Model2

$$\frac{dx}{dt} = \alpha * \frac{P^{n_p}}{K_p^{n_p} + P^{n_p}} - \gamma * x \qquad (12)$$

Where K_p is the FC of Pol II that corresponds to half of the maximum mRNA production rate and n_p is the Hill coefficient. The bound for K_p is set based on the Pol II FC range and n_p is set from 1 to 5. The non-linear Pol II model can fit 66 genes, which explains about 34.34% of all 198 differentially regulated genes.

In summary, due to the different magnitude of fold changes between Pol II density and mRNA, the linear or delay models can only fit the data for a few target genes. While the model using the nonlinear model can largely increase the number of fitted genes from about 10% to 34%. Further regulations would need to be taken account in order to fit mRNA data for other genes.

4.7 Kinetic models for the expression of 70 genes with Smad2 binding density

Previous work has shown that Smad protein play important role in regulating gene expression in the TGF-beta signaling. We next aimed at developing new models that include the regulation of mRNA production by Smad2 binding. After annotating the Smad2 peaks to the closet gene within 100 kb from the gene's TSS/TTS already. We found the majority of these peaks are located within a window of 10kb or even only 1kb around the TSS of their regulated genes (Figure 27). We identified 70 genes that are associated with 133 Smad2 peaks after Activin stimulation. For further analysis, we first compared the FC of Smad2 activity at different times with the mRNA abundance and Pol II activities of the target genes (Figure 28). Although it seems that both mRNA and Smad2 reached its respective highest level at the 8h time point, the Smad2 binding level is not correlated with the transcription abundance. This result is agreement with the findings in Coda *et al.* (Coda, Gaarenstroom et al. 2017).

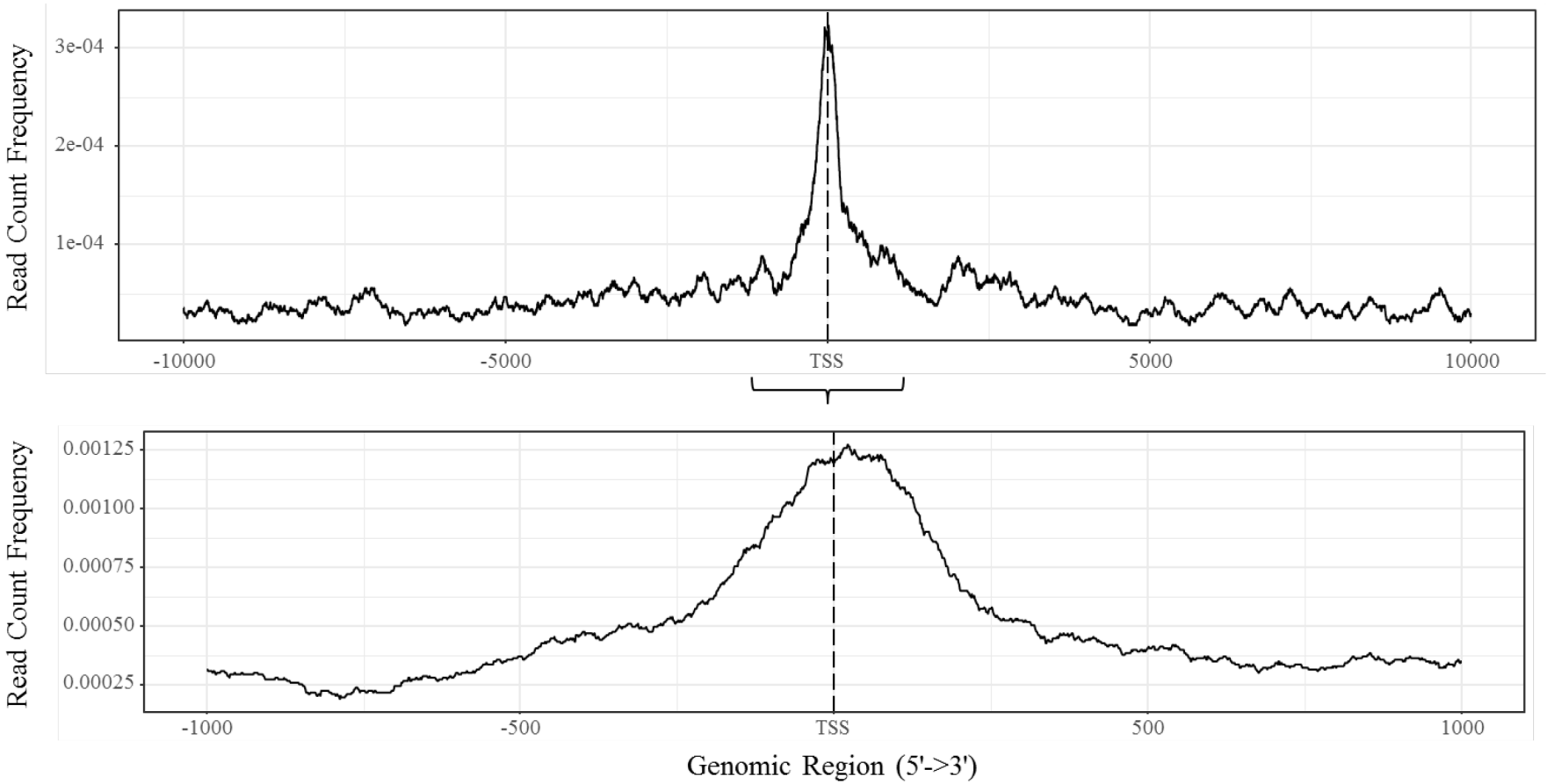


Figure 27 | The density of Smad2 peaks and their distance from the annotated TSS of the nearest regulated target gene within a ± 10kb or ± 1kb window.

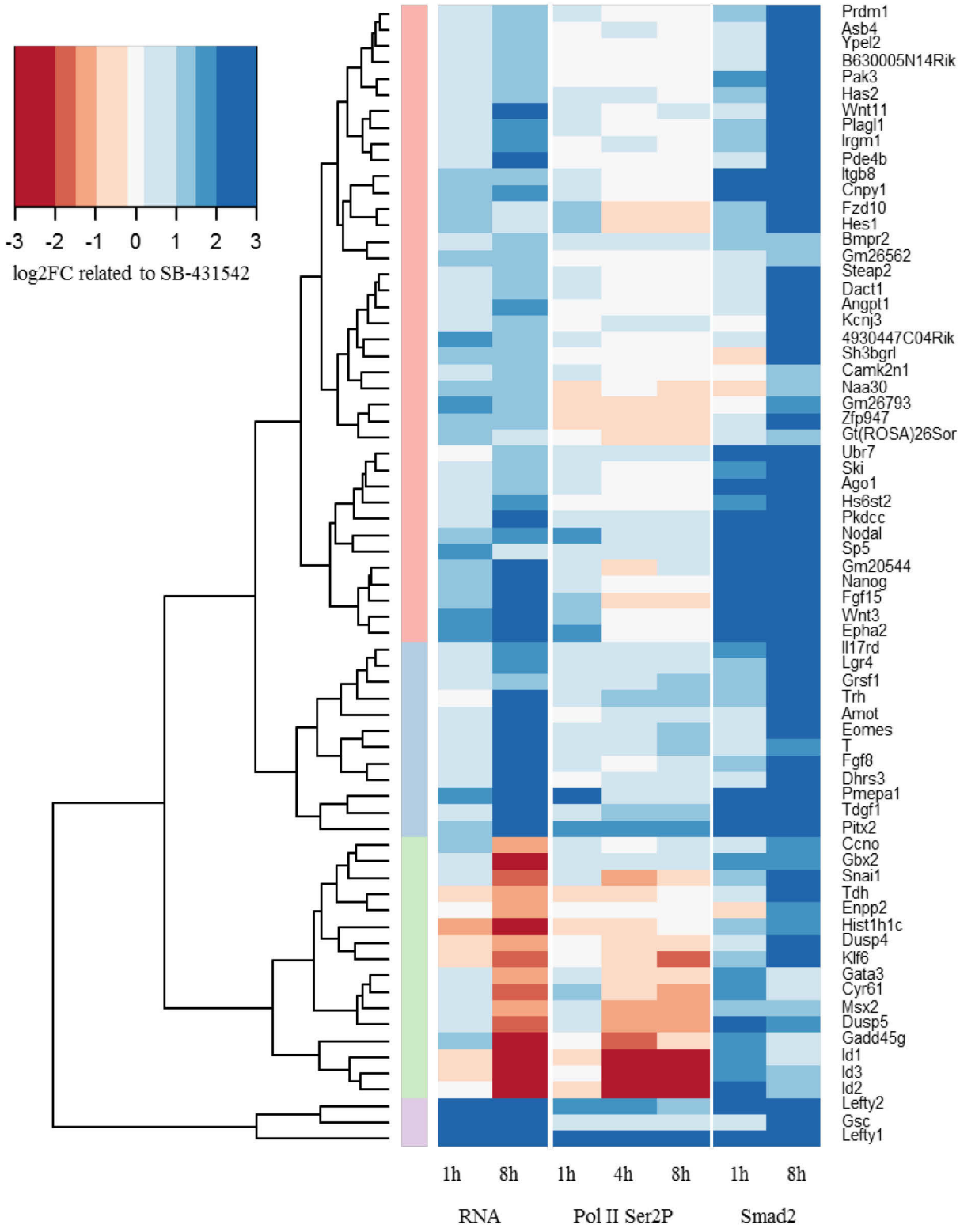


Figure 28 | Hierarchically-clustered heatmap for each target gene showing log2FC values, relative to SB-431542 for gene expression, as determined by RNA-seq (left), Pol II Ser2P binding level (middle) and Smad2 binding level (right).

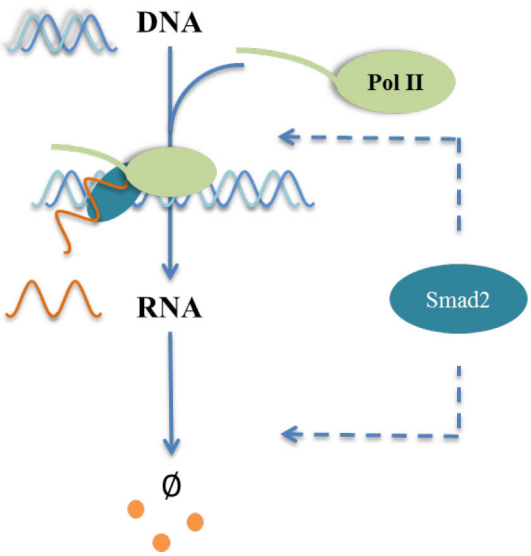


Figure 29 | An overview of the generative model with Smad2 activities.

Through the regulation of Smads, TGF-β superfamily signal can induce activation or repression of target genes (Ross and Hill 2008). In addition, many researches have shown that TGF-β superfamily can regulate transcription processes by mRNA alternative splicing or mRNA degradation through Smads (Chen, Zhou et al. 2016, Tripathi, Sixt et al. 2016). Therefore, we hypothesize that Smad2 could have both positive or negative effects on mRNA production or degradation steps (Figure 29). Correspondingly, we devloped 8 new models, which are shown in equation 13-20.

Model3
$$\frac{dx}{dt} = \alpha * \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * P - \gamma * x \tag{13}$$

Model4
$$\frac{dx}{dt} = \alpha / \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * P - \gamma * x \tag{14}$$

Model5
$$\frac{dx}{dt} = \alpha * P - \gamma * \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * x \tag{15}$$

Model6
$$\frac{dx}{dt} = \alpha * P - \gamma / \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * x \tag{16}$$

Model7 $$\frac{dx}{dt} = \alpha * \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * \frac{P^{n_p}}{K_p^{n_p} + P^{n_p}} - \gamma * x \quad (17)$$

Model8 $$\frac{dx}{dt} = \alpha / \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * \frac{P^{n_p}}{K_p^{n_p} + P^{n_p}} - \gamma * x \quad (18)$$

Model9 $$\frac{dx}{dt} = \alpha * \frac{P^{n_p}}{K_p^{n_p} + P^{n_p}} - \gamma * \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * x \quad (19)$$

Model10 $$\frac{dx}{dt} = \alpha * \frac{P^{n_p}}{K_p^{n_p} + P^{n_p}} - \gamma / \left(1 + \frac{\beta * S^{n_s}}{K_s^{n_s} + S^{n_s}}\right) * x \quad (20)$$

Where *S* is the Smad2 binding signal. The meaning and bounds of model parameters are shown below (Table 7).

Table 7 | Bounds for different model parameters. We list the values of the lower and the upper bounds. The *K* was set based on the FC values of ligand.

Parameter symbol	Description	Lower bound	Upper bound
γ	mRNA half-life	$10^{-3.6}$ min^{-1}	$10^{-1.6}$ min^{-1}
β	Scale of Smad2 estimates	0.01 min^{-1}	100 min^{-1}
K_p	Activation coefficient of Pol II FC	0.01	160
K_s	Activation coefficient of Smad2 FC	0.1	10
n_p	Hill coefficient of Pol II FC	1	5
n_s	Hill coefficient of Smad2 FC	1	5

As with the previous models, we used the chi-square distribution and the *p*-value to evaluate model fitting results. After incorporating Smad2 in the models， we were able to obtain good model fits for 59 out of 70 (84.23%) genes, which has Smad2 binding peaks. In summary, after Smad2 binding signal is added to the models, the fitting results of target genes greatly improved. It can be concluded that Smad2 plays an important role in gene regulation. However, many genes can be fitted by more than one model. In the next stage, we would like to choose the best order for the mRNA regulation of each gene.

4.8 Model selection using Akaike's Information Criterion

Now we have developed a pool of kinetic models to fit the mRNA datasets for Activin regulated genes, we next ask which model is most likely true given the data. The model with a large number of parameters may fit the data better, but it has a high risk of overfitting. To select the best model, we used Akaike Information Criterion (AIC) (Burnham, Anderson et al. 2002), which is defined as

$$AIC = -2(log - likelihood) + 2K \tag{21}$$

Where K is the number of parameters.

If the sample size is small, the AIC may select the model that overfits the data (McQuarrie and Tsai 1998). To address this problem, AICc was developed by Sugiura in 1978, which is AIC with a correction for small sample sizes.

$$AIC_c = AIC + \frac{2K^2 + 2K}{n - K - 1} \tag{22}$$

Where n is the sample size. If $n / K < 40$, it is recommended to use AICc (Burnham, Anderson et al. 2002). With least squares fitting, the log-likehood is calculated by the residual sum of squares, which is called *RSS*,

$$Log - likelihood = -(\frac{n}{2})log\left(\frac{RSS}{n}\right) \tag{23}$$

thus

$$AIC_c = nlog\left(\frac{RSS}{n}\right) + 2K + \frac{2K^2 + 2K}{n - K - 1} \tag{24}$$

Based on AICc value, we can select the model that has minimal information loss. However, the model with the smallest AICc value may not always be the best one.

To put AICc into practice, we also need to estimate the information loss among the models to find the probability for each model as the best model. Let's take the set of m models as an example. First, we need to account for the AICc differences Δ over all of the candidate models in the sets. The AICc difference Δ_i is the AICc of model i minus the AICc of the best model. The AICc differences are easy to interpret and provide a quick comparison of the candidate models. For the best model, $\Delta_i = \Delta_{min} = 0$. Then, we can calculate the model probabilities ω_i (Burnham, Anderson et al. 2002)

$$\omega_i = \exp\left(-\frac{\Delta_i}{2}\right) / \sum_{i=1}^{m} \exp\left(-\frac{\Delta_r}{2}\right) \tag{25}$$

The sum of the ω among the candidate models is 1. ω_i represents the weight of evidence when the i^{th} model is the best model that minimizes the information loss in the set of m models.

It is also called Akaike weight. For example, if one model's ω is 0.3, it means that this model has 30% probability to be the best model in the set. Although we used AICc to screen the best model, the results we get are not necessarily accurate. The AICc weight can help us judge the possibility of other best models for analysis.

4.9 Model selection results

4.9.1 Activin regulated transcriptional responses are explained by different kinetic models

Model selection results for 10 models ranked by AICc weight are shown below (Figure 30) (Appendix A, B). For most target genes, the fitting results for mRNA half-life data are very good. Fitting of the best models to mRNA datasets are shown in Appendix C.

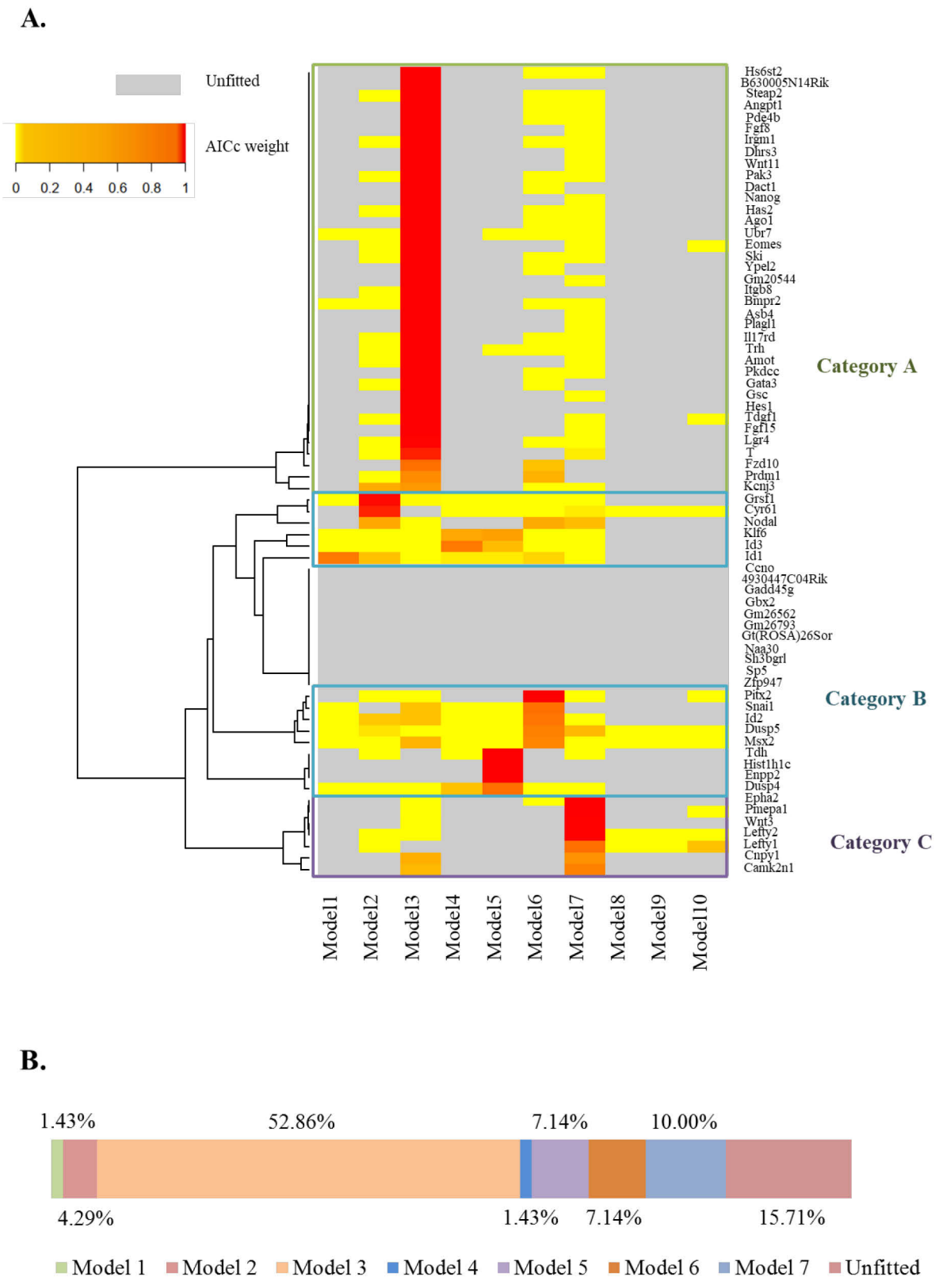


Figure 30 | Model selection results. A. Hierarchically-clustered heatmap for each target gene showing the AICc weight of fitted models (colour bar) and unfitted models (grey). B. Percentage of the best model based on AICc weight.

