NODAL/Activin signalling to chromatin: mechanisms of SMAD2-regulated transcription

Davide Martino Coda

University College London

and

The Francis Crick Institute

PhD Supervisor: Dr. Caroline Hill

A thesis submitted for the degree of

Doctor of Philosophy

University College London

January 2018

Declaration

I Davide Martino Coda confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Abstract

NODAL/Activin signalling regulates key processes during embryonic development via SMAD2. How SMAD2 activates programmes of gene expression that are modulated over time, however, is not known. In this thesis, using the P19 embryonic teratoma cell line as a model system, I delineate the sequence of events that occur from SMAD2 binding to transcriptional activation, and the underlying mechanisms. I show that NODAL/Activin signalling induces dramatic changes in the chromatin landscape, and orchestrates a dynamic transcriptional network regulated by SMAD2, which acts via multiple mechanisms. By combining different genome-wide approaches, I have discovered two modes of SMAD2 binding. SMAD2 can bind pre-acetylated nucleosome-depleted sites, where it promotes a further increase in H3K9ac/H3K27ac. However, SMAD2 also binds to unacetylated, closed chromatin, independently of pioneer factors, where it induces nucleosome displacement and H3 acetylation. For a subset of genes, this requires cooperation with the remodeller SMARCA4 and the transcription factor FOXH1. I demonstrate that SMAD2 regulates RNA Polymerase II via *de novo* recruitment to target promoters, and that long term modulation of the transcriptional responses requires continued NODAL/Activin signalling. Moreover, SMAD2 binding does not necessarily equate with transcriptional kinetics, and my data suggest that SMAD2 recruits multiple co-factors during sustained signaling to shape the downstream transcriptional programme. I have used ATAC-seq to identify specific transcription factor footprints at SMAD2 binding sites, and future work will aim to unveil and characterise the network of transcription factors that collaborate with SMAD2 and enable cells to correctly interpret NODAL/Activin signaling over time.

Impact Statement

The signalling molecules NODAL and Activin control a broad spectrum of biological processes in both normal and pathological contexts through activation of their intracellular effector SMAD2. In my research, I identified universal mechanisms through which SMAD2 transmits the NODAL/Activin signal in cells. During my PhD, I had the opportunity to present my results to the scientific community working in the transcription and epigenetic fields at several international meetings, most notably the 2015 EMBL Transcription meeting in Heidelberg and the 2016 EpiGenesys meeting in Paris. Since my work has provided concepts and methodologies which can be used to understand how cells interpret extracellular signals in general, I anticipate that my findings will have a great impact in setting future directions of research in these fundamental areas of biology. My work has also been published in January 2017 in eLife, which is a well-regarded open access journal. Hence, the results and the data I have generated are available online to a broad audience.

In the long term, I am confident that my findings have the potential to be used directly for medical application. During development and *in vitro* differentiation of stem cells, NODAL/Activin signalling drives the formation of endoderm, the embryonic tissue which ultimately gives rise to the digestive and respiratory tubes. Since my research contributes to the elucidation of how this process is implemented, I believe that it could be of great benefit in the context of regenerative medicine. This innovative approach aims to revolutionise patient care in the 21st century by replacing damaged human cells in the body with their healthy counterparts obtained *in vitro*. Being able to generate specific tissues starting from undifferentiated cells is the key for this technique to succeed.

NODAL/Activin signalling is also crucial in cancer, where its misregulation is now known to have prominent roles in both tumour development and metastasis. In particular, Activin has recently been shown to promote tumour progression in different types of cancers, amongst which is the highly aggressive, poorly treatable pancreatic cancer. Therefore, Activin represents a very attractive target for anticancer therapies. My finding that, in a cancer cell line, Activin signalling controls cell behaviour over long period of time has already contributed the lab embarking in a ‘spin-out’ project in collaboration with a large pharmaceutical company. Finally, I plan to identify the enzymes and the other factors which help the SMADs to execute

Activin signalling in cancer cells, and this could ultimately provide a list of candidates to use for drug discovery approaches.

Acknowledgements

This thesis would not have been possible without the support of many people, and I would like to give them credit here.

I am immensely grateful to my supervisor Dr. Caroline Hill, for giving me the opportunity to work in her lab, for the exceptional guidance, and precious advice which enabled me to hugely develop during my PhD, both as a person and as a scientist.

I would like to express my sincere gratitude to my colleagues Ila, Claire, Anna, Izzie, Tessa, Danny, Andrew, Thijs, John and Rob for always being helpful, for the scientific discussions, but also for all those fun moments shared together at work and on many other occasions. I would also like to acknowledge Philip East and Harshil Patel from the Bioinformatics and Biostatistics Service of the Francis Crick Institute for their essential contributions to the project.

Thank you to my family, my aunts, my uncle and my cousins. My mum and dad, Claudia and Alberto, and my brothers, Emanuele and Carlo, for always believing in me, supporting my choices and being there when it matters the most. This thesis is also in memory of Ester, Riccardo, Ancella and Maurizio, who could not see me finish the PhD, but have been precious sources of inspiration.

Having moved to London for the PhD, I was blessed to find a family away from home in the friends of ‘Parchetto’ and in the frequent visits from Paola, Bolzo and Michi. Thank you.

For being part of my life every day, thanks to Carlotta. Without you I simply would have never made it through the PhD. This thesis is dedicated to you.

Table of Contents

Table of figures

Abbreviations

Ac	Acetyl
AMH	Anti-Müllerian hormone
APS	Ammonium persulphate
ATAC	Assay for transposase accessible chromatin
ATP	Adenosine triphosphate
BMP	Bone morphogenetic protein
bp	Basepair
BSA	Bovine serum albumin
CDK	Cyclin-dependent kinase
cDNA	Complementary DNA
ChIP	Chromatin immunoprecipitation
cm	Centimetre
COFs	Transcriptional cofactors
CRISPR	Clustered regularly interspaced short palindromic repeats
Ct	Cycle threshold
DMEM	Dulbecco's Modified Eagle Medium
DMSO	Dimethyl sulfoxide
DNA	Deoxyribonucleic acid
dNTP	Deoxy-nucleotide triphosphate
DTT	Dithiothreitol
EpiSC	Epiblast-derived stem cells
ERK	Extracellular signal-regulated kinase
ESC	Embryonic stem cell
EDTA	Ethylenediaminetetraacetic acid
EGTA	Ethylene glycol-bis(β-aminoethyl ether)-N,N,N',N'-tetraacetic acid
FACS	Fluorescent activated cell sorting
FAIRE	Formaldehyde-assisted isolation of regulatory elements
FBS	Foetal bovine serum
FGF	Fibroblast growth factor
Fwd	Forward
g	Gravity
GDF	Growth and differentiation factor
GFP	Green fluorescent protein
GS	Glycine- and serine-rich
GTFs	General transcription factors
GTP	Guanosine triphosphate