As it shown in Figure 30, model3 can fit 37 of 70 (52.86%) target genes that have Smad2 binding information. This result indicates that Smad2 plays a role in promoting mRNA production rate of these genes. In addition, model4 is selected as the best model for only 1 gene (Id3), indicating that Smad2 rarely repress mRNA production rate of the target genes. Furthermore, both model5 and model6 are selected as the best model for 5 genes, suggesting that Smad2 might regulate the expression of these genes through promoting or inhibiting their mRNA degradation rates. The models assuming mRNA production rate is nonlinear to Pol II activity can only account for 10 of these 70 genes that have Smad2 binding density. Interestingly, although Smad2 binds to the promoter of some genes, it doesn't contribute to the best models (model1 and model2) that fit these genes. This result suggest that Smad2 binding does not always have function on the regulation of the target gene.

We next analyzed the biological function of the genes that are best explained by model 3 (i.e.: Smad2 promoted transcription). The GO term enrichment analysis for biological process (BP) (Figure 31) and molecular function (MF) (Figure 32) was applied to identify the biological characteristics of the genes. Genes with Smad2 promoted transcription mainly involved in biological development processes, such as genes associated with metanephros, kidney, and midbrain-hindbrain boundary development or cellular developmental processes, like cell fate specification, cell migration or cell differentiation. For example, the model3 genes include members of the fibroblast growth factors (FGF) family; e.g., Fgf8 and Fgf15, which is critical in regulating cell proliferation, migration and differentiation during embryonic development (Ornitz and Itoh 2001). The FGF pathway reportedly cooperates with Activin/Nodal pathway to maintain pluripotency of hESCs (Vallier, Alexander et al. 2005). The molecular functions of model3 genes included Wnt-protein binding, growth factor receptor binding and RNA polymerase II proximal promoter sequence-specific DNA binding. All of the results showed that genes with Smad2 promoted transcription, which exhibited variable distribution of developmental processes and relationship with protein and nucleic acid bindings.

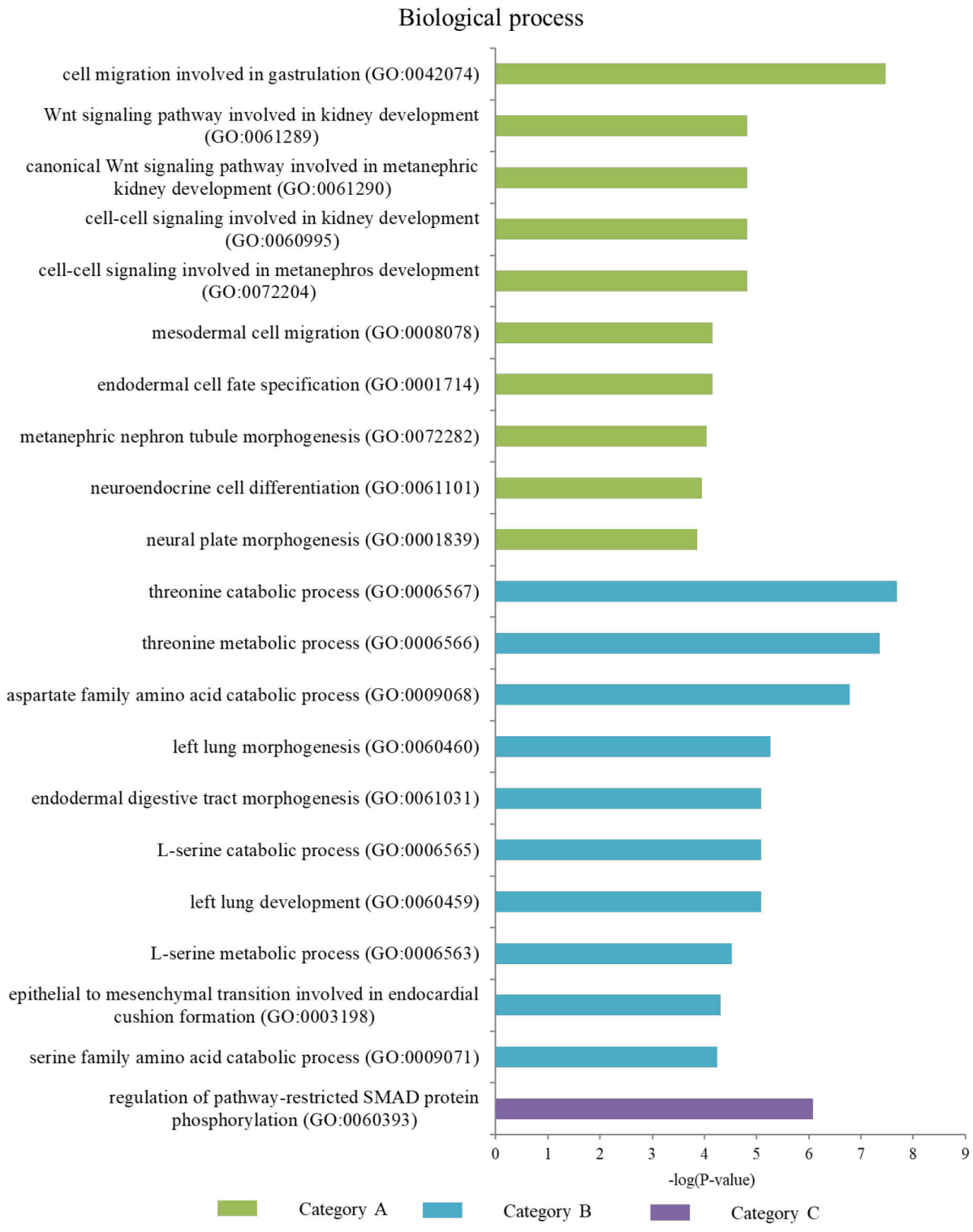


Figure 31 | Gene Ontology enrichment analysis of biological processes for fitted target genes among three categories. The data was analyzed by GO Enrichment Analysis (Mi, Muruganujan et al. 2019).

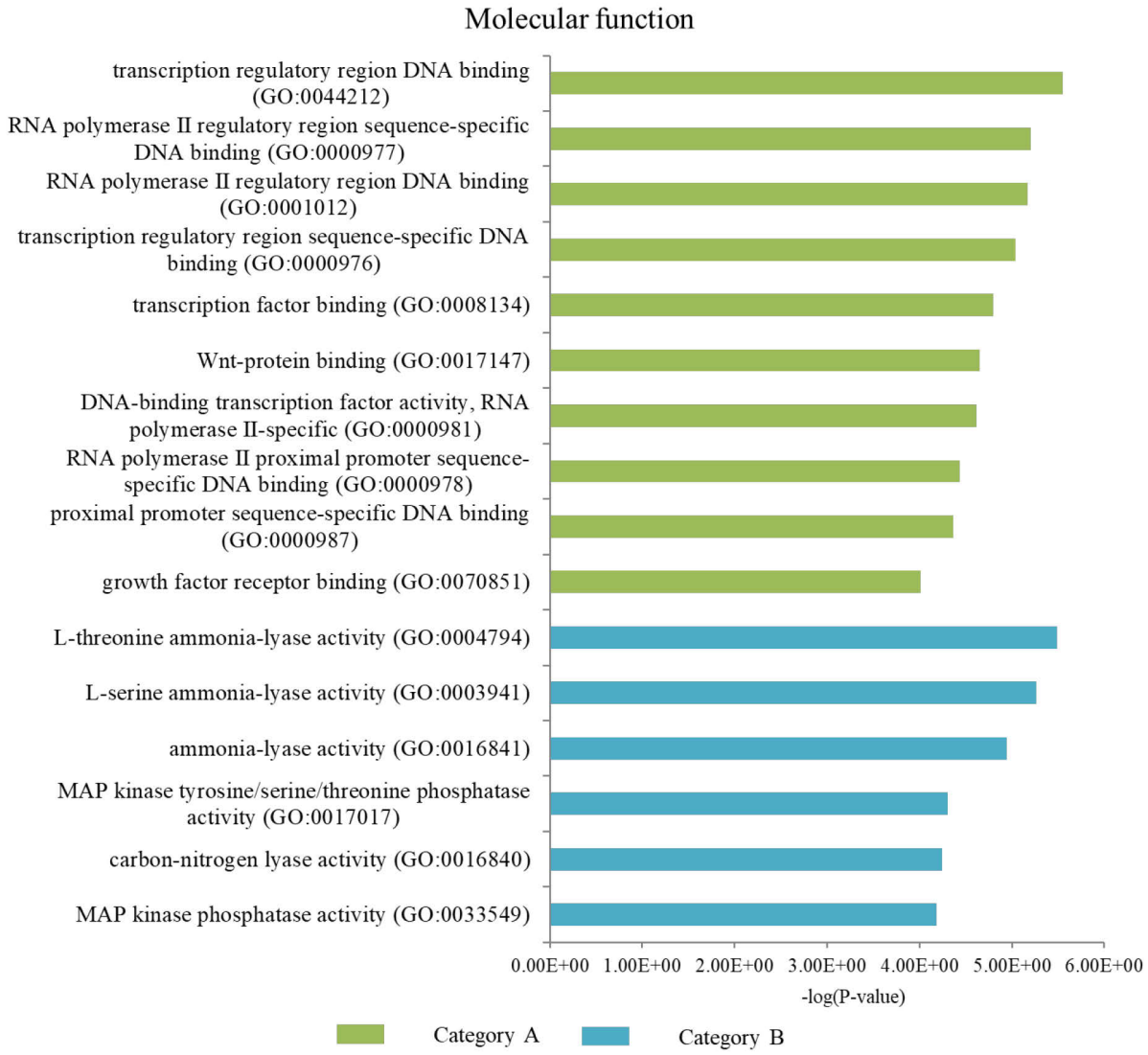


Figure 32 | Gene Ontology enrichment analysis of molecular functions for fitted target genes among two categories. The data was analyzed by GO Enrichment Analysis (Mi, Muruganujan et al. 2019).

4.9.2 Linear and non-linear Pol II regulated genes

The kinetic models can be classified into two groups: (i) linear Pol II model, in which mRNA production rate is linear to the activity of Pol II (model 1, 3-6) and (ii) nonlinear Pol II model, in which mRNA production rate is nonlinear to the activity of Pol II (model 2, 7-10). As shown in Figure 33, most of genes with Smad2 binding are best explain by the linear Pol II model.

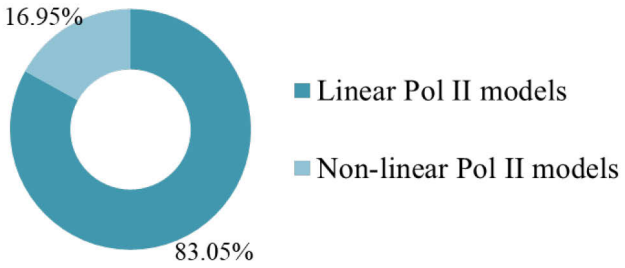


Figure 33 | Percentage of the linear and non-linear Pol II models that best explain genes with Smad2 binding activity.

4.9.3 Smad2 is not required for the regulation of some genes

There are 5 genes (Cyr61, Grsf1, Id1, Kcnj3, Nodal), whose expression data are not best fitted by the models that are related to Smad2 activity. These genes can be regarded as Smad2-independent genes. Among these genes, Id1 is a famous TGF-β superfamily target genes. ID genes are mediated by Smads and down-regulated in epithelial cells (Kang, Chen et al. 2003). Our modeling analysis result is consistent with previous studies. For example, it has been reported that Smad3, but not Smad2, mediates early expression of Id1 by TGF-β1 (Liang, Brunicardi et al. 2009). Expression of Id1 was regulated differently as compared to Id2 and Id3 in healing skin wounds, during which Activin is highly expressed (Hubner, Hu et al. 1996, Rotzer, Krampert et al. 2006). In the P19 cell line, it also seems that the Id1 and Id2 are regulated differently by Nodal/Activin (Coda, Gaarenstroom et al. 2017). In addition, Cyr61, also known as cysteine-rich angiogenic inducer 61, is a member of the CCN family. After the specific knockdown of Smad2 and Smad3 by RNA interference, the TGF-β1-induced secreted CCN protein is not affected by Smad2 knockdown (Phanish, Wahab et al. 2006).

CHAPTER 5

Identifying the features for predicting the types of gene expression induced by Activin

In Chapter 4, we presented ODE models to understand the temporal dynamics of mRNA transcription and degradation. As Activin induced gene expressions can be classified into different clusters depending on their dynamics. We use logistic regression model to identify which kind of gene features (e.g.: epigenetic modifications, Pol II binding, Smad2 binding etc.) can help us to predict the types of mRNA transcription dynamics. We find the most important feature associated with gene expression labels was Pol II occupancy and there is no correlation between mRNA half-life and transcription profile.

5.1 Classification of Activin induced gene expression dynamics

As Activin triggers different temporal patterns of gene expression, we classified Activin regulated genes into two groups according to the FC of mRNA at 8h:

1. Active genes. This kind of gene has been up-regulated by signaling. The genes were selected into this cluster if RNA 8h FC ≥ 1.

2. Repressed genes. This kind of gene has been down-regulated by signaling. The genes were selected into this cluster if RNA 8h FC < 1.

We first focused on 70 genes that have Smad2 binding peaks (Figure 34).

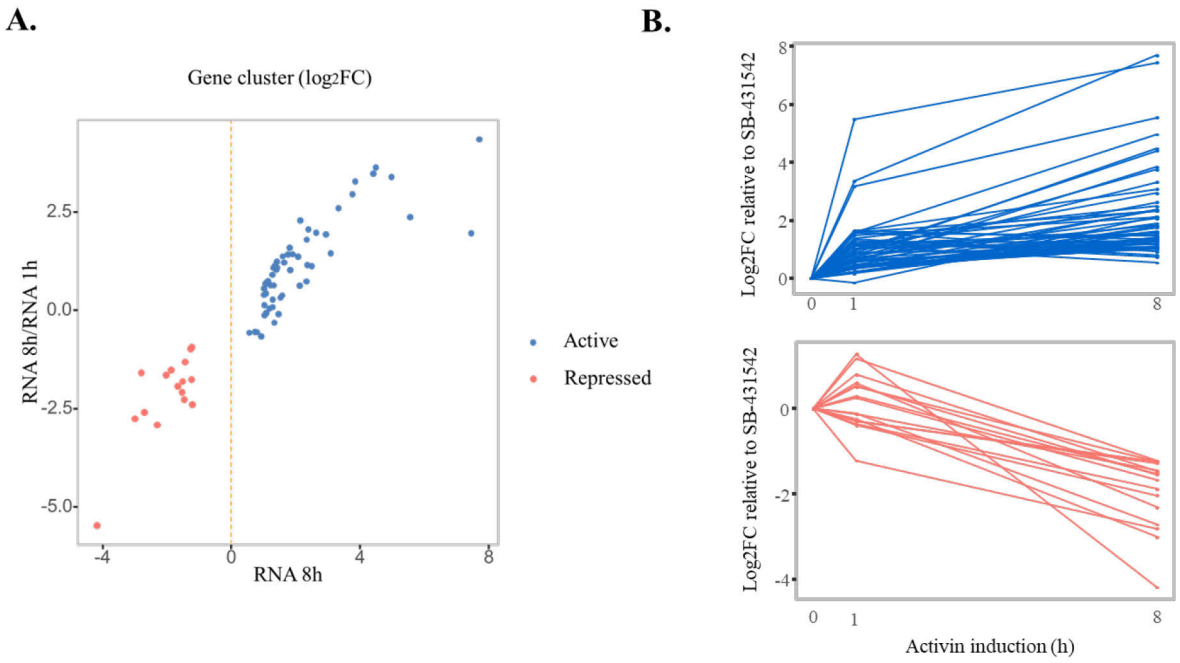


Figure 34 | Gene clusters. A. The 70 target genes with Smad2 binding were classified into the two clusters defined above. Active genes = 54; repressed genes = 16. B. Log2FC relative to SB-431542 at each time point plotted for three clusters.

5.2 Epigenetic features selection

Now based on the sequencing data, we obtained mRNA half-life, Pol II Ser2p, Smad2, H3K9ac and H3K27ac binding profiles for each gene (Table 8). In order to understand which features might determine the classification of the gene into one of the three distinct kinetic clusters, we evaluated these features for predicting the type of gene expression pattern using logistic regression algorithm.

Table 8 | Gene features used for logistic regression modeling.

Gene features	Data		
General transcription regulators	RNA Pol II Ser2p 1h	RNA Pol II Ser2p 4h	RNA Pol II Ser2p 8h
Transcription factors	Smad2 1h	Smad2 8h	
Chromatin modifications	H3K27ac 1h	H3K27ac 4h	H3K27ac 8h
	H3K9ac 1h	H3K9ac 4h	H3K9ac 8h
Others	Half-life time		

One feature or the combination of 2 features were systematically analyzed in this work. For the combination of two features, there may be significant associations between 2 features. This problem is known as collinearity (Miles and Shevlin 2001). In this situation, it can be difficult to have reliable estimates of individual coefficients for variables (Neter, Wasserman et al. 1989, Kang 2013). To avoid this problem in logistic regression modeling analysis, we used variance inflation factors (VIF) to exam whether the presence of collinearity in the selected gene features or not. The VIF for an independent variable x_i defined as

$$VIF_i = \frac{1}{1 - R_i^2} \tag{26}$$

where R_i^2 is the R^2 for a covariate x_i regressed on the remaining covariates in the model. A rule of thumb is that a VIF more than 10 may indicate the presence of collinearity (Marquandt 1980).

The VIFs of gene features are shown in Figure 35. There is a striking correlation between the "H3K27ca FC 4h" and "H3K27ca FC 8h". More than just H3K27ca, H3K9ac, and Pol II are also highly correlated at 4h and 8h time points. This indicates that histone acetylation status of the target genes may have already been in a steady state at 4h. Based on the VIF values, we

excluded the feature combination with VIF more than 10 - combination of "H3K27ca FC 4h" and "H3K27ca FC 8h". In total, we evaluated 12 single-feature and 65 two-feature combinations.

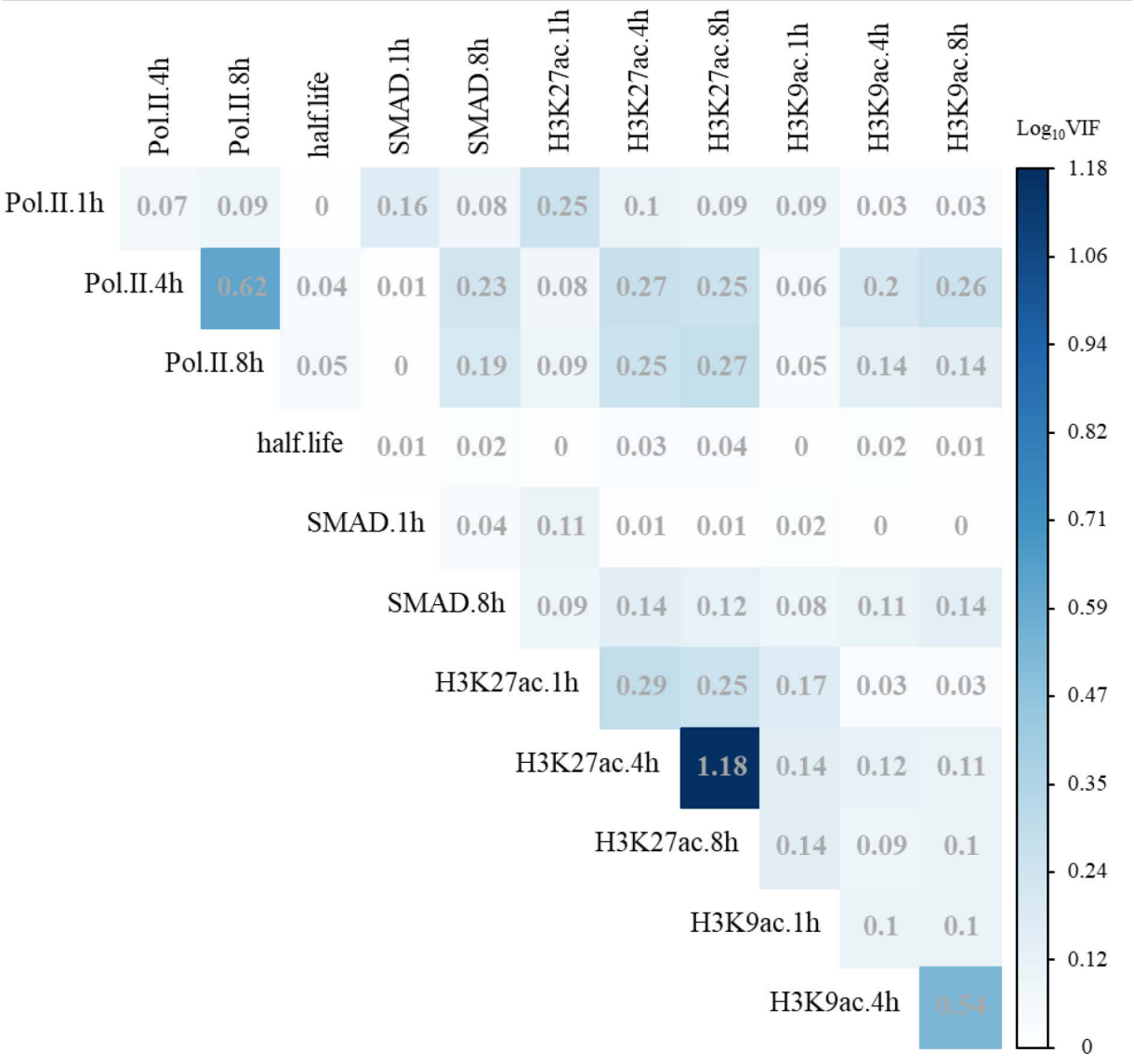


Figure 35 | Variance inflation factors (VIF) for features. The plot shows the log10 VIF based on Spearman's rank correlation coefficient.

5.3 Logistic regression methods

For each feature or combination of 2 features, we use a logistic regression function to predict the type of gene expression. Logistic regression is a process of modeling the probability of a

certain class or event existing given a set of independent variables. Here we applied binary logistic regression, which was developed primarily by Cox and Walker and Duncan to predict two categories (Cox 1958, Walker and Duncan 1967). The training algorithm is described in Figure 36 and elaborated as below:

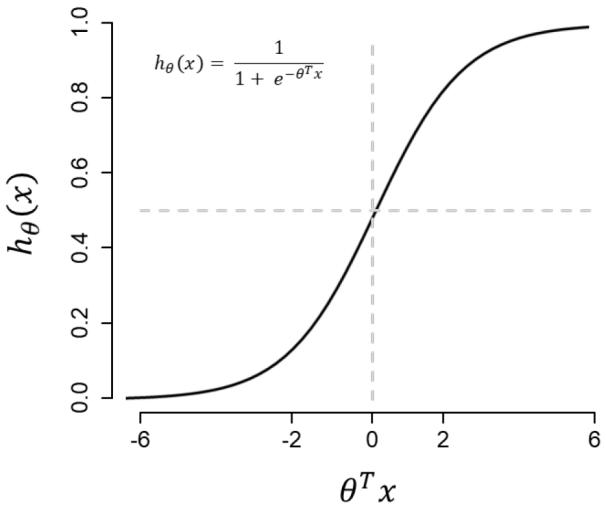


Figure 36 | Logistic function. In the logistic model, $h\theta\ (x)$ is interpreted as the probability of the dependent variable $y = 1$ than $y = 0$.

1. We started by considering the *group* of "repressed genes" as a binary value depending on whether it is classified into the type of expression kinetics. For convenience, we encode 1 or 0. For example, we defined

$$y_i = \begin{cases} 1, & if\ the\ i^{th}\ gene\ is\ classified\ in\ "repressed\ genes" \\ 0, & otherwise. \end{cases}$$

2. For a parameter θ, the logistic regression estimates a probability that $y = 1$ by applying the logistic function:

$$h_\theta(x) = \frac{1}{1 + e^{-\theta^T x}} \tag{27}$$

so $h_\theta\ (x)$ presents the probability of the outcome being 1 and 1 - $h_\theta\ (x)$ presents the probability of the outcome being 0.

3. Given a vector of binary class label $y = (y_1, y_{2,\ldots,} y_m)$, parameter θ can be estimated by maximum likelihood estimation (Menard 2002). For logistic regression, the log-likelihood is

$$Log - likelihood(\theta) = \sum_{i=1}^{m} y_i log h_\theta(x_i) + (1 - y_i) log(1 - h_\theta(x_i)) \quad (28)$$

where m is the sample abundance.

4. The maximum likelihood estimator of θ is applied to make class predictions on test data to measure model performance.

Here, we used *glmfit* function (MATLAB Statistics Toolbox, Mathworks, Natick, MA, USA) for fitting logistic models on the train data. We first separated target genes into three groups randomly. To identify the features that predict the types of target gene expression kinetics and assess the generalization power of features, the logistic regression classifier was trained on each feature or combination of 2 features in a round-robin fashion for cross-validation. Training on two groups and testing on the remaining group, we repeated this procedure three times in a round-robin and evaluated logistic regression classifier performance by following methods (Buggenthin, Buettner et al. 2017) (Figure 37).

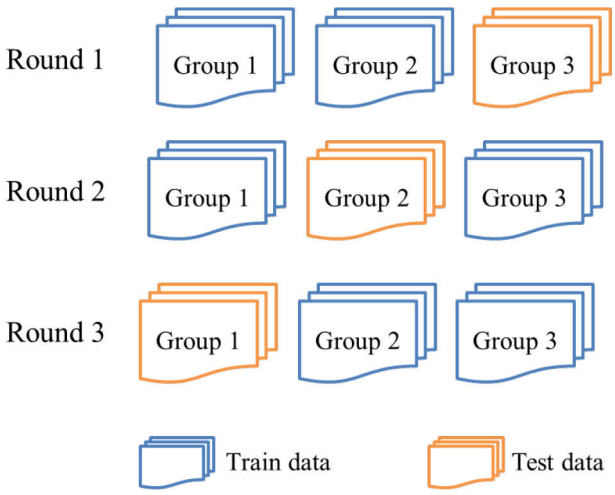


Figure 37 | Schematic of round-robin training and testing. The data was separated into three groups. At each round, two groups would be treated as train data and the third group would be the test data. This procedure will be repeated for three times in a round-robin.

5.4 Quantifying the quality of predictions

. In order to compare the logistic regression models based on different features, we used Cohen's kappa coefficient to quantify the quality of model predictions. The Cohen's kappa is a statistical coefficient that shows the degree of accuracy and reliability in a statistical classification. It measures the degree of agreement among raters (Cohen 1960).

Table 9 | Sample data for counting Cohen's kappa.

A	B		Total
	+	-	
+	*a*	*b*	R_1
_	*c*	*d*	R_2
Total	C_1	C_2	*N*

Suppose we have a dataset that contains only two categories. Each data is evaluated by two methods and each method either clusters "+" or "-" to the data. Suppose the count data is as Table 9, where both A and B are methods, *a* and *d* count the number of agreements, along with *b* and *c* count the number of disagreements. R_1 and R_2 show the number of two categories clustered by A. C_1 and C_2 show the number of two categories clustered by B. *N* is the total number of the dataset.

The definition of Cohen's kappa is:

$$Cohen's\ kappa = \frac{p_0 - p_e}{1 - p_e} = \frac{N(a+d) - (R_1C_1 + R_2C_2)}{N^2 - (R_1C_1 + R_2C_2)} \quad (29)$$

Where p_0 is the ratio of total observed agreement among raters to all observations, and p_e is the agreement that is expected to occur by chance. The Cohen's kappa value can range from -1.00 to +1.00, and its interpretations is shown in Table 10.

Table 10 | Interpretation of the Kappa value according to the reference (Landis and Koch 1977).

Kappa Value	Strength of agreement
< 0	Poor
0–0.20	Slight
0.21–0.40	Fair
0.41–0.60	Moderate
0.61–0.80	Substantial
0.81–1.00	Almost perfect

However, Cohen's kappa has some limitations (Cicchetti and Feinstein 1990). If there is a huge discrepancy between the positive agreement rate (a / N) and the negative agreement rate (d / N), the calculated Cohen's kappa value will be low, as well as the true degree of the agreement will be underestimated. For example, if $a = 98$, $b = c = 1$ and $d = 0$, though the accuracy equals 0.98, the Cohen's kappa will be -0.01 for this result. In addition, when a, d and the sum of b and c are the same, a higher difference between b and c of the higher Cohen's kappa will be produced.

Due to the limitations of Cohen's kappa, we also use other forms of measurement to evaluate the classification results. Because transient induced genes make up more than half of the target genes, the quantity of "accuracy" seems unsuitable for this uneven class distribution. In order to measure the accuracy of the test, we employed the F1 score that is derived from precision and recall for classification results evaluation (Sasaki 2007).

To calculate the F1 score, the precision and the recall of the results are first derived based on a confusion matrix (Table 11).

Table 11 | Four outcomes of a classification result (a 2×2 contingency table or confusion matrix)

True condition / Predicted condition	Positive	Negative
Positive	True positive (TP)	False positive (FP)
Negative	False negative (FP)	True negative (TN)

Precision is the ratio of the correctly predicted positive observations to all of the predicted positive observations.

$$Precision = \frac{TP}{(TP + FP)} \tag{30}$$

Recall is the ratio of the correctly predicted positive observations to the total observations in the positive class.

$$Recall = \frac{TP}{(TP + FN)} \tag{31}$$

The F1 score is defined as the harmonic mean (average) of the precision and recall. The F1 score can range from 0 to 1, with 1 for a best result 1 and 0 for a worst one.

$$F1\ score = (\frac{Recall^{-1} + Precision^{-1}}{2})^{-1} = 2 * \frac{Recall * Precision}{Recall + Precision} \tag{32}$$

The macro-averaged F1 score calculated the performance metrics for each cluster (C), and found their unweighted mean:

$$macro\ F1 = \frac{1}{|C|}\sum_{i=1}^{|C|} F1(C_i) \tag{33}$$

For each feature or combination of 2 features, we reported the macro-averaged F1 score of two classes: active genes and repressed genes.

5.5 Pol II binding features associated with target gene expression patterns

The F1 score and Cohen's kappa for the best five features or combinations are shown below (Table 12) (Appendix C). If only one feature is used, the best three features for predicting gene expression patterns are the same: “Pol II FC 4h”, “Pol II FC 8h” and “H3K27ac FC 8h”, no matter whether F1 score or Cohen’s kappa is used for evaluation. “Pol II FC 4h” received a slightly higher F1 score and Cohen's kappa. While the results of other features like chromatin, modification or transcription factors are quite different from Pol II. For two features, the combination of “Pol II FC 4h” and “H3K9ac FC 1h showed a similar F1 score and Cohen's kappa as other combinations of 2 features. No combinations of 2 features had outstanding better performance compared to the model with “Pol II FC 4h” feature. Therefore, the feature of “Pol II FC 4h” is the best marker to predict the type of gene expression kinetics induced by Activin. The confusion matrices of “Pol II FC 4h” for the three test data are shown in Figure 38. It is worth noting that the model with “mRNA half-life time” feature alone has the worst performance on predicting the type of target gene expression kinetics, which show similar Cohen's kappa score as a random classification method (Appendix D, -0.025). This suggests that mRNA degradation of these genes were not substantially changed by Activin stimulation.

Table 12 | Feature importance for distinguishing clusters.

Feature	F1 score	Kappa	Combination of 2 features			F1 score	Kappa
Pol II FC 4h	0.875	0.750	Pol II FC 4h	and	H3K9ac FC 1h	0.882	0.764
Pol II FC 8h	0.814	0.629	Pol II FC 4h	and	Pol II FC 8h	0.878	0.757
H3K27ac FC 8h	0.750	0.516	Pol II FC 4h	and	Half-life time	0.878	0.755
H3K9ac FC 4h	0.713	0.432	Pol II FC 4h	and	SMAD2 FC 1h	0.875	0.750
H3K9ac FC 8h	0.713	0.432	Pol II FC 4h	and	H3K27ac FC 4h	0.875	0.750

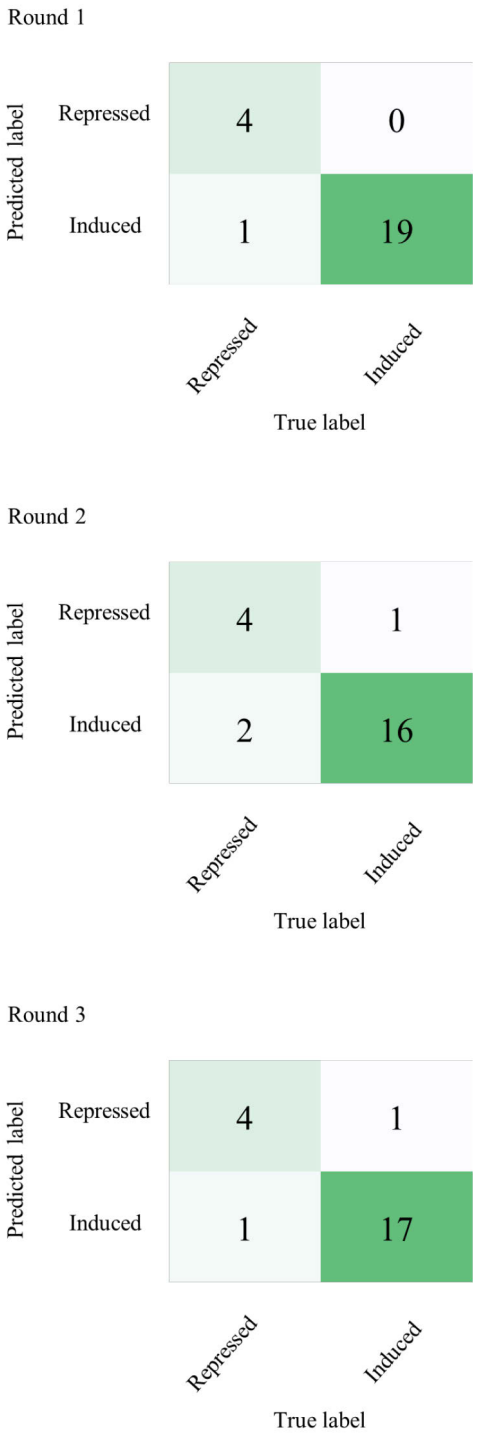


Figure 38 | The confusion matrix of round-robin of "Pol II FC 4h" feature.

5.6 Robustness of gene features for predicting the subcategories of active gene expression with different classification standards

The active genes are further classified into two subcategories. One is "induced sustained". This type of gene has been up-regulated and persists over time. The other is "transient induced". This type of gene has been up-regulated and slowly rises after prolonged signaling. For classification between "induced sustained" and "transient induced", a gene was designated as "induced sustained" if its RNA 8h FC/RNA 1h FC was more than the cutoff. To test how robust gene features can predict the types of target gene expression, we classified the target genes by trying three different cutoff of RNA 8h FC/RNA 1h FC (1.5, 2 and 3) for distinguishing "induced sustained" and "transient induced" genes. The different rules lead to different numbers of genes for two groups (Figure 39).

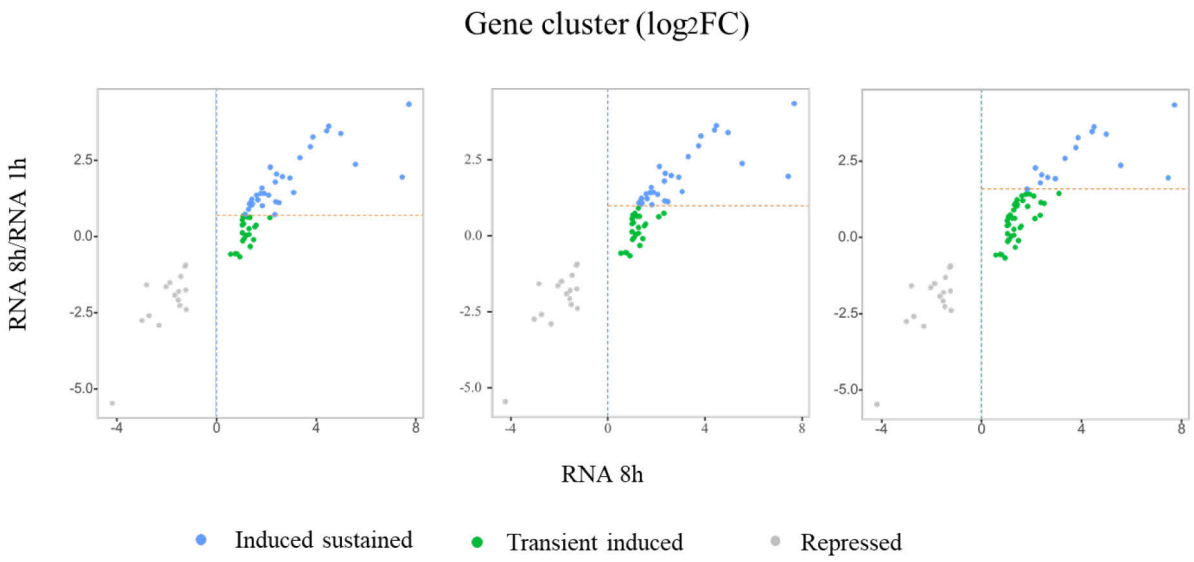


Figure 39 | Subcategories in active genes. Left: Active genes were clustered using RNA 8h FC/ RNA 1h FC filter of 1.5. Induced sustained genes = 37; transient induced genes = 17. Middle: Active genes were clustered using RNA 8h FC/ RNA 1h FC filter of 2. Induced sustained genes = 29; transient induced genes = 25. Right: Active genes were clustered using RNA 8h FC/ RNA 1h FC filter of 3. Induced sustained genes = 15; transient induced genes = 39.

Differently from the analysis by Coda et al. (Coda, Gaarenstroom et al. 2017), the “delayed genes” were not considered because only a few genes would fall into this group and their expression pattern may be categorized into “induced sustained” or “transient induced” group. We next evaluated the performance of logistic regression models with different gene features again (Table 13). Compared with the results of three rules, the top five models with a single feature are similar and the best single feature for predicting the types of active gene expression kinetics is “Pol II FC 8h”. A combination of this feature with Smad2 binding or H3K9ac modification data can slightly increase F1 and/or Kappa scores when using RNA 8h FC/ RNA 1h FC filter of 1.5 or 3. The combinations of Smad2 binding and H3K27ac modification data showed a similar F1 score and Cohen's kappa as other combinations of 2 features when using RNA 8h FC/ RNA 1h FC filter of 2.

Table 13 | Feature importance for distinguishing active genes.

FC filter	Feature	F1 score	Kappa	Combination of 2 features		F1 score	Kappa
	Pol II FC 8h	0.693	0.415	Pol II FC 8h	and Smad2 FC 8h	0.732	0.503
	SMAD2 FC 8h	0.676	0.365	Pol II FC 8h	and Smad2 FC 1h	0.726	0.418
1.5	H3K27ac FC 8h	0.560	0.219	H3K27ac FC 8h	and H3K9ac FC 4h	0.713	0.271
	H3K9ac FC 4h	0.546	0.159	Smad2 FC 8h	and H3K9ac FC 4h	0.702	0.372
	Pol II FC 4h	0.532	0.156	Half-life time	and Smad2 FC 8h	0.702	0.294
	Pol II FC 8h	0.699	0.415	Smad2 FC 8h	and H3K27ac FC 4h	0.766	0.536
	H3K27ac FC 4h	0.684	0.387	Smad2 FC 8h	and H3K27ac FC 8h	0.749	0.502
2	SMAD2 FC 8h	0.678	0.372	Pol II FC 8h	and Smad2 FC 8h	0.748	0.503
	H3K27ac FC 8h	0.656	0.318	Smad2 FC 1h	and H3K27ac FC 8h	0.735	0.474
	Pol II FC 4h	0.620	0.282	Pol II FC 1h	and Pol II FC 8h	0.734	0.482
	Pol II FC 8h	0.796	0.598	Pol II FC 8h	and H3K9ac FC 4h	0.830	0.671
	Pol II FC 4h	0.714	0.436	Pol II FC 8h	and H3K9ac FC 8h	0.830	0.671
3	H3K27ac FC 8h	0.703	0.450	Pol II FC 1h	and Pol II FC 8h	0.811	0.625
	SMAD2 FC 8h	0.695	0.398	Pol II FC 8h	and H3K27ac FC 1h	0.791	0.588
	H3K27ac FC 4h	0.686	0.452	Pol II FC 4h	and H3K9ac FC 4h	0.777	0.561

The rules for the classification of target genes (Figure 39) is not objective as the cutoffs are set arbitrary. Instead of setting cutoff, we applied k-means to create subcategories for active genes (Forgy 1965, MacQueen 1967, Hartigan and Wong 1979, Lloyd 1982) (Figure 40).