hESC	Human embryonic stem cell
H3	Histone H3
H3K…	Histone H3 Lysine…
hi-C	High-throughput chromosome conformation capture
hr	Hours
HAT	Histone acetyl transferase
HP1	Heterochromatin protein 1
HRP	Horseradish peroxidase
IGV	Integrative genome viewer
I-SMAD	Inhibitory SMAD
LB	Lysogeny broth
log_2FC	log_2 Fold Change
MACS	Model-based analysis for ChIP-seq
mAmp	Milli-ampere
MAPK	Mitogen-activated protein kinase
Me	Methyl
MEME	Multiple Em for Motif Elicitation
mESC	Mouse embryonic stem cell
MH1	Mad homology domain 1
MH2	Mad homology domain 2
mins	Minutes
ml	Millilitre
mM	Millimolar
mg	Milligram
mRNA	Messenger RNA
NGS	Next generation sequencing
NEB	New England Biosciences
nl	Nanolitre
nm	Nanomolar
ng	Nanogram
NT	Non-targeting
OD	Optical density
OSN	OCT4, SOX2 and NANOG
P	Phosphorylated
PBS	Phosphate buffered saline
PcG	Polycomb group proteins
PCR	Polymerase chain reaction
PenStrep	Penicillin/Streptomycin
PI3K	Phosphatidylinositol-4,5-bisphosphate 3-kinase
Pol II	RNA polymerase II
PVDF	Polyvinylidene fluoride
qPCR	Quantitative real-time PCR

R-SMAD	Receptor-regulated SMAD
Rev	Reverse
RNA	Ribonucleic acid
rpm	Revolutions per minute
SBS	SMAD binding site
SD	Standard deviation
SDS	Sodium dodecyl sulphate
seq	Sequencing
siRNA	Small interfering RNA
snRNA	Small nuclear RNA
SWI/SNF	SWItch/Sucrose Non-Fermentable
TAD	Topologically Associated Domain
TAE	Tris-acetate EDTA
TEMED	Tetramethylethylenediamine
TF	Transcription factor
TGF-β	Transforming growth factor β
Tris	Tris(hydroxymethyl)aminomethane
TSS	Transcription start site
TTS	Transcription termination site
UCSC	University of Santa Cruz, California
UV	Ultraviolet
WB	Western blot
Wnt	Wingless-related integration site
v/v	Volume per volume
w/v	Weight per volume
µl	Microlitre
µM	Micromolar
µg	Microgram
4-C	Circularized chromosome conformation capture
5-C	Carbon copy chromosome conformation capture

Chapter 1. Introduction

The set of genes that are transcribed in a cell defines and maintains its specific identity (Lee and Young, 2013). During embryonic development, in adult organisms and in several diseases, such as cancer and congenital disorders, programmes of gene expression are established by extracellular signals via activation of transcriptional effectors functioning as signal-responsive transcription factors (Perrimon et al., 2012). In recent years, the explosion of experimental and computational techniques has led to critical insights into how transcription is regulated (van Dijk et al., 2014). However, the sequence of events that occur from transcriptional effectors activation to control of gene expression is still largely unknown. The molecular mechanisms underlying this process are also poorly understood. The Transforming Growth Factor β (TGF-β) superfamily produces signals crucial for many physiological and pathological processes, from embryogenesis and tissue homeostasis, to cancer development and metastasis (Caja et al., 2012, Massague, 2008, Massague, 2012, Wakefield and Hill, 2013, Wu and Hill, 2009). Thus, it provides an excellent model system to study how cells interpret extracellular signals with respect to complex programmes of gene expression. Here, I will begin with a brief overview of the general modes of transcription regulation, with a focus on the role of chromatin and histone modifications. Then, after introducing the TGF-β superfamily signalling pathway I will review the biological role of the two TGF-β superfamily members studied in this thesis, Activin and NODAL. Finally, the state of the art of how the transcriptional effectors of the pathway, the SMADs, regulate gene expression will be discussed.

1.1 Modes of transcription regulation

1.1.1 General principles of transcription

Transcription is the result of complex interplay between DNA-binding proteins called transcription factors (TFs), the chromatin architecture and the general transcription apparatus (Figure 1.1). According to the classical model, TFs recognise *cis*-regulatory elements in the genome, such as enhancers and promoters, to control the

activity of target genes, in both a positive and negative manner (Voss and Hager, 2014).

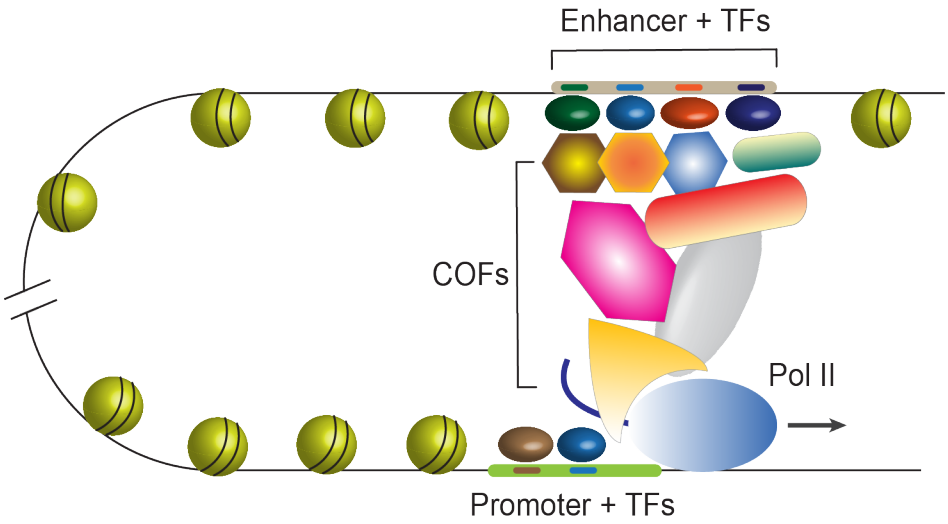


Figure 1.1. The main players of transcriptional regulation.
Transcription factors (TFs) recognise short sequence motifs present at enhancers/promoters and recruit transcriptional cofactors (COFs), which in turn mediate the recruitment of RNA Polymerase II (Pol II) at core-promoters to initiate transcription. Adapted from (Reiter et al., 2017).

In eukaryotes, especially in mammals, site-specific recognition of DNA sequences is not sufficient for TFs to drive the distinct transcriptional programmes that define individual cell types. For multicellular organisms, the organisation of the genome into complex nucleoproteins structure is also crucial, since it enables regulated access of TFs to regulatory sites (Spitz and Furlong, 2012, Voss and Hager, 2014). In general, chromatin transitions affecting gene expression can occur at three levels: histone post-translational modifications (PTMs); nucleosome mobilisation; changes to higher order structures (Li et al., 2007). These processes are regulated by enzymes and structural proteins, and how they interact with TFs is a central question in biology. Once bound to enhancers, the ability of TFs to activate transcription depends on the recruitment of coactivator proteins (COFs). COFs do not have DNA-binding properties on their own, and include chromatin remodellers, histone modifiers and factors such as the Mediator complex which bring enhancers in close proximity to promoters by forming DNA loops (Lee and Young, 2013). Together, these regulatory proteins coordinate the recruitment of RNA polymerase II (Pol II) and general transcription factors (GTFs) to transcription start sites (Figure

1.1). In some cases, enhancers are also bound by GTFs and Pol II, and this results in the production of enhancer originating RNAs known as eRNAs (Goldstein and Hager, 2017, Reiter et al., 2017). eRNAs function through poorly understood mechanisms to regulate gene expression either in *trans* at other genomic loci or in *cis* at the enhancer site itself (Mousavi et al., 2013, Orom and Shiekhattar, 2011).