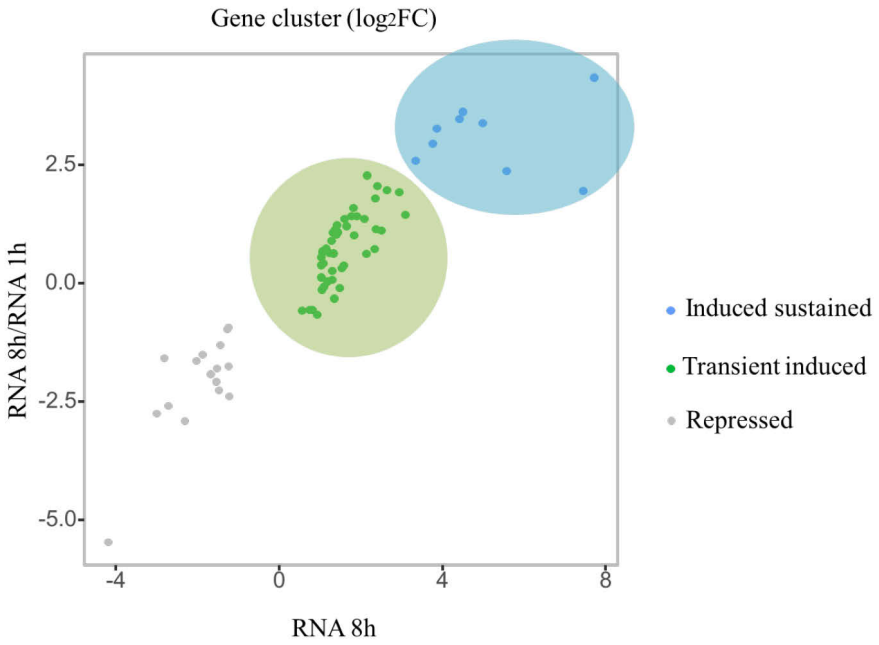


Figure 40 | Subcategories in active genes by applying k-means. Active genes were clustered using k-means. Induced sustained genes = 9; transient induced genes = 45.

For subcategories generated by k-means, the "H3K27ac FC 8h" is the best single feature to predict the types of active gene expression kinetics, followed by "Pol II FC 8h" (Table 14). Combinations of "H3K27ac FC 8h" with H3K9ac modification data can slightly increase F1 and Kappa scores. According to all of the above results, it seems that good predictions to distinguish active genes can be achieved at different cutoffs or with k-means clustering algorithm. As the results, the subcategories of active genes can be well predicted from gene features.

Table 14 | Feature importance for distinguishing active genes by applying k-means.

Feature	F1 score	Kappa	Combination of 2 features			F1 score	Kappa
H3K27ac FC 8h	0.884	0.775	H3K27ac FC 8h	and	H3K9ac FC 1h	0.884	0.775
Pol II FC 8h	0.772	0.551	H3K27ac FC 8h	and	H3K9ac FC 4h	0.884	0.775
H3K27ac FC 4h	0.669	0.354	H3K27ac FC 8h	and	H3K9ac FC 8h	0.865	0.736
Pol II FC 4h	0.582	0.182	H3K27ac FC 1h	and	H3K27ac FC 8h	0.862	0.726
H3K27ac FC 1h	0.573	0.186	Pol II FC 4h	and	H3K27ac FC 8h	0.860	0.726

We also applied method for predicting active and repressed genes to analyze all the differentially-regulated genes by Activin without Smad2 binding. However, none of the features is good for predicting the type of gene expression kinetics (Table 15).

Table 15 | Feature importance for clusters of differentially-regulated genes.

Feature	F1 score	Kappa	Combination of 2 features			F1 score	Kappa
Half-life time	0.628	0.290	Pol II FC 4h	and	Half-life time	0.639	0.306
H3K9ac FC 8h	0.506	0.080	Pol II FC 1h	and	Half-life time	0.628	0.290
H3K9ac FC 4h	0.492	0.060	Pol II FC 8h	and	Half-life time	0.628	0.290
Pol II FC 4h	0.475	0.027	Half-life time	and	H3K27ac FC 1h	0.628	0.290
Pol II FC 1h	0.450	0.000	Half-life time	and	H3K9ac FC 1h	0.628	0.290

The highest Cohen's kappa is about 0.31 for the combination of 2 features of “Pol II FC 4h” and “Half-life time”, which indicate that it is not stronger than expected result from random

classification. We noticed that the majority of 128 differentially-regulated genes are classified into the group of "active" genes (82.03%) (Figure 41).

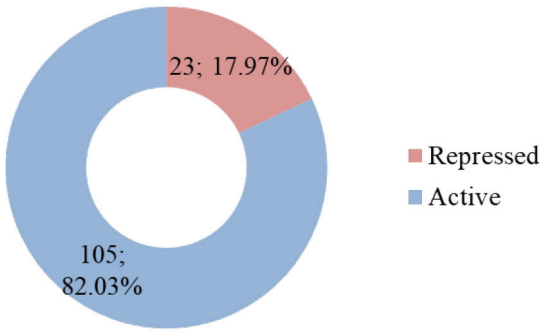


Figure 41 | Number and percentage of gene clusters for differentially-regulated genes.

The extremely imbalanced gene cluster data might affect the performance of classification predictions made from logistic regression models. To address this issue, we tried an over-sampling approach to deal with the extremely imbalanced data. Random oversampling simply randomly choose the minority class of the training data and copy them until the two classes had the same number of data (Ling and Li 1998). However, random oversampling had almost no improvement in the results for predicting the types of target genes: the best F1 score increased from 0.64 to 0.68, as well as the Cohen's kappa increased from 0.31 to 0.37 at the same time (Table 16).

Table 16 | Feature importance for clusters of differentially-regulated genes (oversampling).

Feature	F1 score	Kappa	Combination of 2 features			F1 score	Kappa
H3K9ac FC 8h	0.663	0.342	Half-life time	and	H3K9ac FC 8h	0.682	0.373
H3K27ac FC 4h	0.612	0.290	Pol II FC 1h	and	H3K9ac FC 8h	0.676	0.365
H3K9ac FC 4h	0.586	0.197	Pol II FC 4h	and	H3K9ac FC 8h	0.670	0.355
H3K27ac FC 8h	0.583	0.239	Half-life time	and	H3K27ac FC 8h	0.666	0.350
Half-life time	0.558	0.121	H3K9ac FC 1h	and	H3K9ac FC 8h	0.662	0.334

Taken together, our results indicate that the dynamics of target genes without Smad2 binding cannot be well predicted from histone modifications and or pol II binding features. As TGF-β superfamily signals often crosstalk to other signaling pathways and recruit the binding of other transcriptional factors. Therefore, the prediction of these genes maybe improved by including the binding profiles of specific enhances or repressors for the target genes.

CHAPTER 6

Conclusion & Discussion

Cell signaling networks are complex and dynamic systems that ensure cells sense and respond to their environment. The TGF-β is one of the most important signaling events. It regulates many different cellular processes such as cell proliferation, migration, and death (Shi and Massague 2003). The Nodal/Activin is a member of the TGF-β superfamily, which is involved in many cellular processes. Abnormal activity of the Nodal/Activin has been connected to human diseases such as cancers (Pauklin and Vallier 2015).

In this study, we used mathematical modeling approaches to study transcriptional responses to Activin signal. As a result, the mathematical models helped us to understand how Activin signal induces different kinetic profiles of gene expression. The results suggest that Activin regulates target genes through different mechanisms. In addition, using logistic regression models, we found that the kinetic patterns of the genes bound with Smad2 can be well predicted by their Pol II binding profiles.

6.1 Smad2 activities linked to TGF-β induced transcription regulation

It is known that Smads play important roles in regulating downstream responses of TGF-β signaling. Smads are thought to interact with a number of co-activators and co-repressors to induce chromatin remodeling. Ross et al. show that Smad2-mediated transcription requires specific histone modifications and chromatin remodeling (Ross, Cheung et al. 2006). The repression of Runx2 function by Smad3/HDAC complex depended on the regulation of the histone acetylation levels guided by Smad3 binding specificity (Kang, Alliston et al. 2005). These findings suggest that Smads regulate gene responses by regulating Pol II activities.

However, our model selection results indicate that Smad2 regulates target genes not only through the recruitment of Pol II transcription machinery, but also through participation in the degradation for some genes. For example, the top model of Dusp4 and Enpp2 is model5, which indicates that Smad2 binding level promotes mRNA degradation rates. Through the model selection, we analyzed the regulation of each gene. The results produced by some of these genes are consistent with previous literature or experiments.

6.2 Signaling crosstalk between Activin and other signaling pathways

TGF-β super family signal talks with other pathways at many development stages. The interaction between TGF- β and Wnt pathway has been known for a long time (Attisano and Labbe 2004). Our results show that Wnt3 and Wnt11, which are best explained by model3 and model7, are regulated by Activin/Nodal signaling via Smad2 (Figure 30). In additional, Msx2 genes that are induced by Wnt and BMP pathways, play a crucial role in neural development (Willert, Epping et al. 2002). These genes also show strong FC after Activin/Nodal stimulation, which indicate that other pathway regulators could contribute to the control of target gene expression.

6.3 Correlation between gene features and the dynamics of transcription responses

Degradation is an important step for mRNA stability regulation. Previous studies have suggested that mRNA degradation plays an important role in the regulation of gene expression by phosphatidylinositol 3-kinase signaling (Graham, Hendershott et al. 2010). In some of our kinetic models, we assume that mRNA degradation rate is not changed after Activin stimulation. It may be unrealistic, but the 'constant degradation' model reportedly can fit the majority of genes well (Rabani, Levin et al. 2011). By comparing different models, we found that Smad2 is possibly involved in regulating mRNA degradation of a few target genes. However, when we look at 70 genes that are bound with Smad2 after Activin stimulation, we found the mRNA half-life time is a poor feature for predicting the dynamics of mRNA

transcription. In contrast, the binding profile of Pol II Ser2P at 4h or 8h is a very good feature for predicting the types of transcription dynamics for genes with Smad2 binding profile. It is interesting that this feature is not good for the prediction of other genes that are not bound with Smad2. This indicate that other transcriptional factors (co-activator or co-repressor) might play a more important role in shaping the dynamics of these genes. Future works are needed to elucidate the mechanism for controlling the transcriptional dynamics of these genes.

6.4 The limitations of this work

In this work, we aimed to understand the regulation of transcription dynamics for all Activin-regulated genes (in total 198 genes). However, we found that the kinetic models or logistic regression models could explain well only for the group of target genes that have Smad2 binding activities. The insights from this study might be due to the following limitations.

Limited samples in RNA-seq data

The RNA-seq data for gene expression are only available at 1h and 8h after Activin stimulation. Due to the limited samples, the change of some genes may not be detected. For example, the jun proto-oncogene (c-jun) has been reported to be decreased very rapidly after an initial rise (Mauviel, Chung et al. 1996) (Figure 42A). In our data set, the dynamics of c-jun is reduced due to lack of early sampling point. In contrast, the Pol II activity shows an increase at the beginning and then a decrease (Figure 42B). That may explain why our model could not explain the dynamics of c-jun transcription data.

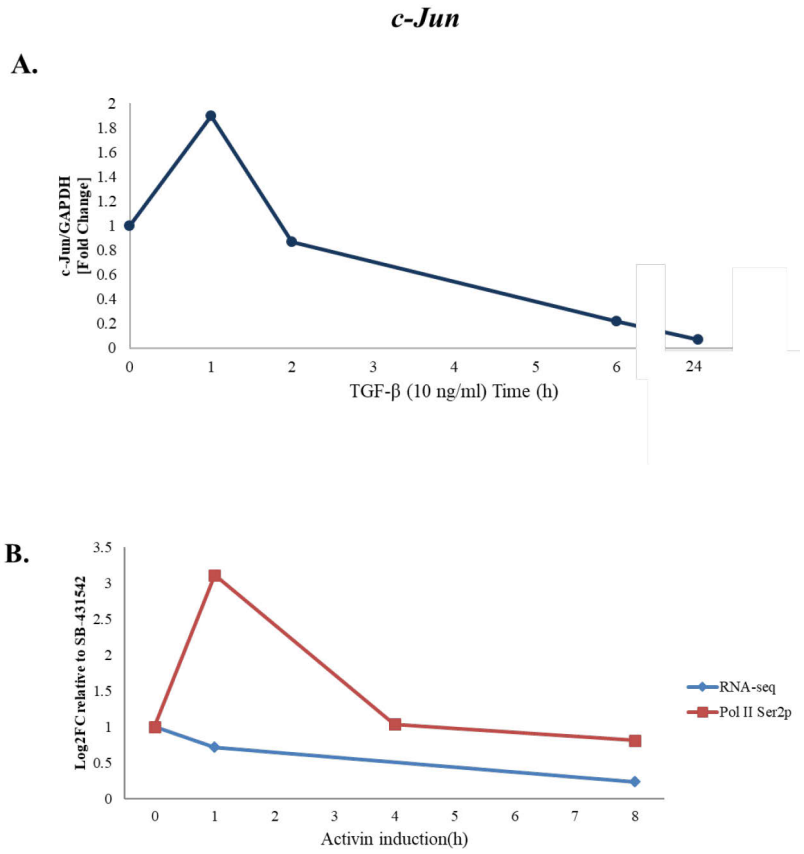


Figure 42 | Expression of *c-jun*. A. A rough plot for the original data shown in the paper of Mauviel 1996. B. Transcription dynamic and Pol II activity of *c-jun* via sequencing.

The lack of Smad3 activities

Nodal/Activin signaling pathway is mediated specially by Smad2/3 and Smad4. In this study, as we only have the ChIP-seq data for Smad2, Smad3 binding activities for the target genes were not considered. Although the western blot data we used did not detect strong Smad3 signal, the level of phosphor-Smad3 proteins in P19 cells has been reported to have increased after treatment with Activin (Mazerbourg, Klein et al. 2004). Due to the lack of available Smad3 binding data, Smad3-depedent genes may not be well predicted by the models present in this study.

The limitations of annotation

Most of the recent studies have demonstrated that 1-20% of SMAD2 binding sites are involved in the regulation of nearby gene expression (Morikawa, Koinuma et al. 2013). The level of mature mRNA is regulated by a variety of processes, such as splicing and degradation.

Previous studies has indicated that Smads binding sites can be more than 50 kb away from the TSS of target genes (Landry, Bonadies et al. 2009, Morikawa, Koinuma et al. 2011, Trompouki, Bowman et al. 2011). Although we assigned Smad2 binding sites to the nearest genes within 100 kb, only 70 of the 198 differentially-regulated genes were annotated with Smad2 binding. In addition to the lack of data for Smad3 binding, Activin stimulation may induce certain TFs that are involved in a feedforward regulatory loop and cooperatively regulate gene expression, especially at late time points (Yan, Xiong et al. 2017).

6.5 Future directions

Future work might be able to address above-mentioned limitations. First, in order to accurately identify the target genes of Smad-binding sites, the chromatin architecture needs to be well characterized. Chromosome conformation capture approaches such as Chromosome conformation capture (3C), Chromosome conformation capture-on-chip (4C), Chromosome conformation capture carbon copy (5C) or HI-C, make it possible to analyze the spatial organization of chromatin in a cell. Application of these technologies will help to capture the Smads binding activities for target genes more accurately and completely. In addition, more time points can improve the resolution of transcriptional dynamics for the target genes.

In order to identify the features for predicting the type of transcriptional dynamics, we could analyze more features such as other sequence-specific TFs, structural proteins, inactivation chromatin modifications and genomic elements etc. when such kind of datasets are available in the future.

Appendix

A. The AICc values of the models.

The genes that are highlighted in red cannot be fitted into any of the models. The last column shows the best model based on the AICc value.

Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Best Model
4930447C04Rik	59.0	59.1	30.1	66.9	66.2	60.3	44.8	73.4	73.7	73.5	m3
Ago1	-11.8	-23.9	-166.0	-3.8	-3.8	-114.0	-33.4	4.2	5.3	2.8	m3
Amot	38.9	-69.6	-99.9	46.9	34.4	38.7	-62.8	54.9	56.1	55.5	m3
Angpt1	6.9	11.0	-152.4	14.8	14.7	-13.7	-37.2	21.3	21.4	20.4	m3
Asb4	-27.1	-34.8	-102.0	-19.2	-30.4	-29.2	-69.7	-12.6	-12.1	-12.2	m3
B630005N14Rik	9.2	12.6	-82.2	17.1	17.2	-26.0	3.6	21.9	21.9	20.1	m3
Bmpr2	-38.5	-100.5	-161.5	-30.5	-31.4	-126.6	-107.2	-17.6	-13.5	-16.9	m3
Camk2n1	-3.2	-14.4	-22.5	4.8	4.8	-9.5	-25.1	13.8	15.5	13.9	m7
Ccno	4.9	-6.5	-11.0	12.8	12.8	-3.3	0.2	20.7	22.3	20.8	m3
Cnpy1	36.7	27.8	-57.0	44.6	44.5	13.4	-58.3	51.3	52.0	50.9	m7
Cyr61	-81.0	-128.1	-73.1	-98.3	-93.8	-90.9	-119.8	-89.5	-84.2	-95.1	m2
Dact1	-15.0	-12.0	-148.0	-7.0	-7.0	-92.5	-13.7	0.4	0.6	-2.6	m3
Dhrs3	89.8	82.1	-90.6	97.7	96.9	84.6	-12.5	104.3	104.9	104.2	m3
Dusp4	-112.1	-145.6	-104.2	-169.0	-174.6	-103.8	-136.1	-67.7	-61.7	-67.0	m5
Dusp5	-122.6	-130.9	-124.2	-114.6	-114.3	-138.1	-135.1	-89.7	-84.7	-94.6	m6
Enpp2	-71.2	-66.2	-63.2	-65.4	-103.2	-63.2	-56.9	-57.0	-56.9	-56.7	m5
Eomes	136.8	49.2	-68.1	144.7	143.5	133.9	-18.6	151.7	152.8	75.5	m3
Epha2	64.0	51.5	-29.1	72.0	72.0	-20.7	-48.2	82.9	84.8	80.6	m7

Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Best Model
Fgf15	60.7	38.6	-66.6	68.7	68.7	28.9	-51.2	76.9	77.6	75.9	m3
Fgf8	109.6	103.9	-40.7	117.5	116.8	108.4	43.5	124.1	124.5	124.0	m3
Fzd10	10.3	5.5	-54.0	18.3	18.4	-48.4	-15.2	31.7	31.8	27.3	m3
Gadd45g	24.5	24.0	-38.1	32.5	32.5	12.7	-22.1	33.3	33.4	32.7	m3
Gata3	-40.1	-54.3	-94.7	-32.1	-32.1	-73.4	-14.5	-20.5	-20.4	-23.8	m3
Gbx2	-54.9	-50.1	-47.0	-53.0	-54.5	-46.8	-40.8	-39.9	-40.8	-39.0	m1
Gm20544	66.2	42.6	-59.7	74.2	74.6	47.6	-24.8	81.2	82.0	80.9	m3
Gm26562	28.6	31.4	-18.3	36.6	35.5	19.0	-3.1	43.1	43.2	42.6	m3
Gm26793	65.2	67.8	49.9	73.2	73.2	67.4	65.5	77.1	77.1	76.8	m3
Grsf1	-60.3	-207.2	-196.1	-52.3	-105.1	-144.9	-130.0	-16.2	-10.1	-13.6	m2
Gsc	198.1	196.1	54.4	206.0	205.8	203.0	73.3	212.4	212.7	205.0	m3
Gt(ROSA)26Sor	41.2	42.9	-0.9	49.2	49.2	33.3	27.7	52.2	52.2	51.5	m3
Has2	-23.0	-107.6	-166.1	-15.0	-15.4	-114.5	-112.8	-3.0	1.0	-2.8	m3
Hes1	31.5	37.7	-36.7	39.5	39.5	-19.2	8.2	47.0	47.0	45.6	m3
Hist1h1c	-15.1	-21.9	-7.1	-24.5	-89.0	-7.0	-12.8	-0.8	0.6	0.5	m5
Hs6st2	10.2	9.4	-185.7	18.2	16.3	-39.6	-24.2	24.7	25.1	23.8	m3
Id1	-127.6	-123.1	-100.5	-119.6	-119.6	-121.2	-113.5	-72.7	-70.0	-71.8	m1
Id2	-135.7	-157.6	-158.6	-127.8	-127.2	-163.5	-148.3	-78.8	-73.7	-82.4	m6
Id3	-119.8	-121.3	-111.8	-160.1	-156.6	-111.4	-111.8	-59.3	-55.8	-61.2	m4
Il17rd	-6.5	-47.0	-152.4	1.5	1.1	-122.1	-84.4	17.0	21.4	17.0	m3
Irgm1	17.8	-7.1	-150.7	25.7	24.8	-70.3	-51.5	32.8	34.2	32.2	m3
Itgb8	24.2	-13.1	-60.4	32.2	32.0	-11.3	-25.7	39.5	41.1	39.1	m3
Kcnj3	-34.8	-95.1	-96.0	-26.8	-32.1	-66.4	-48.3	-17.3	-14.6	-16.6	m3
Klf6	-113.0	-140.7	-105.1	-155.5	-155.8	-104.7	-131.3	-54.9	-49.9	-57.2	m5

Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Best Model
Lefty1	283.2	126.1	135.0	291.2	291.1	249.2	100.2	144.3	147.1	105.7	m7
Lefty2	283.1	204.7	88.8	291.1	290.8	290.2	75.5	214.1	214.1	152.2	m7
Lgr4	-16.9	-67.2	-164.3	-9.0	-10.0	-151.4	-33.3	5.2	9.4	5.5	m3
Msx2	-77.4	-76.0	-114.8	-69.4	-69.3	-117.0	-87.3	-64.5	-64.2	-80.8	m6
Naa30	23.8	26.6	31.8	23.3	31.8	29.5	31.7	35.8	35.9	35.9	m4
Nanog	56.6	34.1	-116.6	64.6	64.4	35.0	-64.3	71.6	72.5	71.4	m3
Nodal	4.2	-68.5	-56.4	12.2	12.2	-68.3	-66.0	41.3	47.2	37.5	m2
Pak3	-10.9	-29.4	-135.6	-3.0	-5.4	-69.7	-44.1	4.0	5.3	3.6	m3
Pde4b	33.9	32.2	-129.6	41.9	41.6	-29.9	-10.2	48.4	49.0	47.3	m3
Pitx2	51.3	-17.3	-29.5	59.3	56.9	-95.2	-35.1	83.4	88.1	12.5	m6
Pkdcc	54.8	27.6	-117.3	62.8	62.5	11.6	-93.1	70.5	72.0	70.1	m3
Plagl1	25.2	-0.5	-110.5	33.2	31.8	18.5	-78.2	40.1	40.9	40.4	m3
Pmepa1	162.9	146.1	48.0	170.8	170.8	139.6	5.3	177.6	179.0	123.6	m7
Prdm1	-25.5	-42.5	-134.5	-17.5	-17.7	-132.7	-30.7	-9.9	-8.2	-12.2	m3
Sh3bgrl	29.2	33.4	20.6	37.2	37.2	33.4	33.8	42.7	42.7	42.5	m3
Ski	-22.8	-35.6	-174.8	-14.9	-14.9	-133.6	-28.1	-5.4	-3.1	-6.4	m3
Snai1	-66.1	-64.1	-77.0	-58.1	-58.0	-82.8	-57.8	-54.1	-54.0	-54.6	m6
Sp5	44.0	-4.1	-5.4	52.0	49.9	22.6	3.0	62.7	65.5	61.6	m3
Steap2	-11.4	-40.6	-179.3	-3.4	-4.0	-49.1	-24.8	4.1	5.6	4.2	m3
T	109.9	-7.6	-44.6	117.9	112.9	113.5	-36.2	125.4	126.4	120.8	m3
Tdgf1	140.3	-8.7	-51.9	148.3	145.2	144.1	-36.6	155.4	156.1	37.3	m3
Tdh	-75.6	-110.5	-67.7	-147.3	-225.7	-67.6	-101.3	-60.4	-57.0	-57.0	m5
Trh	4.1	-136.5	-165.7	12.0	-15.7	-30.9	-64.4	31.6	35.2	32.7	m3
Ubr7	-29.8	-95.0	-181.7	-21.8	-27.3	-131.4	-62.8	-7.2	-3.4	-5.4	m3

Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Best Model
Wnt11	41.5	20.0	-121.5	49.4	49.2	29.3	-46.3	56.7	57.7	57.0	m3
Wnt3	90.2	74.3	-42.5	98.2	98.2	68.1	-56.3	105.9	106.9	104.8	m7
Ypel2	-15.8	-33.2	-96.6	-7.9	-9.3	-57.0	-44.7	-0.9	0.4	-1.5	m3
Zfp947	43.4	40.3	-2.9	51.4	51.4	30.4	28.4	49.6	49.6	48.6	m3

B. The AICc weights of the models.

The genes that are highlighted in red cannot be fitted into any of the models. For the fitted genes, if the AICc weight of any Smad2 independent models (model1 and model 2) has a value greater than 0.05, this gene may not be Smad2 dependent. And these Smad2 independent genes are shown in blue background.

Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Smad2 Dependent
4930447C04Rik	0.000	0.000	0.999	0.000	0.000	0.000	0.001	0.000	0.000	0.000	-
Ago1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Amot	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Angpt1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Asb4	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
B630005N14Rik	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Bmpr2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Camk2n1	0.000	0.004	0.207	0.000	0.000	0.000	0.789	0.000	0.000	0.000	Y
Ccno	0.000	0.094	0.884	0.000	0.000	0.019	0.003	0.000	0.000	0.000	-
Cnpy1	0.000	0.000	0.343	0.000	0.000	0.000	0.657	0.000	0.000	0.000	Y
Cyr61	0.000	0.984	0.000	0.000	0.000	0.000	0.016	0.000	0.000	0.000	N
Dact1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Dhrs3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Dusp4	0.000	0.000	0.000	0.058	0.942	0.000	0.000	0.000	0.000	0.000	Y
Dusp5	0.000	0.022	0.001	0.000	0.000	0.800	0.176	0.000	0.000	0.000	Y
Enpp2	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	Y
Eomes	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Epha2	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	Y
Fgf15	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y

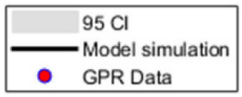
Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Smad2 Dependent
Fgf8	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Fzd10	0.000	0.000	0.942	0.000	0.000	0.058	0.000	0.000	0.000	0.000	Y
Gadd45g	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-
Gata3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Gbx2	0.437	0.039	0.008	0.163	0.343	0.007	0.000	0.000	0.000	0.000	-
Gm20544	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Gm26562	0.000	0.000	0.999	0.000	0.000	0.000	0.001	0.000	0.000	0.000	-
Gm26793	0.000	0.000	0.999	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-
Grsf1	0.000	0.996	0.004	0.000	0.000	0.000	0.000	0.000	0.000	0.000	N
Gsc	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Gt(ROSA)26Sor	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-
Has2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Hes1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Hist1h1c	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	Y
Hs6st2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Id1	0.846	0.087	0.000	0.016	0.016	0.035	0.001	0.000	0.000	0.000	N
Id2	0.000	0.045	0.075	0.000	0.000	0.880	0.000	0.000	0.000	0.000	Y
Id3	0.000	0.000	0.000	0.853	0.147	0.000	0.000	0.000	0.000	0.000	Y
Il17rd	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Irgm1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Itgb8	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Kcnj3	0.000	0.398	0.602	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y/N
Klf6	0.000	0.000	0.000	0.466	0.534	0.000	0.000	0.000	0.000	0.000	Y
Lefty1	0.000	0.000	0.000	0.000	0.000	0.000	0.941	0.000	0.000	0.059	Y

Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Smad2 Dependent
Lefty2	0.000	0.000	0.001	0.000	0.000	0.000	0.999	0.000	0.000	0.000	Y
Lgr4	0.000	0.000	0.998	0.000	0.000	0.002	0.000	0.000	0.000	0.000	Y
Msx2	0.000	0.000	0.253	0.000	0.000	0.747	0.000	0.000	0.000	0.000	Y
Naa30	0.377	0.093	0.007	0.485	0.007	0.021	0.007	0.001	0.001	0.001	-
Nanog	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Nodal	0.000	0.459	0.001	0.000	0.000	0.409	0.131	0.000	0.000	0.000	Y/N
Pak3	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Pde4b	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Pitx2	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	Y
Pkdcc	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Plagl1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Pmepa1	0.000	0.000	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	Y
Prdm1	0.000	0.000	0.707	0.000	0.000	0.293	0.000	0.000	0.000	0.000	Y
Sh3bgrl	0.013	0.002	0.982	0.000	0.000	0.002	0.001	0.000	0.000	0.000	-
Ski	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Snai1	0.000	0.000	0.050	0.000	0.000	0.950	0.000	0.000	0.000	0.000	Y
Sp5	0.000	0.341	0.649	0.000	0.000	0.000	0.010	0.000	0.000	0.000	-
Steap2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
T	0.000	0.000	0.985	0.000	0.000	0.000	0.015	0.000	0.000	0.000	Y
Tdgf1	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Tdh	0.000	0.000	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	Y
Trh	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Ubr7	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Wnt11	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y

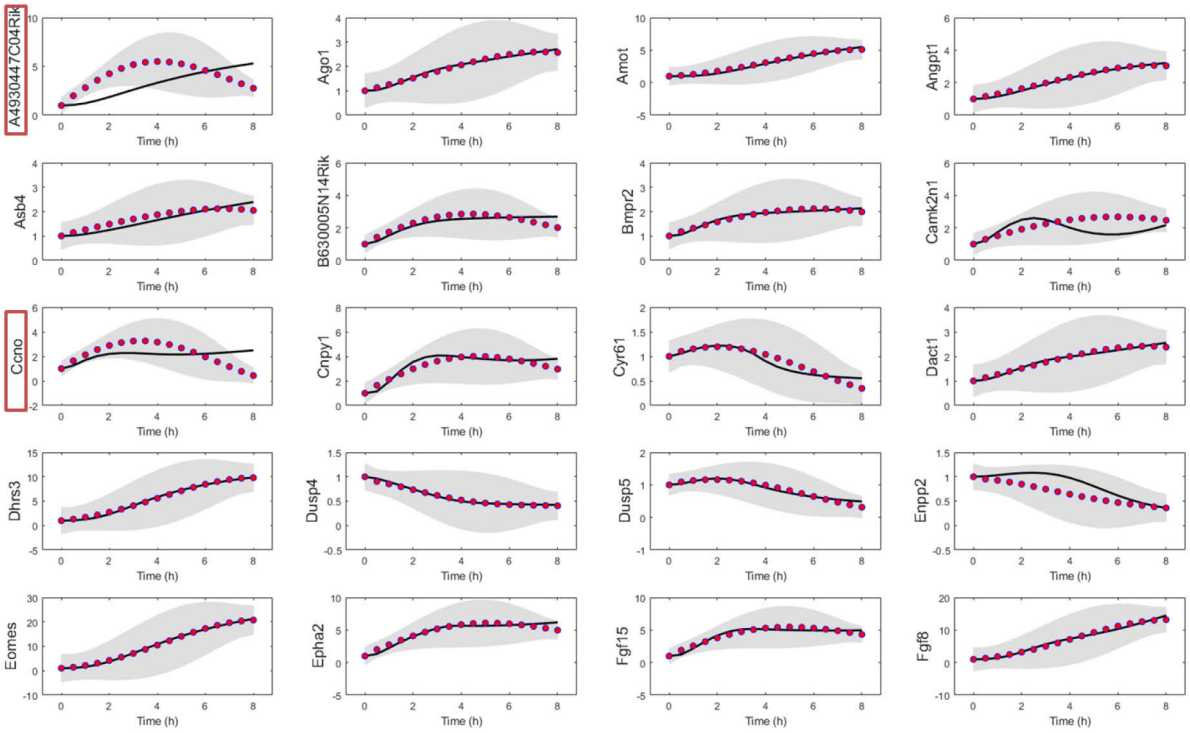
Gene	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6	Model 7	Model 8	Model 9	Model 10	Smad2 Dependent
Wnt3	0.000	0.000	0.001	0.000	0.000	0.000	0.999	0.000	0.000	0.000	Y
Ypel2	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	Y
Zfp947	0.000	0.000	1.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000	-

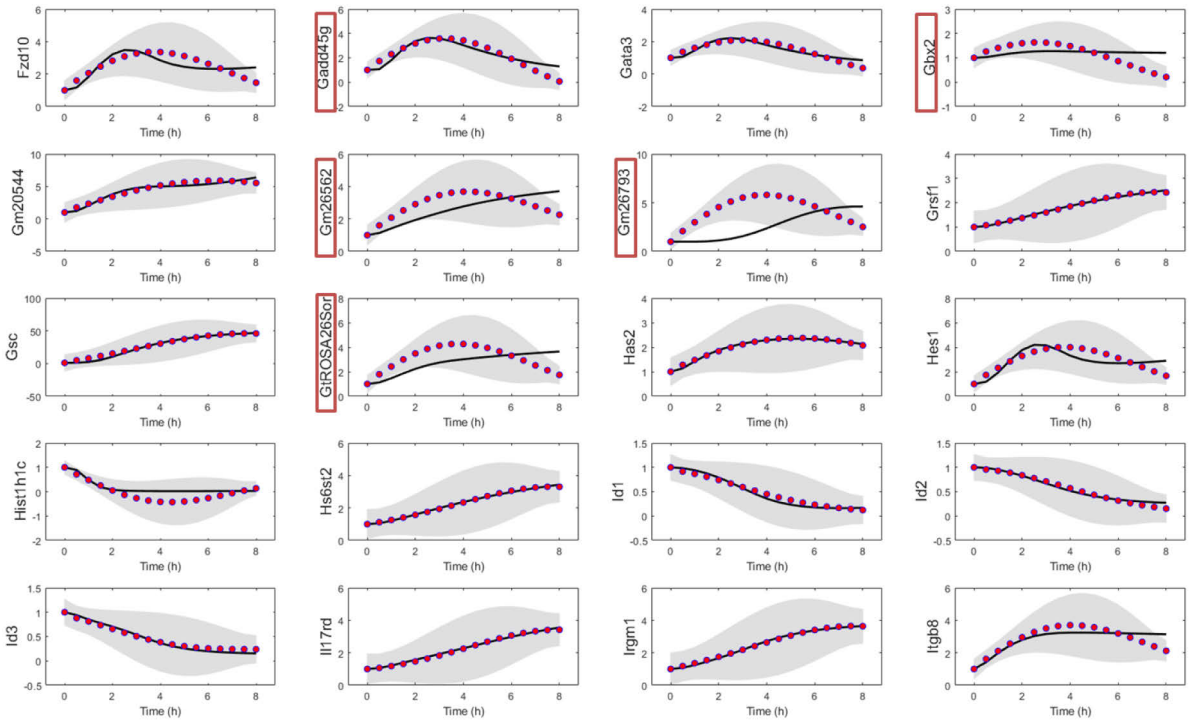
C. Fitting of the best models to mRNA datasets

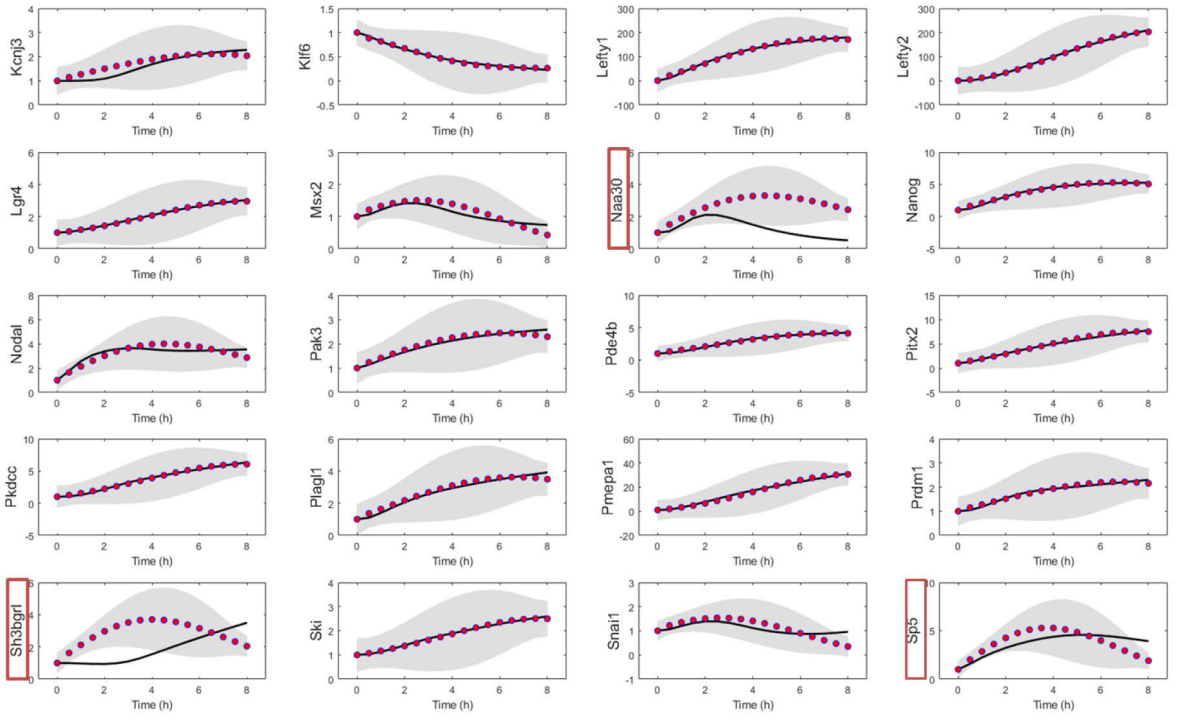
According to the chi-square test, the unfitted data are shown with the red frame. The points show the GPR data. The 95% confidence intervals predicted from GPR model are shown as shaded regions. The simulation from the best model for each gene is shown in a black line.