Since transcription control is the result of a cross-talk between chromatin, TFs and Pol II, modes of function and regulation of these three key players will now be discussed.

1.1.2 Chromatin features and its functions

In the last decade, the advent of high-throughput sequencing technologies made it possible to profile all aspects of chromatin genome-wide in parallel with global analysis of binding of TFs and other proteins (van Dijk et al., 2014). Maps of histone modifications/variants, DNA accessibility, and long-range chromosome interactions have been obtained in different cell types and organisms, using approaches such chromatin immunoprecipitation sequencing (ChIP-seq), DNase I hypersensitive sites sequencing (DNase-seq) and chromatin conformation capture coupled to sequencing (4-C, 5-C, Hi-C) (Dekker et al., 2013). From the integration of these data, the genome has been partitioned into distinct functional elements on the basis of unique chromatin features (Figure 1.2). Transcriptionally inactive, repressed chromatin (Figure 1.2A) is characterised by large domains of H3K27me3, H3K9me2 and/or H3K9me3, and it is bound by Polycomb Group proteins (PcG) and heterochromatin proteins (HP1) (Fisher and Fisher, 2011, Zhou et al., 2011). In contrast, transcribed genes tend to be enriched for H3K36me3, H3K79me2 and have high overall acetylation (Figure 1.2B). The DNA is also readily accessible to Pol II and other factors, since acetylation disrupts the contacts between nucleosomes by reducing the positive charge of the histones (Zentner and Henikoff, 2013, Henikoff and Shilatifard, 2011). Active promoters are also nucleosome depleted, are characterised by the presence of less stable histone variants such as H2A.Z/H3.3 and are commonly marked by acetylation, H3K4me2 and H3K4me3 (Zhou et al., 2011). Moreover, only in embryonic stem cells (ESCs), promoters of key developmental genes (Figure 1.2E) simultaneously display histone modifications that are characteristic of gene repression (H3K27me3) and activation (H3K4me3). These

so-called bivalent domains (see further below) have been proposed to keep the genes in an 'poised' state, allowing timely activation while maintaining repression in the absence of differentiation signals (Bernstein et al., 2006, Voigt et al., 2013). Interestingly, despite evidences exist for asymmetrically modified nucleosomes carrying H3K4me3 and H3K27me3 on opposite H3 tails, whether the two bivalency marks co-occur on the same nucleosome still needs to be clearly demonstrated (Harikumar and Meshorer, 2015, Voigt et al., 2012).

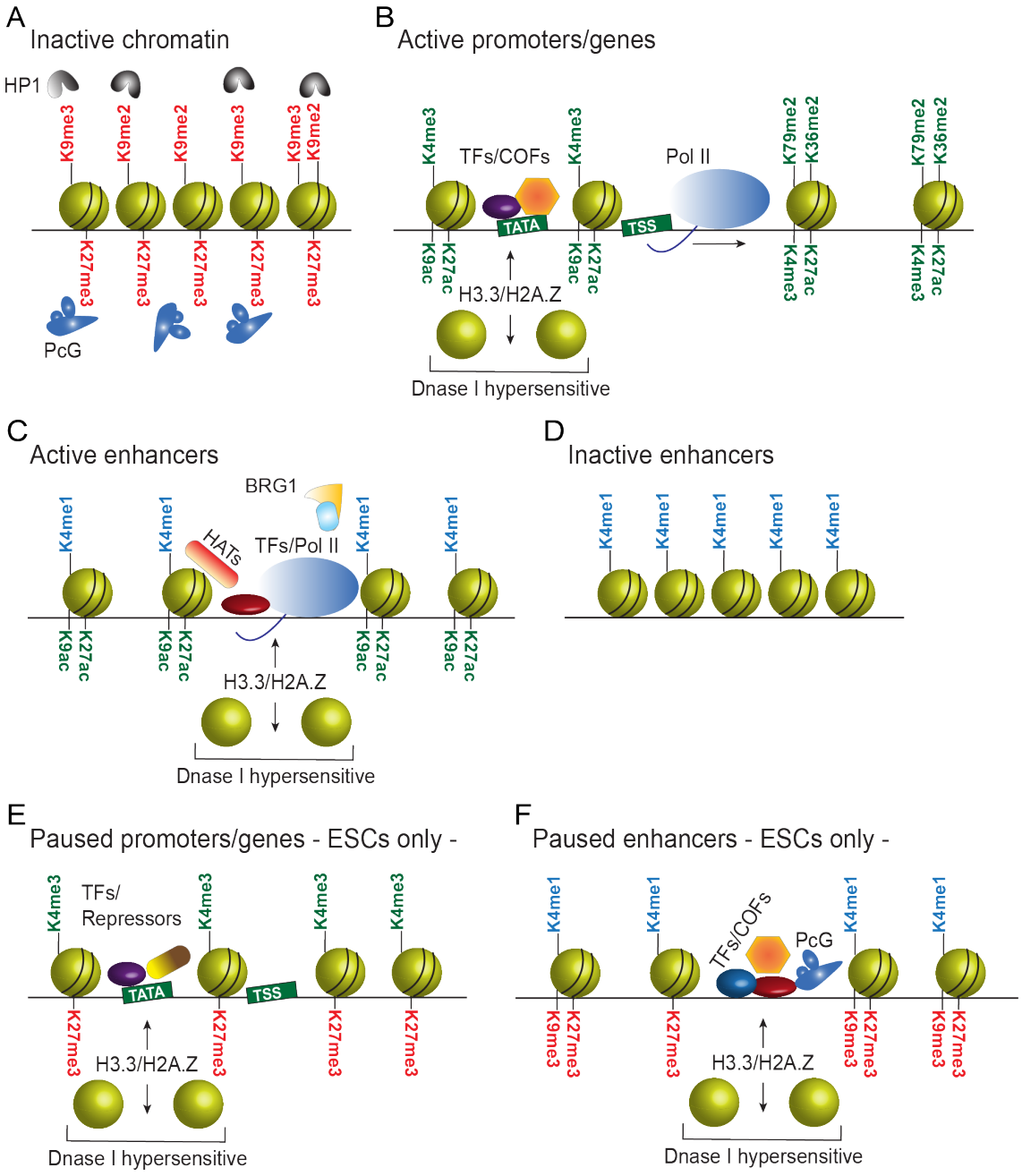


Figure 1.2. See next page for legend.

In contrast to promoters, distal regulatory elements such as enhancers are associated with H3K4me1, and have low levels of H3K4me3 (Figure 1.2C-D, 1.2F). Importantly, the high ratio of H3K4me1 to H3K4me3 is currently regarded as the major epigenetic feature that distinguishes enhancers from promoters (Calo and Wysocka, 2013). The other modifications present at enhancers are more dynamically regulated by the action of enzymes that specifically add ('writers') or remove ('erasers') the histone marks, with distinct chromatin signatures characterising different states of enhancer activity. On this basis, they have been further classified into active, primed and poised enhancers (Calo and Wysocka, 2013, Rada-Iglesias et al., 2011, Zentner and Scacheri, 2012).

Active enhancers (Figure 1.2C) are marked with both H3K4me1, H3K27ac and H3K9ac, contain H2A.Z/H3.3 and are highly sensitive to DNA nucleases such as DNase I (Bernstein et al., 2006, Rada-Iglesias et al., 2011). As with other histone modifications, acetylation functions not just by perturbing the chromatin structure as discussed above, but also by recruiting proteins ('readers') that recognise and bind the modified residues via specific domains. Thus, active enhancers are also bound by bromodomain-containing proteins, which include HATs themselves (e.g., EP300, CBP, PCAF, and Gcn5) and ATP-dependent remodellers (e.g. BRG1, and BRM) (Kouzarides, 2007, Bannister and Kouzarides, 2011). In addition, TFs and Pol II will also be present at these sites as already mentioned. Prior to activation, enhancers are instead devoid of any acetylation and are characterised only by the presence of H3K4me1, which 'primes' them for future use (Calo and Wysocka, 2013). However, how h3k4me1 is established at these sites in the first place is still unclear (Zentner and Scacheri, 2012). Chromatin at inactive enhancer is typically closed, with nucleosomes acting as 'gatekeepers' for TF binding (Figure 1.2D) (Barozzi et al., 2014, Tillo et al., 2010). At least in mouse and human ESCs, transcriptionally inactive

---

Figure 1.2. Chromatin functional elements.
(A-F) Different histone modifications, histone variants, DNA accessibility and protein binding events identify chromatin elements with distinct transcriptional activity. Note that promoters/genes and enhancers can be found in a paused state only in embryonic stem cells (ESCs). Light blue, H3K4me1; red, H3K27me3/ H3K9me2/H3K9me3; green, H3K4me3/H3K27ac/H3K9ac/H3K79me2/H3K36me2. HP1, Heterochromatin protein 1; PcG, Polycomb group proteins; TFs, transcription factors; COFs, transcriptional cofactors; Pol II, RNA polymerase II; HAT, histone acetyl transferase.

enhancers typically associated with developmental genes can also be found in a different state termed poised (Creyghton et al., 2010, Zentner et al., 2011).

Poised enhancers share many of the properties of active enhancers, such as low nucleosomal density, the presence of TFs and coactivators, and enrichment of H3K4me1 (Calo and Wysocka, 2013). However, they lack H3K27ac/H3K9ac and are characterised by the presence of the repressive marks H3K27me3/H3K9me3 and PcG proteins (Figure 1.2F). During differentiation into specific lineages, following external stimuli, the repressive histone modifications are removed by histone demethylases such as JMJD3, and the enhancers acquire the ability to drive gene expression (Buecker and Wysocka, 2012). Regardless of the different states identified, it is important to note that when considering enhancers and histone modifications in general, a question of causality still remains. To what extent histone marks define enhancers, or if they are a consequence of the establishment of the enhancer state itself is unclear.

In recent years, it became evident that the 3D chromatin structure also plays a key role in regulating gene expression. Enhancers are organised within larger scale loops that partition the chromosomes into segments, known as topological associated domains (TADs) (Ciabrelli and Cavalli, 2015). The compartmentalization of the genome into TADs is mediated by the chromosome-structuring proteins CTCF and cohesins, which bind to insulator/boundary elements largely conserved across cell types and species (Figure 1.3) (Long et al., 2016). Multiple studies have now demonstrated that TADs function as 'regulatory neighbourhoods' which confine enhancers activity to genes falling within the same TAD and limit the spread of chromatin modifications (Dixon et al., 2012, Nora et al., 2012). Indeed, the regulatory changes occurring at the level of enhancers-promoters interactions take place exclusively within the same TAD, and mutations that break the TAD boundaries have been associated with genetic diseases and cancer development (Ji et al., 2016, Lupianez et al., 2016). Some examples are congenital skeletal disorders, where TAD boundary duplication/inversion/deletion events cause the misregulation of genes such as *EPHA4*, *PAX3 or WNT6*; or the activation of the *PDGFRA* oncogene in gliomas, which results from loss of CTCF binding at the *PDGFRA* TAD boundaries upon aberrant DNA methylation of these sites (Flavahan et al., 2016, Valton and Dekker, 2016, Lupianez et al., 2015).

Considering the complex level of organisation just described, a central question in biology is how regions of chromatin are identified and bound by TFs, or, in other words, how enhancers function.

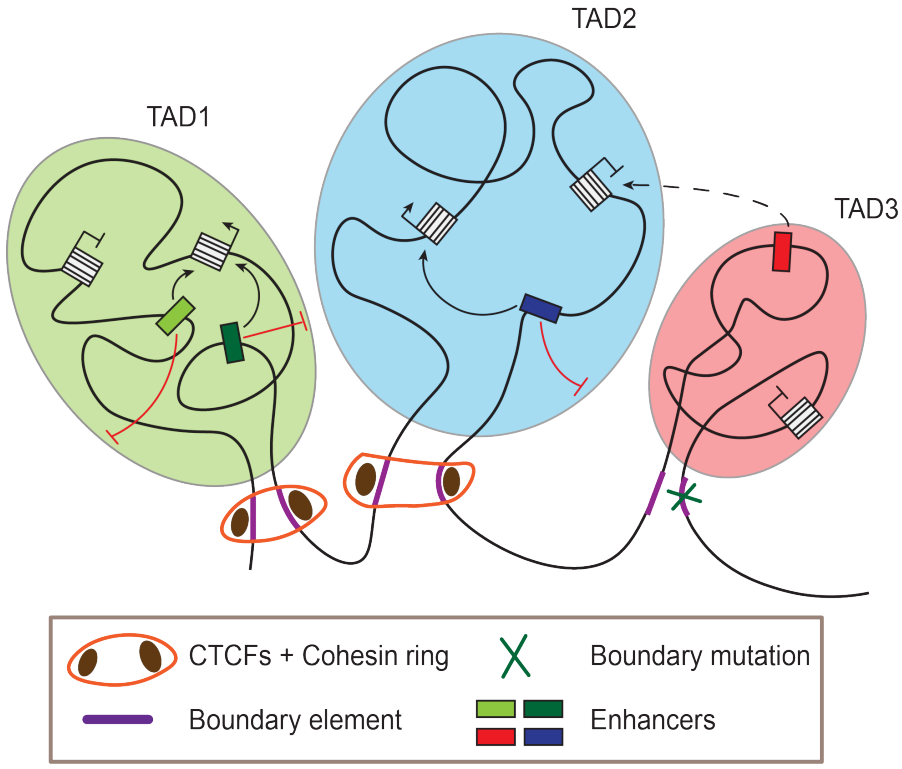


Figure 1.3. Organisation of chromatin into topological associated domains. Chromosomes are compartmentalized into topological associated domains (TADs) by CTCF and cohesins, which bind to highly conserved boundary elements. TAD boundaries restrict enhancers activity to genes within the same TAD (black lines versus red lines), and TAD disruption following boundary mutations can result in misregulated gene expression (dotted line). Adapted from (Long et al., 2016).