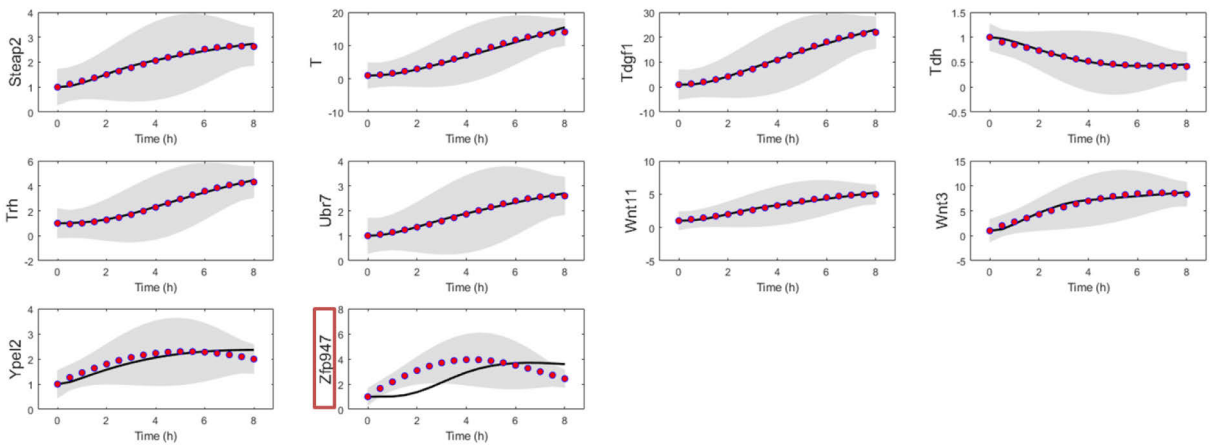


A. Activin/Nodal simulation mRNA FC

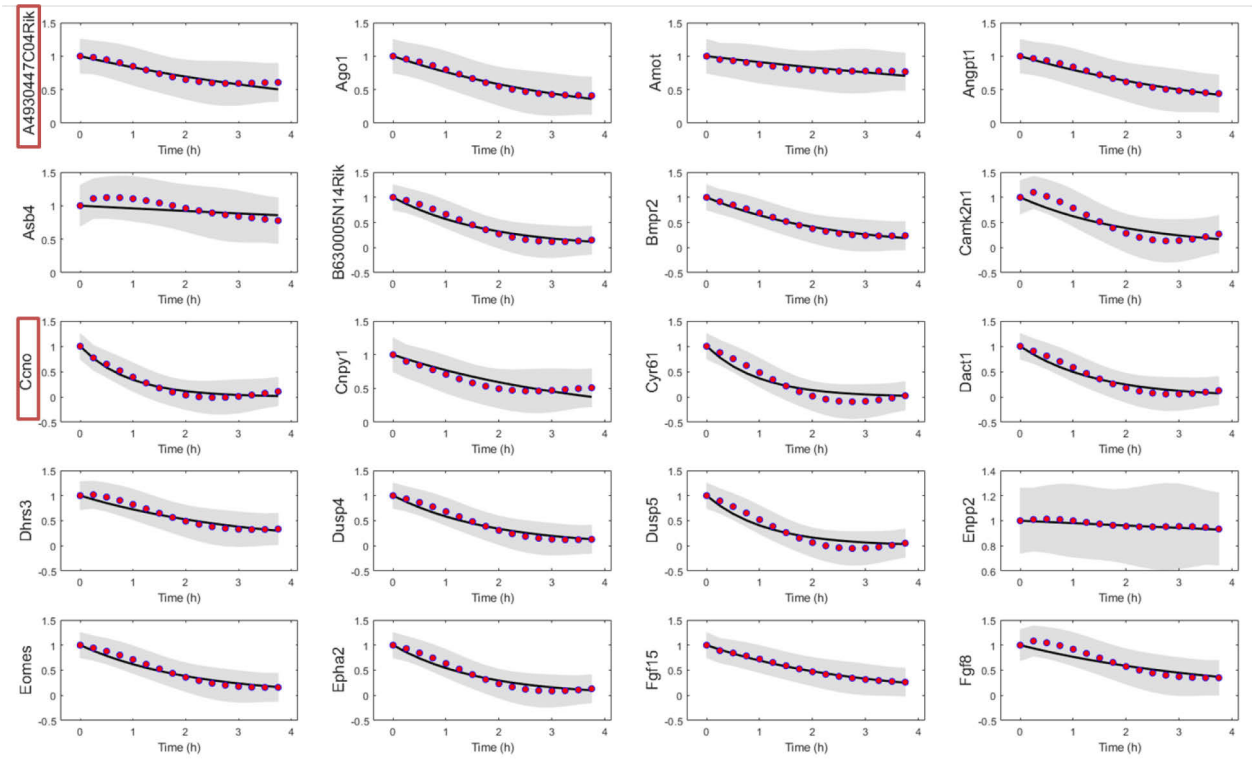


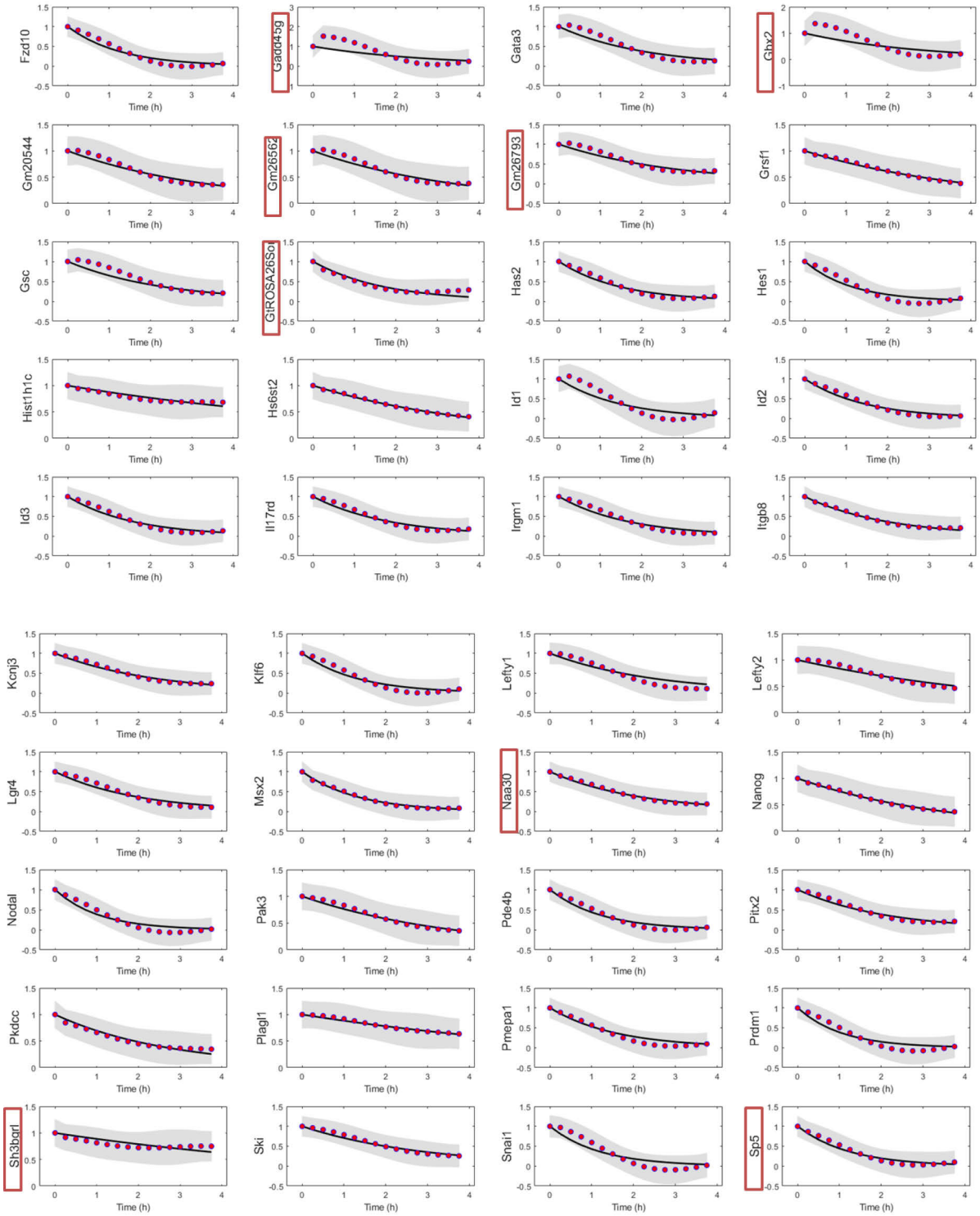

Fzd10
Gadd45g
Gata3
Gbx2
Gm20544
Gm26562
Gm26793
Grsf1
Gsc
GtROSA26Sor
Has2
Hes1
Hist1h1c
Hs6st2
Id1
Id2
Id3
Il17rd
Irgm1
Itgb8
Time (h)

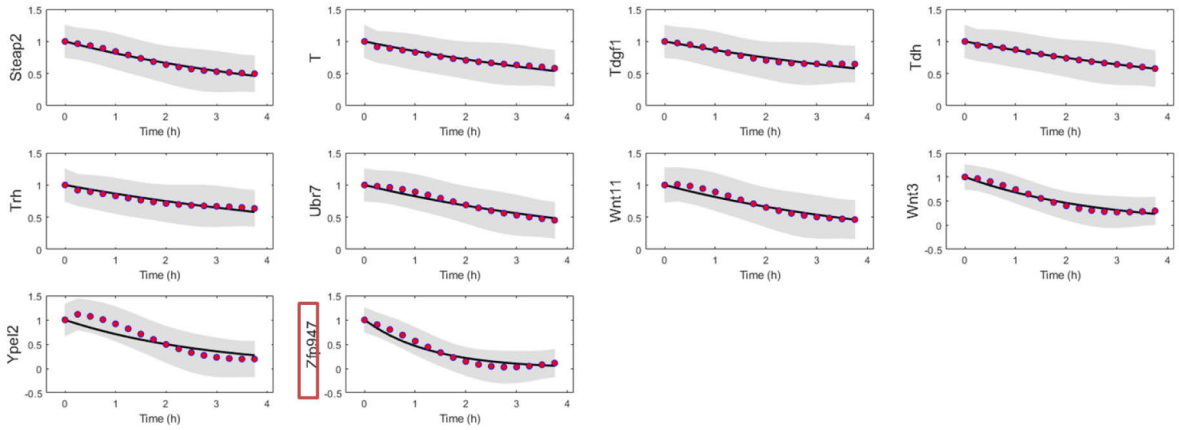

Kcnj3
Klf6
Lefty1
Lefty2
Lgr4
Msx2
Naa30
Nanog
Nodal
Pak3
Pde4b
Pitx2
Pkdcc
Plagl1
Pmepa1
Prdm1
Sh3bgrl
Ski
Snai1
Sp5
Time (h)



B: Half-life mRNA FC




Fzd10
Gadd45g
Gata3
Gbx2
Gm20544
Gm26562
Gm26793
Grsf1
Gsc
GtROSA26Sor
Has2
Hes1
Hist1h1c
Hs6st2
Id1
Id2
Id3
Il17rd
Irgm1
Itgb8
Kcnj3
Klf6
Lefty1
Lefty2
Lgr4
Msx2
Naa30
Nanog
Nodal
Pak3
Pde4b
Pitx2
Pkdcc
Ptagl1
Pmepa1
Prdm1
Sh3bgrl
Ski
Snai1
Sp5
Time (h)


Steap2
T
Tdgf1
Tdh
Trh
Ubr7
Wnt11
Wnt3
Ypel2
Zfp947
Time (h)

D. The performance of logistic regression models with different features

The features were ranked by the *F*1 score. The top three *F*1 scores or Cohen's kappa values are highlighted in aquamarine.

Feature	*F*1 score	Kappa	Combination of 2 features			*F*1 score	Kappa
Pol II FC 4h	0.875	0.750	Pol II FC 4h	and	H3K9ac FC 1h	0.882	0.764
Pol II FC 8h	0.814	0.629	Pol II FC 4h	and	Pol II FC 8h	0.878	0.757
H3K27ac FC 8h	0.750	0.516	Pol II FC 4h	and	Half-life time	0.878	0.755
H3K9ac FC 4h	0.713	0.432	Pol II FC 4h	and	SMAD2 FC 1h	0.875	0.750
H3K9ac FC 8h	0.713	0.432	Pol II FC 4h	and	H3K27ac FC 4h	0.875	0.750
SMAD2 FC 8h	0.703	0.418	Pol II FC 4h	and	H3K27ac FC 8h	0.875	0.750
H3K27ac FC 4h	0.692	0.393	Pol II FC 4h	and	H3K9ac FC 4h	0.875	0.750
H3K9ac FC 1h	0.588	0.220	Pol II FC 4h	and	H3K27ac FC 1h	0.852	0.705
H3K27ac FC 1h	0.522	0.111	H3K27ac FC 1h	and	H3K9ac FC 8h	0.845	0.691
Pol II FC 1h	0.466	0.019	Pol II FC 4h	and	H3K9ac FC 8h	0.832	0.665
SMAD2 FC 1h	0.435	0.000	Pol II FC 1h	and	Pol II FC 4h	0.822	0.644
Half-life time	0.431	-0.025	Pol II FC 8h	and	H3K9ac FC 1h	0.804	0.613
			Pol II FC 1h	and	Pol II FC 8h	0.802	0.605
			H3K27ac FC 1h	and	H3K9ac FC 4h	0.795	0.592
			Pol II FC 8h	and	Half-life time	0.795	0.591
			H3K27ac FC 4h	and	H3K9ac FC 8h	0.789	0.583
			SMAD2 FC 8h	and	H3K27ac FC 4h	0.789	0.581
			SMAD2 FC 8h	and	H3K27ac FC 8h	0.789	0.581
			Pol II FC 4h	and	SMAD2 FC 8h	0.787	0.577
			H3K27ac FC 8h	and	H3K9ac FC 4h	0.772	0.550
			H3K27ac FC 8h	and	H3K9ac FC 8h	0.772	0.550
			Pol II FC 8h	and	SMAD2 FC 1h	0.768	0.542
			Pol II FC 8h	and	H3K9ac FC 4h	0.765	0.532
			Pol II FC 8h	and	H3K27ac FC 1h	0.764	0.530
			Pol II FC 8h	and	H3K27ac FC 4h	0.764	0.530
			Pol II FC 8h	and	H3K27ac FC 8h	0.764	0.530
			H3K27ac FC 4h	and	H3K9ac FC 4h	0.755	0.513
			Pol II FC 8h	and	SMAD2 FC 8h	0.751	0.506
			SMAD2 FC 8h	and	H3K9ac FC 8h	0.739	0.485
			SMAD2 FC 8h	and	H3K9ac FC 1h	0.733	0.477
			SMAD2 FC 8h	and	H3K27ac FC 1h	0.733	0.473
			Pol II FC 8h	and	H3K9ac FC 8h	0.732	0.469
			SMAD2 FC 8h	and	H3K9ac FC 4h	0.719	0.447
			Pol II FC 1h	and	H3K27ac FC 8h	0.717	0.451
			H3K9ac FC 4h	and	H3K9ac FC 8h	0.711	0.432
			H3K27ac FC 1h	and	H3K9ac FC 1h	0.710	0.423

Feature	*F*1 score	Kappa	Combination of 2 features			*F*1 score	Kappa
			SMAD2 FC 1h	and	SMAD2 FC 8h	0.706	0.415
			SMAD2 FC 1h	and	H3K27ac FC 8h	0.704	0.425
			H3K27ac FC 1h	and	H3K27ac FC 4h	0.704	0.425
			H3K9ac FC 1h	and	H3K9ac FC 4h	0.700	0.405
			H3K27ac FC 8h	and	H3K9ac FC 1h	0.697	0.412
			Half-life time	and	SMAD2 FC 8h	0.696	0.400
			SMAD2 FC 1h	and	H3K9ac FC 8h	0.695	0.397
			Pol II FC 1h	and	H3K27ac FC 4h	0.692	0.393
			SMAD2 FC 1h	and	H3K27ac FC 4h	0.692	0.393
			H3K27ac FC 1h	and	H3K27ac FC 8h	0.692	0.393
			Half-life time	and	H3K9ac FC 4h	0.686	0.382
			Half-life time	and	H3K9ac FC 8h	0.686	0.382
			SMAD2 FC 1h	and	H3K9ac FC 4h	0.686	0.382
			Pol II FC 1h	and	SMAD2 FC 8h	0.684	0.377
			Pol II FC 1h	and	H3K9ac FC 1h	0.683	0.385
			Half-life time	and	H3K27ac FC 8h	0.674	0.366
			Pol II FC 1h	and	H3K9ac FC 8h	0.666	0.338
			H3K27ac FC 4h	and	H3K9ac FC 1h	0.647	0.313
			Half-life time	and	H3K27ac FC 4h	0.646	0.302
			Pol II FC 1h	and	H3K9ac FC 4h	0.642	0.291
			H3K9ac FC 1h	and	H3K9ac FC 8h	0.636	0.278
			SMAD2 FC 1h	and	H3K27ac FC 1h	0.608	0.239
			SMAD2 FC 1h	and	H3K9ac FC 1h	0.591	0.196
			Half-life time	and	H3K9ac FC 1h	0.563	0.172
			Half-life time	and	H3K27ac FC 1h	0.541	0.097
			Pol II FC 1h	and	H3K27ac FC 1h	0.535	0.115
			Pol II FC 1h	and	SMAD2 FC 1h	0.484	0.016
			Pol II FC 1h	and	Half-life time	0.462	-0.012

REFERENCES

Abdollah, S., M. Macias-Silva, T. Tsukazaki, H. Hayashi, L. Attisano and J. L. Wrana (1997). "TbetaRI phosphorylation of Smad2 on Ser465 and Ser467 is required for Smad2-Smad4 complex formation and signaling." J Biol Chem **272**(44): 27678-27685.

Ahlquist, P. (2002). "RNA-dependent RNA polymerases, viruses, and RNA silencing." Science **296**(5571): 1270-1273.

Ahn, S. H., M. Kim and S. Buratowski (2004). "Phosphorylation of serine 2 within the RNA polymerase IIC-terminal domain couples transcription and 3 ' end processing." Molecular Cell **13**(1): 67-76.

Alberts, B. (2017). Molecular Biology of the Cell, CRC Press.

Anders, S. and W. Huber (2010). "Differential expression analysis for sequence count data." Genome Biol **11**(10): R106.

Andres, J. L., K. Stanley, S. Cheifetz and J. Massague (1989). "Membrane-anchored and soluble forms of betaglycan, a polymorphic proteoglycan that binds transforming growth factor-beta." J Cell Biol **109**(6 Pt 1): 3137-3145.

Annes, J. P., J. S. Munger and D. B. Rifkin (2003). "Making sense of latent TGFbeta activation." J Cell Sci **116**(Pt 2): 217-224.

Ashburner, M., C. A. Ball, J. A. Blake, D. Botstein, H. Butler, J. M. Cherry, A. P. Davis, K. Dolinski, S. S. Dwight, J. T. Eppig, M. A. Harris, D. P. Hill, L. Issel-Tarver, A. Kasarskis, S. Lewis, J. C. Matese, J. E. Richardson, M. Ringwald, G. M. Rubin and G. Sherlock (2000). "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium." Nat Genet **25**(1): 25-29.

Attisano, L. and E. Labbe (2004). "TGFbeta and Wnt pathway cross-talk." Cancer Metastasis Rev **23**(1-2): 53-61.

Bakin, A. V., A. K. Tomlinson, N. A. Bhowmick, H. L. Moses and C. L. Arteaga (2000). "Phosphatidylinositol 3-kinase function is required for transforming growth factor β-mediated epithelial to mesenchymal transition and cell migration." Journal of Biological Chemistry **275**(47): 36803-36810.

Becker, P. B. and W. Horz (2002). "ATP-dependent nucleosome remodeling." Annu Rev Biochem **71**: 247-273.

Bennett, C. A. and N. L. Franklin (1954). "Statistical analysis in chemistry and the chemical industry."

Bier, E. and E. M. De Robertis (2015). "EMBRYO DEVELOPMENT. BMP gradients: A paradigm for morphogen-mediated developmental patterning." Science **348**(6242): aaa5838.

Bodenstine, T. M., G. S. Chandler, R. E. Seftor, E. A. Seftor and M. J. Hendrix (2016). "Plasticity underlies tumor progression: role of Nodal signaling." Cancer Metastasis Rev **35**(1): 21-39.

Brennan, J., D. P. Norris and E. J. Robertson (2002). "Nodal activity in the node governs left-right asymmetry." Genes Dev **16**(18): 2339-2344.

Buggenthin, F., F. Buettner, P. S. Hoppe, M. Endele, M. Kroiss, M. Strasser, M. Schwarzfischer, D. Loeffler, K. D. Kokkaliaris, O. Hilsenbeck, T. Schroeder, F. J. Theis and C. Marr (2017). "Prospective identification of hematopoietic lineage choice by deep learning." Nat Methods **14**(4): 403-406.

Burnham, K. P., D. R. Anderson and K. P. Burnham (2002). Model selection and multimodel inference : a practical information-theoretic approach. New York, Springer.

Burrows, M. and D. J. Wheeler (1994). "A block-sorting lossless data compression algorithm."

Cambray, S., C. Arber, G. Little, A. G. Dougalis, V. de Paola, M. A. Ungless, M. Li and T. A. Rodriguez (2012). "Activin induces cortical interneuron identity and differentiation in embryonic stem cell-derived telencephalic neural precursors." Nat Commun **3**: 841.

Camus, A., A. Perea-Gomez, A. Moreau and J. Collignon (2006). "Absence of Nodal signaling promotes precocious neural differentiation in the mouse embryo." Dev Biol **295**(2): 743-755.

Chen, W., S. Zhou, L. Mao, H. Zhang, D. Sun, J. Zhang, J. Li and J. H. Tang (2016). "Crosstalk between TGF-beta signaling and miRNAs in breast cancer metastasis." Tumour Biol **37**(8): 10011-10019.

Choubey, S. (2018). "Nascent RNA kinetics: Transient and steady state behavior of models of transcription." Phys Rev E **97**(2-1): 022402.

Chu, Y. J. and D. R. Corey (2012). "RNA Sequencing: Platform Selection, Experimental Design, and Data Interpretation." Nucleic Acid Therapeutics **22**(4): 271-274.

Cicchetti, D. V. and A. R. Feinstein (1990). "High agreement but low kappa: II. Resolving the paradoxes." J Clin Epidemiol **43**(6): 551-558.

Coda, D. M., T. Gaarenstroom, P. East, H. Patel, D. S. Miller, A. Lobley, N. Matthews, A. Stewart and C. S. Hill (2017). "Distinct modes of SMAD2 chromatin binding and remodeling shape the transcriptional response to NODAL/Activin signaling." Elife **6**.

Cohen, J. (1960). "A Coefficient of Agreement for Nominal Scales." Educational and Psychological Measurement **20**(1): 37-46.

Cooper, G. M., R. E. Hausman and R. E. Hausman (2000). The cell: a molecular approach, ASM press Washington, DC.

Cox, D. R. (1958). "The regression analysis of binary sequences." Journal of the Royal Statistical Society: Series B (Methodological) **20**(2): 215-232.

Crick, F. (1970). "Central dogma of molecular biology." Nature **227**(5258): 561-563.

Crick, F. H. (1958). "On protein synthesis." Symp Soc Exp Biol **12**: 138-163.

Danko, C. G., N. Hah, X. Luo, A. L. Martins, L. Core, J. T. Lis, A. Siepel and W. L. Kraus (2013). "Signaling pathways differentially affect RNA polymerase II initiation, pausing, and elongation rate in cells." Mol Cell **50**(2): 212-222.

Das, D., S. Dey, R. C. Brewster and S. Choubey (2017). "Effect of transcription factor resource sharing on gene expression noise." PLoS Comput Biol **13**(4): e1005491.

Davidson, E. H. (2006). The regulatory genome : gene regulatory networks in development and evolution. Burlington, MA ; San Diego, Academic.

De Caestecker, M. P., T. Yahata, D. Wang, W. T. Parks, S. Huang, C. S. Hill, T. Shioda, A. B. Roberts and R. J. Lechleider (2000). "The Smad4 activation domain (SAD) is a proline-rich, p300-dependent transcriptional activation domain." Journal of Biological Chemistry **275**(3): 2115-2122.

Derynck, R., J. A. Jarrett, E. Y. Chen, D. H. Eaton, J. R. Bell, R. K. Assoian, A. B. Roberts, M. B. Sporn and D. V. Goeddel (1985). "Human transforming growth factor-beta complementary DNA sequence and expression in normal and transformed cells." Nature **316**(6030): 701-705.

Derynck, R. and Y. E. Zhang (2003). "Smad-dependent and Smad-independent pathways in TGF-beta family signalling." Nature **425**(6958): 577-584.

Dillies, M. A., A. Rau, J. Aubert, C. Hennequet-Antier, M. Jeanmougin, N. Servant, C. Keime, G. Marot, D. Castel, J. Estelle, G. Guernec, B. Jagla, L. Jouneau, D. Laloe, C. Le Gall, B. Schaeffer, S. Le Crom, M. Guedj, F. Jaffrezic and F. S. Consortium (2013). "A comprehensive evaluation of normalization

methods for Illumina high-throughput RNA sequencing data analysis." Briefings in Bioinformatics **14**(6): 671-683.

Drummond, D. R., J. Armstrong and A. Colman (1985). "The effect of capping and polyadenylation on the stability, movement and translation of synthetic messenger RNAs in Xenopus oocytes." Nucleic Acids Res **13**(20): 7375-7394.

Dubois, C. M., F. Blanchette, M. H. Laprise, R. Leduc, F. Grondin and N. G. Seidah (2001). "Evidence that furin is an authentic transforming growth factor-beta1-converting enzyme." Am J Pathol **158**(1): 305-316.

Edlund, S., M. Landstrom, C.-H. Heldin and P. Aspenstrom (2002). "Transforming growth factor-β–induced mobilization of actin cytoskeleton requires signaling by small GTPases Cdc42 and RhoA." Molecular biology of the cell **13**(3): 902-914.

Eijkelenboom, A., M. Mokry, E. de Wit, L. M. Smits, P. E. Polderman, M. H. van Triest, R. van Boxtel, A. Schulze, W. de Laat, E. Cuppen and B. M. Burgering (2013). "Genome-wide analysis of FOXO3 mediated transcription regulation through RNA polymerase II profiling." Mol Syst Biol **9**: 638.

Ewing, B. and P. Green (1998). "Base-calling of automated sequencer traces using phred. II. Error probabilities." Genome Res **8**(3): 186-194.

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

Fei, T., S. Zhu, K. Xia, J. Zhang, Z. Li, J. D. Han and Y. G. Chen (2010). "Smad2 mediates Activin/Nodal signaling in mesendoderm differentiation of mouse embryonic stem cells." Cell Res **20**(12): 1306-1318.

Feldman, B., M. A. Gates, E. S. Egan, S. T. Dougan, G. Rennebeck, H. I. Sirotkin, A. F. Schier and W. S. Talbot (1998). "Zebrafish organizer development and germ-layer formation require nodal-related signals." Nature **395**(6698): 181-185.

Feng, X.-H. and R. Derynck (2005). "Specificity and versatility in TGF-β signaling through Smads." Annu. Rev. Cell Dev. Biol. **21**: 659-693.

Feng, X. H. and R. Derynck (2005). "Specificity and versatility in tgf-beta signaling through Smads." Annu Rev Cell Dev Biol **21**: 659-693.

Ferragina, P. and G. Manzini (2000). Opportunistic data structures with applications. Proceedings 41st Annual Symposium on Foundations of Computer Science, IEEE.

Forgy, E. W. (1965). "Cluster Analysis of Multivariate Data - Efficiency Vs Interpretability of Classifications." Biometrics **21**(3): 768-&.

Friess, H., Y. Yamanaka, M. Buchler, M. S. Kobrin, E. Tahara and M. Korc (1994). "Cripto, a member of the epidermal growth factor family, is over-expressed in human pancreatic cancer and chronic pancreatitis." Int J Cancer **56**(5): 668-674.

Fu, X., N. Fu, S. Guo, Z. Yan, Y. Xu, H. Hu, C. Menzel, W. Chen, Y. Li, R. Zeng and P. Khaitovich (2009). "Estimating accuracy of RNA-Seq and microarrays with proteomics." BMC Genomics **10**: 161.

Fuda, N. J., M. S. Buckley, W. Wei, L. J. Core, C. T. Waters, D. Reinberg and J. T. Lis (2012). "Fcp1 dephosphorylation of the RNA polymerase II C-terminal domain is required for efficient transcription of heat shock genes." Mol Cell Biol **32**(17): 3428-3437.

Furey, T. S. (2012). "ChIP-seq and beyond: new and improved methodologies to detect and characterize protein-DNA interactions." Nat Rev Genet **13**(12): 840-852.

Gaarenstroom, T. and C. S. Hill (2014). "TGF-beta signaling to chromatin: How Smads regulate transcription during self-renewal and differentiation." Seminars in Cell & Developmental Biology **32**: 107-118.

Gao, P., A. Honkela, M. Rattray and N. D. Lawrence (2008). "Gaussian process modelling of latent chemical species: applications to inferring transcription factor activities." Bioinformatics **24**(16): i70-75.

Garber, M., M. G. Grabherr, M. Guttman and C. Trapnell (2011). "Computational methods for transcriptome annotation and quantification using RNA-seq." Nature Methods **8**(6): 469-477.

Geisler, S. and J. Coller (2013). "RNA in unexpected places: long non-coding RNA functions in diverse cellular contexts." Nat Rev Mol Cell Biol **14**(11): 699-712.

Gentry, L. E. and B. W. Nash (1990). "The pro domain of pre-pro-transforming growth factor beta 1 when independently expressed is a functional binding protein for the mature growth factor." Biochemistry **29**(29): 6851-6857.

Goodnow Jr, R. A. (2014). A handbook for DNA-encoded chemistry: theory and applications for exploring chemical space and drug discovery, John Wiley & Sons.

Graham, J. R., M. C. Hendershott, J. Terragni and G. M. Cooper (2010). "mRNA degradation plays a significant role in the program of gene expression regulated by phosphatidylinositol 3-kinase signaling." Mol Cell Biol **30**(22): 5295-5305.

Greenwald, J., W. H. Fischer, W. W. Vale and S. Choe (1999). "Three-finger toxin fold for the extracellular ligand-binding domain of the type II activin receptor serine kinase." Nat Struct Biol **6**(1): 18-22.

Gritsman, K., W. S. Talbot and A. F. Schier (2000). "Nodal signaling patterns the organizer." Development **127**(5): 921-932.

Hao, S. and D. Baltimore (2013). "RNA splicing regulates the temporal order of TNF-induced gene expression." Proc Natl Acad Sci U S A **110**(29): 11934-11939.

Hartigan, J. A. and M. A. Wong (1979). "Algorithm AS 136: A k-means clustering algorithm." Journal of the Royal Statistical Society. Series C (Applied Statistics) **28**(1): 100-108.

Hawkins, D. M. (2004). "The problem of overfitting." Journal of chemical information and computer sciences **44**(1): 1-12.

Hofmann, T. G., N. Stollberg, M. L. Schmitz and H. Will (2003). "HIPK2 regulates transforming growth factor-beta-induced c-Jun NH(2)-terminal kinase activation and apoptosis in human hepatoma cells." Cancer Res **63**(23): 8271-8277.

Holstege, F. C., U. Fiedler and H. T. Timmers (1997). "Three transitions in the RNA polymerase II transcription complex during initiation." EMBO J **16**(24): 7468-7480.

Honkela, A., C. Girardot, E. H. Gustafson, Y. H. Liu, E. E. Furlong, N. D. Lawrence and M. Rattray (2010). "Model-based method for transcription factor target identification with limited data." Proc Natl Acad Sci U S A **107**(17): 7793-7798.

Honkela, A., J. Peltonen, H. Topa, I. Charapitsa, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence and M. Rattray (2015). "Genome-wide modeling of transcription kinetics reveals patterns of RNA production delays." Proc Natl Acad Sci U S A **112**(42): 13115-13120.

Hrdlickova, R., M. Toloue and B. Tian (2017). "RNA‐Seq methods for transcriptome analysis." Wiley Interdisciplinary Reviews: RNA **8**(1): e1364.

Huang, W., R. Loganantharaj, B. Schroeder, D. Fargo and L. Li (2013). "PAVIS: a tool for Peak Annotation and Visualization." Bioinformatics **29**(23): 3097-3099.

Hubner, G., Q. Hu, H. Smola and S. Werner (1996). "Strong induction of activin expression after injury suggests an important role of activin in wound repair." Dev Biol **173**(2): 490-498.

Huminiecki, L., L. Goldovsky, S. Freilich, A. Moustakas, C. Ouzounis and C. H. Heldin (2009). "Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom." BMC Evol Biol **9**: 28.

Inoue, M. and K. Horimoto (2017). "Relationship between regulatory pattern of gene expression level and gene function." PLoS One **12**(5): e0177430.

Itoh, S., F. Itoh, M. J. Goumans and P. ten Dijke (2000). "Signaling of transforming growth factor-beta family members through Smad proteins." European Journal of Biochemistry **267**(24): 6954-6967.

Jacobson, A. and S. W. Peltz (1996). "Interrelationships of the pathways of mRNA decay and translation in eukaryotic cells." Annu Rev Biochem **65**: 693-739.

James, D., A. J. Levine, D. Besser and A. Hemmati-Brivanlou (2005). "TGF beta/activin/nodal signaling is necessary for the maintenance of pluripotency in human embryonic stem." Development **132**(6): 1273-1282.

Jeziorska, D. M., K. W. Jordan and K. W. Vance (2009). "A systems biology approach to understanding cis-regulatory module function." Semin Cell Dev Biol **20**(7): 856-862.

Jones, C. M., M. R. Kuehn, B. L. M. Hogan, J. C. Smith and C. V. E. Wright (1995). "Nodal-Related Signals Induce Axial Mesoderm and Dorsalize Mesoderm during Gastrulation." Development **121**(11): 3651-3662.

Joseph, E. M. and D. A. Melton (1997). "Xnr4: A Xenopus nodal-related gene expressed in the Spemann organizer." Developmental Biology **184**(2): 367-372.

Juven-Gershon, T. and J. T. Kadonaga (2010). "Regulation of gene expression via the core promoter and the basal transcriptional machinery." Dev Biol **339**(2): 225-229.

Kang, H. (2013). "Appropriate design of research and statistical analyses: observational versus experimental studies." Korean J Anesthesiol **65**(2): 105-107.

Kang, J. S., T. Alliston, R. Delston and R. Derynck (2005). "Repression of Runx2 function by TGF-beta through recruitment of class II histone deacetylases by Smad3." EMBO J **24**(14): 2543-2555.

Kang, Y., C. R. Chen and J. Massague (2003). "A self-enabling TGFbeta response coupled to stress signaling: Smad engages stress response factor ATF3 for Id1 repression in epithelial cells." Mol Cell **11**(4): 915-926.

Kell, D. B. (2005). "Metabolomics, machine learning and modelling: towards an understanding of the language of cells." Biochem Soc Trans **33**(Pt 3): 520-524.

Kim, D., B. Langmead and S. L. Salzberg (2015). "HISAT: a fast spliced aligner with low memory requirements." Nat Methods **12**(4): 357-360.

Kim, D. H., L. M. Villeneuve, K. V. Morris and J. J. Rossi (2006). "Argonaute-1 directs siRNA-mediated transcriptional gene silencing in human cells." Nat Struct Mol Biol **13**(9): 793-797.

Kim, H. D. and E. K. O'Shea (2008). "A quantitative model of transcription factor-activated gene expression." Nat Struct Mol Biol **15**(11): 1192-1198.

Kim, M. R., D. W. Park, J. H. Lee, D. S. Choi, K. J. Hwang, H. S. Ryu and C. K. Min (2005). "Progesterone-dependent release of transforming growth factor-beta1 from epithelial cells enhances the endometrial decidualization by turning on the Smad signalling in stromal cells." Mol Hum Reprod **11**(11): 801-808.

Kiss, T. (2001). "Small nucleolar RNA-guided post-transcriptional modification of cellular RNAs." EMBO J **20**(14): 3617-3622.

Kleeff, J., T. Ishiwata, H. Friess, M. W. Buchler and M. Korc (1998). "Concomitant over-expression of activin/inhibin beta subunits and their receptors in human pancreatic cancer." Int J Cancer **77**(6): 860-868.

Komarnitsky, P., E. J. Cho and S. Buratowski (2000). "Different phosphorylated forms of RNA polymerase II and associated mRNA processing factors during transcription." Genes Dev **14**(19): 2452-2460.

Krebs, J. E., E. S. Goldstein and S. T. Kilpatrick (2018). Lewin's genes XII. Burlington, MA, Jones & Bartlett Learning.

Kuehner, J. N., E. L. Pearson and C. Moore (2011). "Unravelling the means to an end: RNA polymerase II transcription termination." Nat Rev Mol Cell Biol **12**(5): 283-294.

Kusakabe, M., P. L. Cheong, R. Nikfar, I. S. McLennan and K. Koishi (2008). "The structure of the TGF-beta latency associated peptide region determines the ability of the proprotein convertase furin to cleave TGF-betas." J Cell Biochem **103**(1): 311-320.

Landis, J. R. and G. G. Koch (1977). "The measurement of observer agreement for categorical data." Biometrics **33**(1): 159-174.

Landry, J. R., N. Bonadies, S. Kinston, K. Knezevic, N. K. Wilson, S. H. Oram, M. Janes, S. Piltz, M. Hammett, J. Carter, T. Hamilton, I. J. Donaldson, G. Lacaud, J. Frampton, G. Follows, V. Kouskoff and B. Gottgens (2009). "Expression of the leukemia oncogene Lmo2 is controlled by an array of tissue-specific elements dispersed over 100 kb and bound by Tal1/Lmo2, Ets, and Gata factors." Blood **113**(23): 5783-5792.

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

Lee, T. I. and R. A. Young (2000). "Transcription of eukaryotic protein-coding genes." Annu Rev Genet **34**: 77-137.

Lemon, B. and R. Tjian (2000). "Orchestrated response: a symphony of transcription factors for gene control." Genes Dev **14**(20): 2551-2569.

Liang, Y. Y., F. C. Brunicardi and X. Lin (2009). "Smad3 mediates immediate early induction of Id1 by TGF-beta." Cell Res **19**(1): 140-148.

Lifton, R., M. Goldberg, R. Karp and D. Hogness (1978). The organization of the histone genes in Drosophila melanogaster: functional and evolutionary implications. Cold Spring Harbor symposia on quantitative biology, Cold Spring Harbor Laboratory Press.

Ling, C. X. and C. Li (1998). Data mining for direct marketing: Problems and solutions. Kdd.

Liu, W. and M. Niranjan (2012). "Gaussian process modelling for bicoid mRNA regulation in spatio-temporal Bicoid profile." Bioinformatics **28**(3): 366-372.

Lloyd, S. (1982). "Least squares quantization in PCM." IEEE transactions on information theory **28**(2): 129-137.

Lo, R. S., Y. G. Chen, Y. Shi, N. P. Pavletich and J. Massagué (1998). "The L3 loop: a structural motif determining specific interactions between SMAD proteins and TGF - β receptors." The EMBO journal **17**(4): 996-1005.

Lonardo, E., P. C. Hermann, M. T. Mueller, S. Huber, A. Balic, I. Miranda-Lorenzo, S. Zagorac, S. Alcala, I. Rodriguez-Arabaolaza, J. C. Ramirez, R. Torres-Ruiz, E. Garcia, M. Hidalgo, D. A. Cebrian, R. Heuchel, M. Lohr, F. Berger, P. Bartenstein, A. Aicher and C. Heeschen (2011). "Nodal/Activin signaling drives self-renewal and tumorigenicity of pancreatic cancer stem cells and provides a target for combined drug therapy." Cell Stem Cell **9**(5): 433-446.

MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. Proceedings of the fifth Berkeley symposium on mathematical statistics and probability, Oakland, CA, USA.

Manova, K., B. V. Paynton and R. F. Bachvarova (1992). "Expression of activins and TGF beta 1 and beta 2 RNAs in early postimplantation mouse embryos and uterine decidua." Mech Dev **36**(3): 141-152.

Marquandt, D. (1980). "You should standardize the predictor variables in your regression models. Discussion of: A critique of some ridge regression methods." Journal of the American Statistical Association **75**(369): 87-91.

Massague, J. (1998). "TGF-beta signal transduction." Annu Rev Biochem **67**: 753-791.

Massague, J. (2008). "TGFbeta in Cancer." Cell **134**(2): 215-230.

Matzuk, M. M., M. J. Finegold, J. G. Su, A. J. Hsueh and A. Bradley (1992). "Alpha-inhibin is a tumour-suppressor gene with gonadal specificity in mice." Nature **360**(6402): 313-319.

Mauviel, A., K. Y. Chung, A. Agarwal, K. Tamai and J. Uitto (1996). "Cell-specific induction of distinct oncogenes of the Jun family is responsible for differential regulation of collagenase gene expression by transforming growth factor-beta in fibroblasts and keratinocytes." J Biol Chem **271**(18): 10917-10923.

Mazerbourg, S., C. Klein, J. Roh, N. Kaivo-Oja, D. G. Mottershead, O. Korchynskyi, O. Ritvos and A. J. Hsueh (2004). "Growth differentiation factor-9 signaling is mediated by the type I receptor, activin receptor-like kinase 5." Mol Endocrinol **18**(3): 653-665.

McQuarrie, A. D. R. and C.-L. Tsai (1998). Regression and time series model selection. Singapore ; River Edge, N.J., World Scientific.

Menard, S. W. (2002). Applied logistic regression analysis. Thousand Oaks, Calif., Sage Publications.

Mi, H., A. Muruganujan, D. Ebert, X. Huang and P. D. Thomas (2019). "PANTHER version 14: more genomes, a new PANTHER GO-slim and improvements in enrichment analysis tools." Nucleic Acids Res **47**(D1): D419-D426.

Miles, J. and M. Shevlin (2001). Applying regression and correlation: A guide for students and researchers, Sage.

Miyazawa, K., M. Shinozaki, T. Hara, T. Furuya and K. Miyazono (2002). "Two major Smad pathways in TGF-beta superfamily signalling." Genes Cells **7**(12): 1191-1204.

Mokry, M., P. Hatzis, J. Schuijers, N. Lansu, F. P. Ruzius, H. Clevers and E. Cuppen (2012). "Integrated genome-wide analysis of transcription factor occupancy, RNA polymerase II binding and steady-state RNA levels identify differentially regulated functional gene classes." Nucleic Acids Res **40**(1): 148-158.