1.1.3 Transcription factors and models of enhancer function

Typically, enhancers contain clusters of TF recognition motifs and to become active they require the binding of multiple TFs, often including lineage-specific factors and signal-responsive transcription factors which are effectors of signal transduction pathways. The combinatorial function is essential to ensure spatiotemporal control of gene expression, and it is based on the fact that individual TFs are not able to successfully compete with nucleosomes for the underlying DNA (Reiter et al., 2017). In contrast, the binding of multiple TFs can be achieved by different mechanisms (Figure 1.4). According to the cooperativity model, the presence of several TFs involves a net increase in the affinity of the individual TFs for their motifs (Spitz and

Furlong, 2012). If the nucleosomes eviction is the result of the direct physical interaction between TFs, before or concurrent with DNA binding, the cooperativity will be considered 'direct' (Figure 1.4A). In the absence of protein-protein interactions, TFs might still be able to displace nucleosomes together via a sort of 'mass action', relying solely on the individual TF DNA affinities (Figure 1.4B). This 'collaborative' mechanism is referred to as 'indirect' cooperativity (Long et al., 2016). In contrast to what was just discussed, a third model envisages sequential rather than simultaneous TF binding, and is based on specialized TFs called 'pioneers' priming the enhancers for activation (Figure 1.4C). Pioneer TFs are argued to possess distinct biochemical properties that enable them to bind nucleosomal DNA within closed chromatin, thus facilitating the subsequent binding of additional TFs. Examples of TFs regarded as pioneers are FoxA or PU1 in mammals or Zelda in *Drosophila* (Zaret and Carroll, 2011).

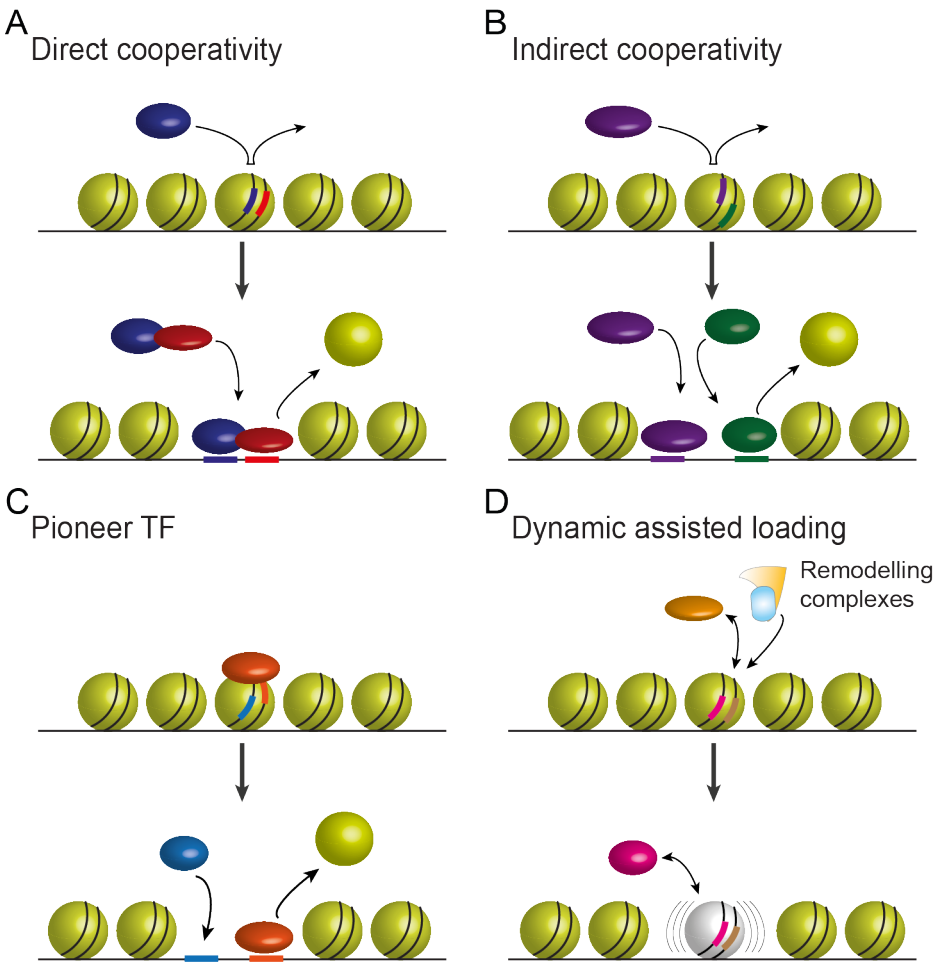


Figure 1.4. See next page for legend.

It is important to note that all these models imply long residency times for TFs at their binding sites and also long-lived nucleosome states. None of these models invokes a necessary role for ATP-dependent remodelling complexes in the process of chromatin penetration (Voss and Hager, 2014). According to the classical paradigm, once bound, the TFs recruit EP300, then other histones modifiers and additional COFs to finally mediate the formation of a stable, multi-protein complex in contact with Pol II at the core promoter. This static view of enhancer function has recently been challenged by *in vivo* imaging studies, which revealed that most TFs have residence times in the order of seconds, rather than minutes or hours (Sung et al., 2016). Following these findings, an alternative model, termed dynamic assisted loading has been proposed (Swinstead et al., 2016a). It suggests that at enhancer sites the initiating factor recruits ATP-dependent chromatin remodellers to transiently open chromatin, providing a window of opportunity for other TFs to bind (Figure 1.4D). As a consequence, the system is highly dynamic and involve a continuous cycling of nucleosomes, TFs and chromatin remodellers through different states. In such a scenario, interactions with histone modifiers and other COFs are also transient and not structurally constrained, leading to a more flexible form of enhancer–core promoter communication (Figure 1.5).

According to a recent interpretation of this view, gene transcription is the result of increased local concentration of cooperating factors (that is, TFs, COFs, Pol II, non-coding RNA, chromatin regulators and other proteins), that are 'phase separated' at gene regulatory elements (Reiter et al., 2017). In the cytoplasm and nucleolus, high densities of nucleic acids and proteins interacting with each other form membraneless organelles where essential biochemical reactions are confined (Bergeron-Sandoval et al., 2016). Similarly, multi-molecular complexes would assembly at enhancers to compartmentalise transcriptional activity, providing a mechanistic explanation to the findings that enhancers activate genes in bursts, and

---

Figure 1.4. Mechanisms for transcription factor binding and nucleosome eviction at enhancer sites.
(A-D) Different models to explain the binding of multiple transcription factors (TFs) at inactive enhancers in the context of chromatin. Note that the mechanism in (D) is the only one which involves the requirement of chromatin remodellers to mediate nucleosome eviction. The continuous cycling of nucleosomes and TFs on and off DNA is represented with grey lines around the remodelled nucleosome (grey) and bi-directional arrows, respectively.

that one enhancer can simultaneously activate two core promoters (Fukaya et al., 2016, Hnisz et al., 2017). In this model, the concentration of components and their valency, that is the repertoire of reversible modifications (such as phosphorylation or acetylation) which can affect their dynamic interactions, are key parameters in determining the phase separation. As a consequence, changes in component number or valency directly affect the transcriptional output, and this seems to be particularly the case of super-enhancers (SE), which are cluster of enhancers thought to regulate the expression of cell-specific genes (Hnisz et al., 2013, Hnisz et al., 2017). According to this hypothesis, at SE but not at typical enhancers, high levels of transcriptional activity would be rapidly achieved when one of the variables just defined exceed specific threshold values, and then stably maintained over time. In turn, this mechanism would limit transcriptional noise and ensure that genes responsible to maintain cell identity are robustly transcribed (Hnisz et al., 2017).

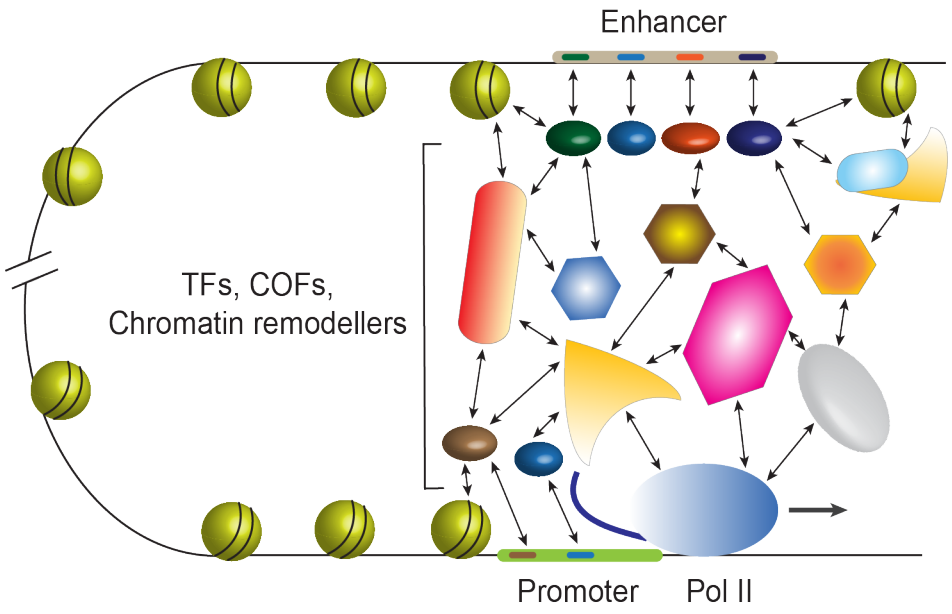


Figure 1.5. A dynamic model for transcriptional regulation.
Recent findings suggest a more flexible model of transcriptional regulation compared to the classic one illustrated in Figure 1.1. Transcription activation is seen as the result of transient protein-protein interactions, increased local concentrations and chromatin remodelling events. TFs, transcription factors; COFs, transcriptional cofactors; Pol II, RNA Polymerase II. Adapted from (Reiter et al., 2017).

1.1.4 Regulation of RNA Polymerase II by pause-release

The simple presence of the transcriptional apparatus at a core promoter does not necessarily signify the expression of the related gene. At the level of Pol II activity, transcription is regulated at either of two stages: recruitment of Pol II or Pol II promoter proximal pausing and subsequent release (Figure 1.6). For genes regulated via Pol II recruitment, the limiting step is the formation of the pre-initiation complex at their promoter DNA, which is mediated by sequence-specific TFs as discussed above (Nechaev and Adelman, 2011). The preinitiation complex consists of Pol II and the associated GTFs (such as TFIID) that recognise specific DNA sequence elements within the core promoter, such as the TATA element and the downstream promoter element (DPE) (Juven-Gershon and Kadonaga, 2010). Pol II is composed of 12 subunits (RPB1-12), the largest of which (RPB1) contains a carboxyl-terminal region (CTD) made of several heptapeptide repeats (YSPTSPS) (Meinhart et al., 2005, Cramer, 2004). Upon phosphorylation of the CTD at Serine 5 by the CDK7 kinase subunit of the TFIIH complex, the promoter is cleared and transcription of the gene body rapidly initiates (Compe and Egly, 2012). In some cases, however, Pol II arrests just after the promoter escape, typically 20-60 bp downstream of the transcription start site (TSS) (Figure 1.6). Since there is now evidence that this paused Pol II can be found at about one third of all genes, its release from the proximal promoter represent a major mechanism to control gene expression (Adelman and Lis, 2012, Fuda et al., 2009, Levine, 2011).

Typically, paused Pol II is enriched for phosphorylation at Serine 5 of the CTD repeats and is associated with two pausing factors that held it at promoters, the DRB sensitivity inducing factor (DSIF) and the negative elongating factor (NELF) (Figure 1.6). How these factors are recruited and how Pol II pausing is established in the first place is still unclear. A role in the process has been hypothesised for two transcription factors, the GAGA factor (GAF) and the Motif 1 binding protein (M1BP), which recognise short sequences commonly present at promoters such as the GAGA motif and the Motif 1 (Fuda and Lis, 2013, Ohler et al., 2002). GC-rich sequences at the level of promoters have also been suggested to act as ‘speed bumps’ which slow down Pol II and allow the binding of DSIF and NELF (Nechaev et al., 2010). Additionally, other studies have proposed that sequence specific regulatory factors such as the glucocorticoid receptor and the estrogen receptor directly recruit NELF

(Liu et al., 2015), linking the establishment of paused Pol II to the appropriate signalling pathways.