Morikawa, M., R. Derynck and K. Miyazono (2016). "TGF-beta and the TGF-beta Family: Context-Dependent Roles in Cell and Tissue Physiology." Cold Spring Harb Perspect Biol **8**(5).

Morikawa, M., D. Koinuma, K. Miyazono and C. H. Heldin (2013). "Genome-wide mechanisms of Smad binding." Oncogene **32**(13): 1609-1615.

Morikawa, M., D. Koinuma, S. Tsutsumi, E. Vasilaki, Y. Kanki, C. H. Heldin, H. Aburatani and K. Miyazono (2011). "ChIP-seq reveals cell type-specific binding patterns of BMP-specific Smads and a novel binding motif." Nucleic Acids Res **39**(20): 8712-8727.

Morin, R. D., M. Bainbridge, A. Fejes, M. Hirst, M. Krzywinski, T. J. Pugh, H. McDonald, R. Varhol, S. J. M. Jones and M. A. Marra (2008). "Profiling the HeLa S3 transcriptome using randomly primed cDNA and massively parallel short-read sequencing." Biotechniques **45**(1): 81-+.

Mortazavi, A., B. A. Williams, K. McCue, L. Schaeffer and B. Wold (2008). "Mapping and quantifying mammalian transcriptomes by RNA-Seq." Nat Methods **5**(7): 621-628.

Moustakas, A., S. Souchelnytskyi and C. H. Heldin (2001). "Smad regulation in TGF-beta signal transduction." J Cell Sci **114**(Pt 24): 4359-4369.

Nakajima, T., C. Uchida, S. F. Anderson, J. D. Parvin and M. Montminy (1997). "Analysis of a cAMP-responsive activator reveals a two-component mechanism for transcriptional induction via signal-dependent factors." Genes Dev **11**(6): 738-747.

Namigai, E. K. O., N. J. Kenny and S. M. Shimeld (2014). "Right Across the Tree of Life: The Evolution of Left-Right Asymmetry in the Bilateria." Genesis **52**(6): 458-470.

Nelson, D. E., A. E. Ihekwaba, M. Elliott, J. R. Johnson, C. A. Gibney, B. E. Foreman, G. Nelson, V. See, C. A. Horton, D. G. Spiller, S. W. Edwards, H. P. McDowell, J. F. Unitt, E. Sullivan, R. Grimley, N. Benson, D. Broomhead, D. B. Kell and M. R. White (2004). "Oscillations in NF-kappaB signaling control the dynamics of gene expression." Science **306**(5696): 704-708.

Neter, J., W. Wasserman and M. H. Kutner (1989). "Applied linear regression models."

Ornitz, D. M. and N. Itoh (2001). "Fibroblast growth factors." Genome Biol **2**(3): REVIEWS3005.

Parashurama, N., Y. Nahmias, C. H. Cho, D. van Poll, A. W. Tilles, F. Berthiaume and M. L. Yarmush (2008). "Activin alters the kinetics of endoderm induction in embryonic stem cells cultured on collagen gels." Stem Cells **26**(2): 474-484.

Park, P. J. (2009). "ChIP-seq: advantages and challenges of a maturing technology." Nature Reviews Genetics **10**(10): 669-680.

Pauklin, S. and L. Vallier (2015). "Activin/Nodal signalling in stem cells." Development **142**(4): 607-619.

Perlman, R., W. P. Schiemann, M. W. Brooks, H. F. Lodish and R. A. Weinberg (2001). "TGF-beta-induced apoptosis is mediated by the adapter protein Daxx that facilitates JNK activation." Nature Cell Biology **3**(8): 708-714.

Phanish, M. K., N. A. Wahab, P. Colville-Nash, B. M. Hendry and M. E. Dockrell (2006). "The differential role of Smad2 and Smad3 in the regulation of pro-fibrotic TGFbeta1 responses in human proximal-tubule epithelial cells." Biochem J **393**(Pt 2): 601-607.

Pierce, B. A. (2012). Genetics: A conceptual approach, Macmillan.

Poniatowski, L. A., P. Wojdasiewicz, R. Gasik and D. Szukiewicz (2015). "Transforming growth factor Beta family: insight into the role of growth factors in regulation of fracture healing biology and potential clinical applications." Mediators Inflamm **2015**: 137823.

Pop, M. and S. L. Salzberg (2008). "Bioinformatics challenges of new sequencing technology." Trends Genet **24**(3): 142-149.

Press, W. H., B. P. Flannery, S. A. Teukolsky and W. T. Vetterling (1992). Numeric recipes in C: the art of scientific computing, Cambridge: Cambridge University Press.

Price, D. H. (2000). "P-TEFb, a cyclin-dependent kinase controlling elongation by RNA polymerase II." Mol Cell Biol **20**(8): 2629-2634.

Quinlan, A. R. and I. M. Hall (2010). "BEDTools: a flexible suite of utilities for comparing genomic features." Bioinformatics **26**(6): 841-842.

Rabani, M., J. Z. Levin, L. Fan, X. Adiconis, R. Raychowdhury, M. Garber, A. Gnirke, C. Nusbaum, N. Hacohen, N. Friedman, I. Amit and A. Regev (2011). "Metabolic labeling of RNA uncovers principles of RNA production and degradation dynamics in mammalian cells." Nat Biotechnol **29**(5): 436-442.

Rasmussen, C. E. and H. Nickisch (2010). "Gaussian Processes for Machine Learning (GPML) Toolbox." Journal of Machine Learning Research **11**: 3011-3015.

Rasmussen, C. E. and C. K. I. Williams (2006). Gaussian processes for machine learning. Cambridge, Mass., MIT Press.

Raue, A., M. Schilling, J. Bachmann, A. Matteson, M. Schelker, D. Kaschek, S. Hug, C. Kreutz, B. D. Harms, F. J. Theis, U. Klingmuller and J. Timmer (2013). "Lessons learned from quantitative dynamical modeling in systems biology." PLoS One **8**(9): e74335.

Raue, A., B. Steiert, M. Schelker, C. Kreutz, T. Maiwald, H. Hass, J. Vanlier, C. Tonsing, L. Adlung, R. Engesser, W. Mader, T. Heinemann, J. Hasenauer, M. Schilling, T. Hofer, E. Klipp, F. Theis, U. Klingmuller, B. Schoberl and J. Timmer (2015). "Data2Dynamics: a modeling environment tailored to parameter estimation in dynamical systems." Bioinformatics **31**(21): 3558-3560.

Rebagliati, M. R., R. Toyama, P. Haffter and I. B. Dawid (1998). "cyclops encodes a nodal-related factor involved in midline signaling." Proceedings of the National Academy of Sciences of the United States of America **95**(17): 9932-9937.

Reinberg, D., G. Orphanides, R. Ebright, S. Akoulitchev, J. Carcamo, H. Cho, P. Cortes, R. Drapkin, O. Flores and I. Ha (1998). The RNA polymerase II general transcription factors: past, present, and future. Cold Spring Harbor Symposia on Quantitative Biology, Cold Spring Harbor Laboratory Press.

Roberts, A. B., M. A. Anzano, L. C. Lamb, J. M. Smith and M. B. Sporn (1981). "New class of transforming growth factors potentiated by epidermal growth factor: isolation from non-neoplastic tissues." Proc Natl Acad Sci U S A **78**(9): 5339-5343.

Robertson, E., A. Bradley, M. Kuehn and M. Evans (1986). "Germ-Line Transmission of Genes Introduced into Cultured Pluripotential Cells by Retroviral Vector." Nature **323**(6087): 445-448.

Robertson, G., M. Hirst, M. Bainbridge, M. Bilenky, Y. Zhao, T. Zeng, G. Euskirchen, B. Bernier, R. Varhol, A. Delaney, N. Thiessen, O. L. Griffith, A. He, M. Marra, M. Snyder and S. Jones (2007). "Genome-wide profiles of STAT1 DNA association using chromatin immunoprecipitation and massively parallel sequencing." Nat Methods **4**(8): 651-657.

Robinson, J. T., H. Thorvaldsdottir, W. Winckler, M. Guttman, E. S. Lander, G. Getz and J. P. Mesirov (2011). "Integrative genomics viewer." Nat Biotechnol **29**(1): 24-26.

Ropers, D., H. de Jong, M. Page, D. Schneider and J. Geiselmann (2006). "Qualitative simulation of the carbon starvation response in Escherichia coli." Biosystems **84**(2): 124-152.

Ross, M. G., C. Russ, M. Costello, A. Hollinger, N. J. Lennon, R. Hegarty, C. Nusbaum and D. B. Jaffe (2013). "Characterizing and measuring bias in sequence data." Genome Biol **14**(5): R51.

Ross, S., E. Cheung, T. G. Petrakis, M. Howell, W. L. Kraus and C. S. Hill (2006). "Smads orchestrate specific histone modifications and chromatin remodeling to activate transcription." EMBO J **25**(19): 4490-4502.

Ross, S. and C. S. Hill (2008). "How the Smads regulate transcription." International Journal of Biochemistry & Cell Biology **40**(3): 383-408.

Rothberg, J. M. and J. H. Leamon (2008). "The development and impact of 454 sequencing." Nat Biotechnol **26**(10): 1117-1124.

Rotzer, D., M. Krampert, S. Sulyok, S. Braun, H. J. Stark, P. Boukamp and S. Werner (2006). "Id proteins: novel targets of activin action, which regulate epidermal homeostasis." Oncogene **25**(14): 2070-2081.

Rozowsky, J., G. Euskirchen, R. K. Auerbach, Z. D. Zhang, T. Gibson, R. Bjornson, N. Carriero, M. Snyder and M. B. Gerstein (2009). "PeakSeq enables systematic scoring of ChIP-seq experiments relative to controls." Nat Biotechnol **27**(1): 66-75.

Rydenfelt, M., R. S. Cox, 3rd, H. Garcia and R. Phillips (2014). "Statistical mechanical model of coupled transcription from multiple promoters due to transcription factor titration." Phys Rev E Stat Nonlin Soft Matter Phys **89**(1): 012702.

Saharinen, J., J. Taipale and J. Keski-Oja (1996). "Association of the small latent transforming growth factor-beta with an eight cysteine repeat of its binding protein LTBP-1." EMBO J **15**(2): 245-253.

Saijoh, Y., S. Oki, S. Ohishi and H. Hamada (2003). "Left-right patterning of the mouse lateral plate requires Nodal produced in the node." Developmental Biology **256**(1): 160-172.

Sampath, K., A. L. Rubinstein, A. H. S. Cheng, J. O. Liang, K. Fekany, L. Solnica-Krezel, V. Korzh, M. E. Halpern and C. V. E. Wright (1998). "Induction of the zebrafish ventral brain and floorplate requires cyclops/nodal signalling." Nature **395**(6698): 185-189.

Sanger, F., S. Nicklen and A. R. Coulson (1977). "DNA sequencing with chain-terminating inhibitors." Proc Natl Acad Sci U S A **74**(12): 5463-5467.

Sasaki, Y. (2007). "The truth of the F-measure." Teach Tutor mater **1**(5): 1-5.

Schier, A. F. and W. S. Talbot (2001). "Nodal signaling and the zebrafish organizer." Int J Dev Biol **45**(1): 289-297.

Schiffer, M., G. von Gersdorff, M. Bitzer, K. Susztak and E. P. Bottinger (2000). "Smad proteins and transforming growth factor-beta signaling." Kidney Int Suppl **77**: S45-52.

Schmidt, D., M. D. Wilson, C. Spyrou, G. D. Brown, J. Hadfield and D. T. Odom (2009). "ChIP-seq: Using high-throughput sequencing to discover protein-DNA interactions." Methods **48**(3): 240-248.

Schwanhausser, B., D. Busse, N. Li, G. Dittmar, J. Schuchhardt, J. Wolf, W. Chen and M. Selbach (2011). "Global quantification of mammalian gene expression control." Nature **473**(7347): 337-342.

Shi, Y. and J. Massague (2003). "Mechanisms of TGF-beta signaling from cell membrane to the nucleus." Cell **113**(6): 685-700.

Shi, Y., Y.-F. Wang, L. Jayaraman, H. Yang, J. Massagué and N. P. Pavletich (1998). "Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-β signaling." Cell **94**(5): 585-594.

Shum, E. Y., S. H. Jones, A. Shao, J. Dumdie, M. D. Krause, W. K. Chan, C. H. Lou, J. L. Espinoza, H. W. Song, M. H. Phan, M. Ramaiah, L. Huang, J. R. McCarrey, K. J. Peterson, D. G. De Rooij, H. Cook-Andersen and M. F. Wilkinson (2016). "The Antagonistic Gene Paralogs Upf3a and Upf3b Govern Nonsense-Mediated RNA Decay." Cell **165**(2): 382-395.

Tam, P. P. L. and D. A. F. Loebel (2007). "Gene function in mouse embryogenesis: get set for gastrulation." Nature Reviews Genetics **8**(5): 368-381.

Teif, V. B. and K. Rippe (2009). "Predicting nucleosome positions on the DNA: combining intrinsic sequence preferences and remodeler activities." Nucleic Acids Res **37**(17): 5641-5655.

Teng, M. and R. A. Irizarry (2017). "Accounting for GC-content bias reduces systematic errors and batch effects in ChIP-seq data." Genome Res **27**(11): 1930-1938.

The Gene Ontology, C. (2019). "The Gene Ontology Resource: 20 years and still GOing strong." Nucleic Acids Res **47**(D1): D330-D338.

Thisse, C. and B. Thisse (1999). "Antivin, a novel and divergent member of the TGFbeta superfamily, negatively regulates mesoderm induction." Development **126**(2): 229-240.

Tripathi, V., K. M. Sixt, S. Gao, X. Xu, J. Huang, R. Weigert, M. Zhou and Y. E. Zhang (2016). "Direct Regulation of Alternative Splicing by SMAD3 through PCBP1 Is Essential to the Tumor-Promoting Role of TGF-beta." Mol Cell **64**(3): 549-564.

Trompouki, E., T. V. Bowman, L. N. Lawton, Z. P. Fan, D. C. Wu, A. DiBiase, C. S. Martin, J. N. Cech, A. K. Sessa, J. L. Leblanc, P. Li, E. M. Durand, C. Mosimann, G. C. Heffner, G. Q. Daley, R. F. Paulson, R. A. Young and L. I. Zon (2011). "Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration." Cell **147**(3): 577-589.

Vale, W., J. Rivier, J. Vaughan, R. Mcclintock, A. Corrigan, W. Woo, D. Karr and J. Spiess (1986). "Purification and Characterization of an Fsh Releasing Protein from Porcine Ovarian Follicular-Fluid." Nature **321**(6072): 776-779.

Vallier, L., M. Alexander and R. A. Pedersen (2005). "Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells." J Cell Sci **118**(Pt 19): 4495-4509.

Valouev, A., J. Ichikawa, T. Tonthat, J. Stuart, S. Ranade, H. Peckham, K. Zeng, J. A. Malek, G. Costa, K. McKernan, A. Sidow, A. Fire and S. M. Johnson (2008). "A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning." Genome Res **18**(7): 1051-1063.

Vaquerizas, J. M., S. K. Kummerfeld, S. A. Teichmann and N. M. Luscombe (2009). "A census of human transcription factors: function, expression and evolution." Nat Rev Genet **10**(4): 252-263.

Visel, A., M. J. Blow, Z. Li, T. Zhang, J. A. Akiyama, A. Holt, I. Plajzer-Frick, M. Shoukry, C. Wright, F. Chen, V. Afzal, B. Ren, E. M. Rubin and L. A. Pennacchio (2009). "ChIP-seq accurately predicts tissue-specific activity of enhancers." Nature **457**(7231): 854-858.

wa Maina, C., A. Honkela, F. Matarese, K. Grote, H. G. Stunnenberg, G. Reid, N. D. Lawrence and M. Rattray (2014). "Inference of RNA polymerase II transcription dynamics from chromatin immunoprecipitation time course data." PLoS Comput Biol **10**(5): e1003598.

Walker, S. H. and D. B. Duncan (1967). "Estimation of the probability of an event as a function of several independent variables." Biometrika **54**(1-2): 167-179.

Watson, J. D. and F. H. Crick (1953). "Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid." Nature **171**(4356): 737-738.

Werner-Allen, J. W., C. J. Lee, P. Liu, N. I. Nicely, S. Wang, A. L. Greenleaf and P. Zhou (2011). "cis-Proline-mediated Ser(P)5 dephosphorylation by the RNA polymerase II C-terminal domain phosphatase Ssu72." J Biol Chem **286**(7): 5717-5726.

White, R. J. (2009). Gene transcription: mechanisms and control, John Wiley & Sons.

Wilbanks, E. G. and M. T. Facciotti (2010). "Evaluation of algorithm performance in ChIP-seq peak detection." PLoS One **5**(7): e11471.

Wilkes, M. C., H. Mitchell, S. G. Penheiter, J. J. Doré, K. Suzuki, M. Edens, D. K. Sharma, R. E. Pagano and E. B. Leof (2005). "Transforming growth factor-β activation of phosphatidylinositol 3-kinase is independent of Smad2 and Smad3 and regulates fibroblast responses via p21-activated kinase-2." Cancer research **65**(22): 10431-10440.

Willert, J., M. Epping, J. R. Pollack, P. O. Brown and R. Nusse (2002). "A transcriptional response to Wnt protein in human embryonic carcinoma cells." BMC Dev Biol **2**: 8.

Wrana, J. L., L. Attisano, J. Carcamo, A. Zentella, J. Doody, M. Laiho, X. F. Wang and J. Massague (1992). "Tgf-Beta Signals through a Heteromeric Protein-Kinase Receptor Complex." Cell **71**(6): 1003-1014.

Wrana, J. L., L. Attisano, R. Wieser, F. Ventura and J. Massague (1994). "Mechanism of activation of the TGF-beta receptor." Nature **370**(6488): 341-347.

Wray, G. A., M. W. Hahn, E. Abouheif, J. P. Balhoff, M. Pizer, M. V. Rockman and L. A. Romano (2003). "The evolution of transcriptional regulation in eukaryotes." Mol Biol Evol **20**(9): 1377-1419.

Wu, M. Y. and C. S. Hill (2009). "Tgf-beta superfamily signaling in embryonic development and homeostasis." Dev Cell **16**(3): 329-343.

Yagi, K., D. Goto, T. Hamamoto, S. Takenoshita, M. Kato and K. Miyazono (1999). "Alternatively spliced variant of Smad2 lacking exon 3. Comparison with wild-type Smad2 and Smad3." J Biol Chem **274**(2): 703-709.

Yan, X., X. Xiong and Y.-G. Chen (2017). "Feedback regulation of TGF-β signaling." Acta biochimica et biophysica Sinica **50**(1): 37-50.

Yeo, C. Y. and M. Whitman (2001). "Nodal signals to Smads through Cripto-dependent and Cripto-independent mechanisms." Molecular Cell **7**(5): 949-957.

Yu, L., M. C. Hebert and Y. E. Zhang (2002). "TGF-beta receptor-activated p38 MAP kinase mediates Smad-independent TGF-beta responses." EMBO J **21**(14): 3749-3759.

Zavadil, J., M. Bitzer, D. Liang, Y. C. Yang, A. Massimi, S. Kneitz, E. Piek and E. P. Bottinger (2001). "Genetic programs of epithelial cell plasticity directed by transforming growth factor-beta." Proc Natl Acad Sci U S A **98**(12): 6686-6691.

Zawel, L., J. L. Dai, P. Buckhaults, S. Zhou, K. W. Kinzler, B. Vogelstein and S. E. Kern (1998). "Human Smad3 and Smad4 are sequence-specific transcription activators." Mol Cell **1**(4): 611-617.

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

Zhang, Y. E. (2017). "Non-Smad Signaling Pathways of the TGF-beta Family." Cold Spring Harb Perspect Biol **9**(2).

Zhang, Z. D., J. Rozowsky, M. Snyder, J. Chang and M. Gerstein (2008). "Modeling ChIP sequencing in silico with applications." PLoS Comput Biol **4**(8): e1000158.

Zhou, X., H. Sasaki, L. Lowe, B. L. Hogan and M. R. Kuehn (1993). "Nodal is a novel TGF-beta-like gene expressed in the mouse node during gastrulation." Nature **361**(6412): 543-547.

Statement

Ich erkläre, dass ich die vorliegende Arbeit selbständig und nur unter Verwendung der angegebenen Hilfsmittel angefertigt habe.

Dan Shi

Berlin, 07.01.2020