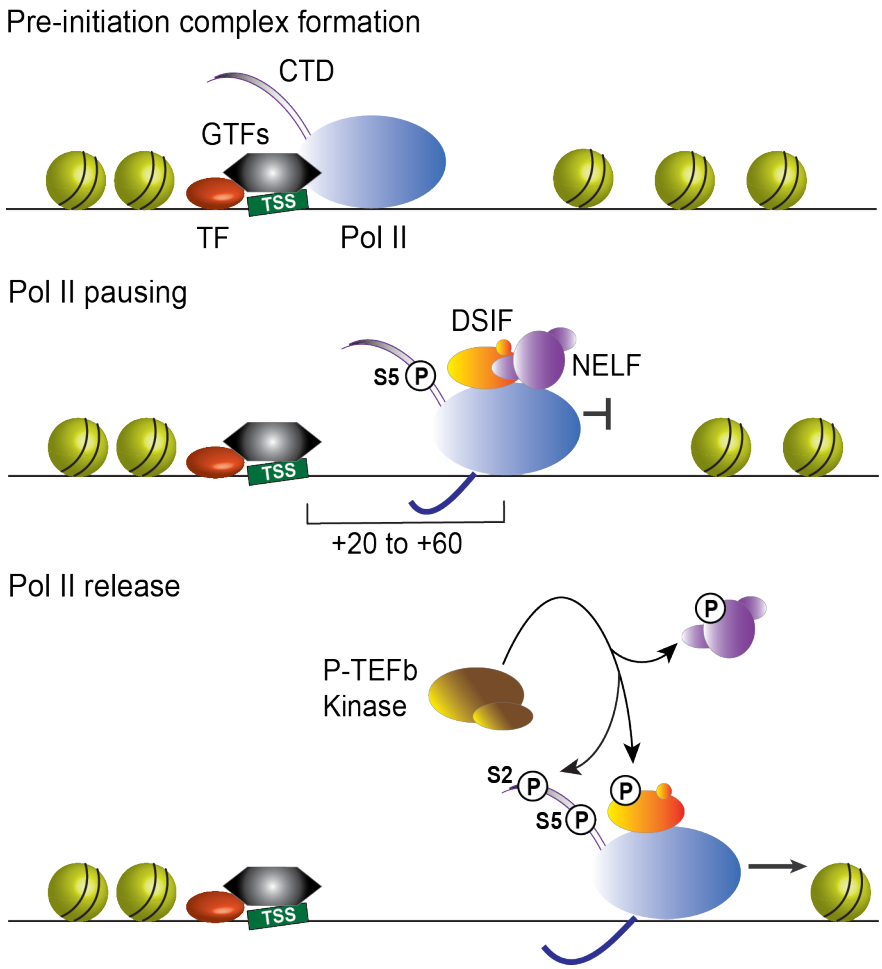


Figure 1.6. Regulation of RNA Polymerase II transcriptional activity by pause-release.
The cartoon illustrates the mechanisms of establishment and release of paused RNA Polymerase II (Pol II). The pre-initiation complex formation involves the recruitment at the promoter region of Pol II and the general transcription factors (GTFs) by gene-specific transcription factors (TFs). The transcription start site (TSS) is shown in green. Pol II pausing occurs after a short transcript (blue) is produced and is mediated by the DRB sensitivity inducing factor (DSIF) and the negative elongating factor (NELF). Paused Pol II is phosphorylated at Serine 5 of its carboxy-terminal domain (CTD). The recruitment of the P-TEFb kinase triggers the release of Pol II from pausing. P-TEFb phosphorylates Serine 2 of the Pol II CTD and the NELF-DSIF complex, causing the dissociation of NELF and transforming DSIF into a positive elongation factor.

For genes with paused Pol II, the rate-limiting step in transcription is represented by the subsequent release, which requires the kinase activity of the positive transcription factor b (P-TEFb). P-TEFb is recruited to gene promoters either directly through the association with a specific TF, or indirectly by transcription co-regulators such as the Mediator complex and chromatin factors such as the bromodomain protein BRD4 (Rahl et al., 2010, Yang et al., 2005, Takahashi et al., 2011). The recruitment of P-TEFb causes the phosphorylation of the repressive DSIF-NELF complex, leading to the dissociation of NELF from Pol II and converting DSIF to a state that promotes Pol II elongation (Figure 1.6). Pol II itself is also phosphorylated by P-TEFb at Serine 2 of the CTD repeats, creating a platform for the binding of RNA-processing factors and chromatin-modifying enzymes that facilitate productive RNA synthesis (Adelman and Lis, 2012). Additionally, P-TEFb and TFIIH/CDK7 can phosphorylate Serine 7 of the Pol II CTD, and this modification seems to be require for the transcription of specific genes, such as the ones encoding snRNAs (Egloff et al., 2012). Importantly, recent studies have found that treatment of cells with a P-TEFb inhibitor blocks the progression of Pol II into elongation for most genes, both in *Drosophila* and in mammals, suggesting that the control of the early elongation complex by DSIF and NELF could represent a general step of transcription. Thus, the rate of P-TEFb recruitment would be crucial for determining the appearance of paused Pol II at some genes but not at others. Paused Pol II would accumulate at those promoters where P-TEFb recruitment is a slow event, but it would not be detected in cases where P-TEFb recruitment to promoters immediately follows transcription initiation, leading to a rapid release of Pol II into the gene body (Adelman and Lis, 2012).

If considering the genes containing paused Pol II at their promoter regions, it can be noted that they mostly fall into the categories of environmental or developmental regulated genes. Classical examples are represented by the heat shock (Hsp) genes in *Drosophila*, for which the recruitment of P-TEFb is dependent on a signal-regulated transcription factor sensing heat shock (HSF). In mammals, paused Pol II has been observed at immediate early genes (IEGs) such as *MYC*, *FOS*, *JUN* and *JUNB*, and the mechanisms of release in response to serum growth factors have been extensively characterised (Liu et al., 2015). The classic functional interpretation of this observation is that paused Pol II ‘poises’ genes for a rapid and synchronous induction of transcription in response to extracellular stimuli. Pioneer

studies of the heat shock genes suggest that the presence of paused Pol II significantly accelerates the timing of induction compared with promoters lacking Pol II. This advantage could be crucial, not just in the context of conferring upon the organism resistance to stress, but also during development and differentiation, when programmes of gene expression need to be efficiently activated in a well-defined spatiotemporal window (Adelman and Lis, 2012). However, not all rapidly induced genes are associated with paused Pol II, rather its presence could even slow-down the rate of mRNA synthesis, according to more recent studies ((Ehrensberger et al., 2013); Patrick Cramer, personal communication). Alternatively, the presence of Pol II at developmental genes might function as 'checkpoint' to suppress the transcriptional noise potentially deriving from a stochastic recruitment of Pol II to the core promoter, thus ensuring a precise and reliable transcription (Levine, 2011). In conclusion, the question of whether pausing serves different roles at different functional classes of genes is unresolved. The understanding of how Pol II in general is regulated on a genome-wide level is also incomplete.

As mentioned earlier, in this thesis I propose to use the NODAL/Activin signalling pathway as a model system to study how the three key players of transcription described so far, that are the chromatin, the TFs and Pol II, function and interact to regulate complex programmes of gene expression. Since NODAL and Activin are members of the TGF-β superfamily of ligands, I will now introduce the TGF-β superfamily signalling pathway, describing the molecular functions and the regulatory mechanisms of the different pathway components.

1.2 The TGF-β superfamily signalling pathway: a brief overview

1.2.1 TGF-β superfamily signalling pathway components

The TGF-β superfamily ligands are a large family of more than 30 growth factors which can be divided into 6 subgroups, the TGF-βs, Activins, NODAL, Bone Morphogenetic Proteins (BMPs), Growth Differentiation factors (GDFs) and anti-Mullerian hormone (AMH), as well as the three antagonists LEFTY1, LEFTY2 and INHA (Figure 1.7). These proteins are essential for embryonic development, tissue homeostasis and stem cell function. When misregulated, they also play crucial roles in both tumour development and metastasis, and in many other diseases such as

inherited connective tissues disorders and fibrosis (Caja et al., 2012, Massague, 2008, Massague, 2012, Wu and Hill, 2009, Miller, 2016).

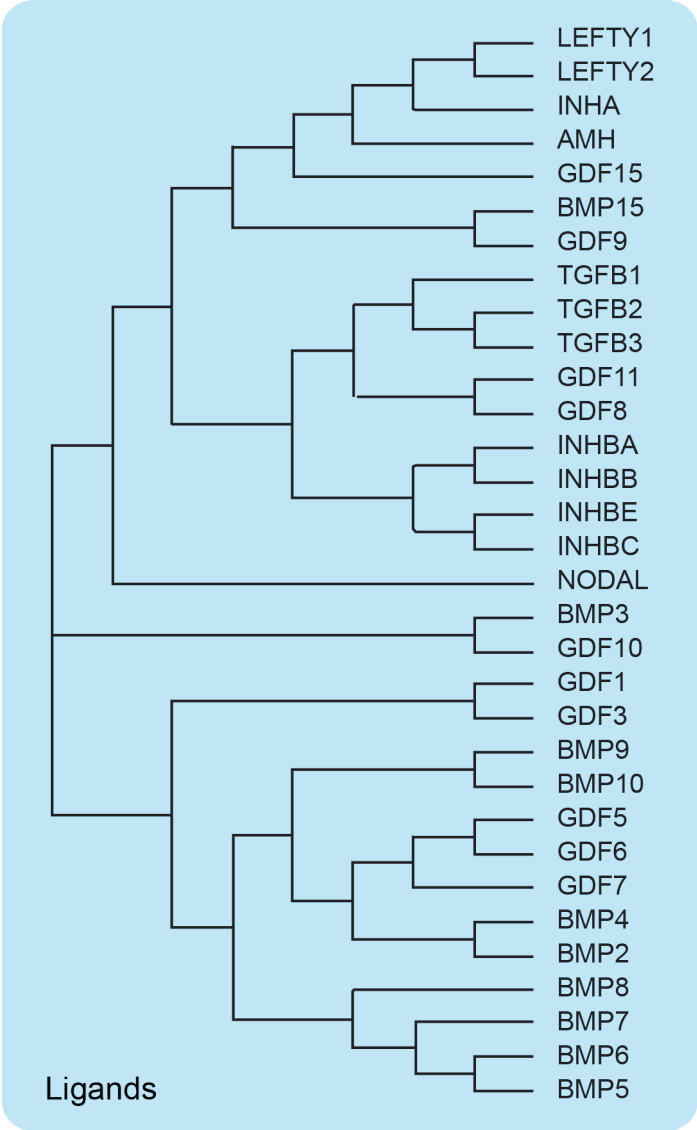


Figure 1.7. The ligands of the TGF-β superfamily signalling pathway.
Phylogenetic tree of the 32 ligands in the TGF-β superfamily. Ligands are classified into 6 subgroups: the TGF-βs, BMPs, GDFs, Activins, NODAL and AMH. LEFTY1, LEFTY2 and INHA are antagonists of the pathway. Adapted from (Miller, 2016).

All the TGF-β superfamily ligands act as dimers, which can be either homo or etherodimers, and signal through the same mechanism by binding to an heterotetrameric complex composed of two types of serine/threonine kinase receptors, a type I and a type II (Antsiferova and Werner, 2012, Mueller and Nickel, 2012, Peng et al., 2013, Tanaka et al., 2007). In mammals, there are seven different

type I receptors, also known as activin receptor like kinases (ALK1-7), and five type II receptors, with individual ligands binding different combinations (Figure 1.8, 1.9). For some ligands, additional co-receptors are required to facilitate this process, and a well characterised example is TDGF1 (also known as CRIPTO) for the NODAL ligand (Wakefield and Hill, 2013). Upon ligand binding, the constitutively active type II receptor phosphorylates the type I receptor on several serines and threonines in the intracellular glycine- and serine-rich (GS) domain, causing a conformation change that activates the type I receptor kinase activity. The phosphorylated residues in the type I GS domain also function as a binding site for the main transcriptional effectors of the pathway, the receptor regulated SMADs (R-SMADs). There are five R-SMADs, SMAD 1, 2, 3, 4, 5 and 9, and the individual type I receptors have distinct preferences for different R-SMADs, thus enabling ligand-specific responses (Wu and Hill, 2009). Once phosphorylated the R-SMADs can form heteromeric complexes with the common mediator SMAD, SMAD4, which accumulate in the nucleus to regulate transcription together with cofactors, both positively and negatively (Figure 1.9). Different activated SMAD complexes have distinct DNA sequence specificities and interact with distinct co-factors in different contexts (see Sections 1.2.2 and 1.4), thus accounting for the diversity observed when comparing the transcriptional responses of individual TGF-β ligands (Massague, 2012).

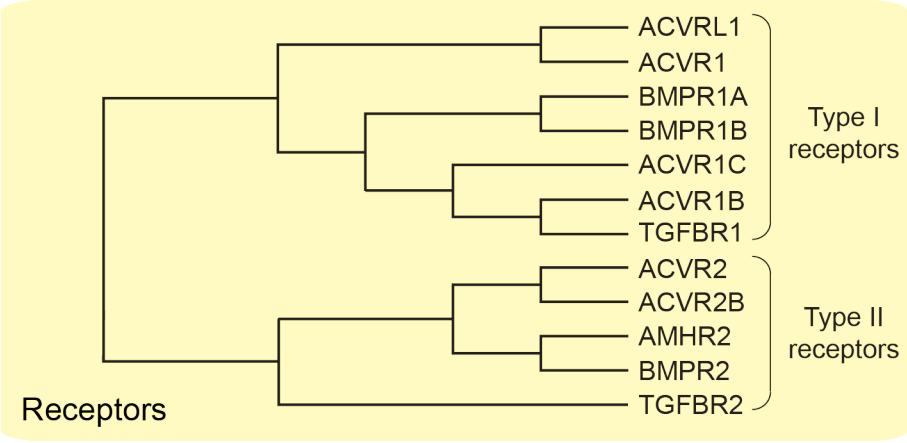


Figure 1.8. The receptors of the TGF-β superfamily signalling pathway.
Phylogenetic tree of the 12 receptors for the TGF-β superfamily ligands. Receptors are divided into 2 classes – Type I and the Type II – as shown. Adapted from (Miller, 2016).

In the canonical view of the TGF-β superfamily signalling pathway, TGF-βs, Activins and NODAL activate SMAD2 and 3, whereas SMAD 1, 5 and 9 are the effectors of BMPs, GDFs and AMH (Figure 1.9). The TGF-βs signal through the type I receptor TGFBR1 (ALK5) and the type II receptor TGFBR2, whilst Activins and NODAL bind the type I receptors ACVR1B (ALK4) and ACVR1C (ALK7) along with the type II receptors ACTR2A and ACTR2B. These last type II receptors are used also by the majority of BMPs and GDFs, in addition to the BMPR2; whilst BMPR1A (ALK3) and BMPR1B (ALK6) act as their primarily type I receptors. For BMP9 and BMP10, the type I receptor is ACVRL1 (ALK1). Finally, ACVR1 (ALK2) and AMHR2 are the type I and type II receptors for AMH (Massague, 2012, Wakefield and Hill, 2013). It is important to note that the division of the pathway into two distinct branches is an oversimplification, since some GDFs, such as GDF8, 9 and 11, signal through SMAD2/3 by binding to bind the type I receptors ACVR1B (ALK4) (Schmierer and Hill, 2007). It is also becoming clear that many different ligand-receptor combinations may occur, creating a great potential diversity in the responses to the individual ligands. TGF-β, for instance, can also lead to the phosphorylation of SMAD1/5/9 through the formation of mixed receptor complexes (Goumans et al., 2003, Gronroos et al., 2012, Wu and Hill, 2009). Moreover, although the TGF-β ligands signal mainly via the R-SMAD, the activation of others ‘non-canonical’ pathways, such as the MAPK and PI3K, has been reported downstream of the receptor complexes (Heldin and Moustakas, 2016, Massague, 2012). Nevertheless, here I will focus only on the SMAD pathway, as that is the subject of my thesis.

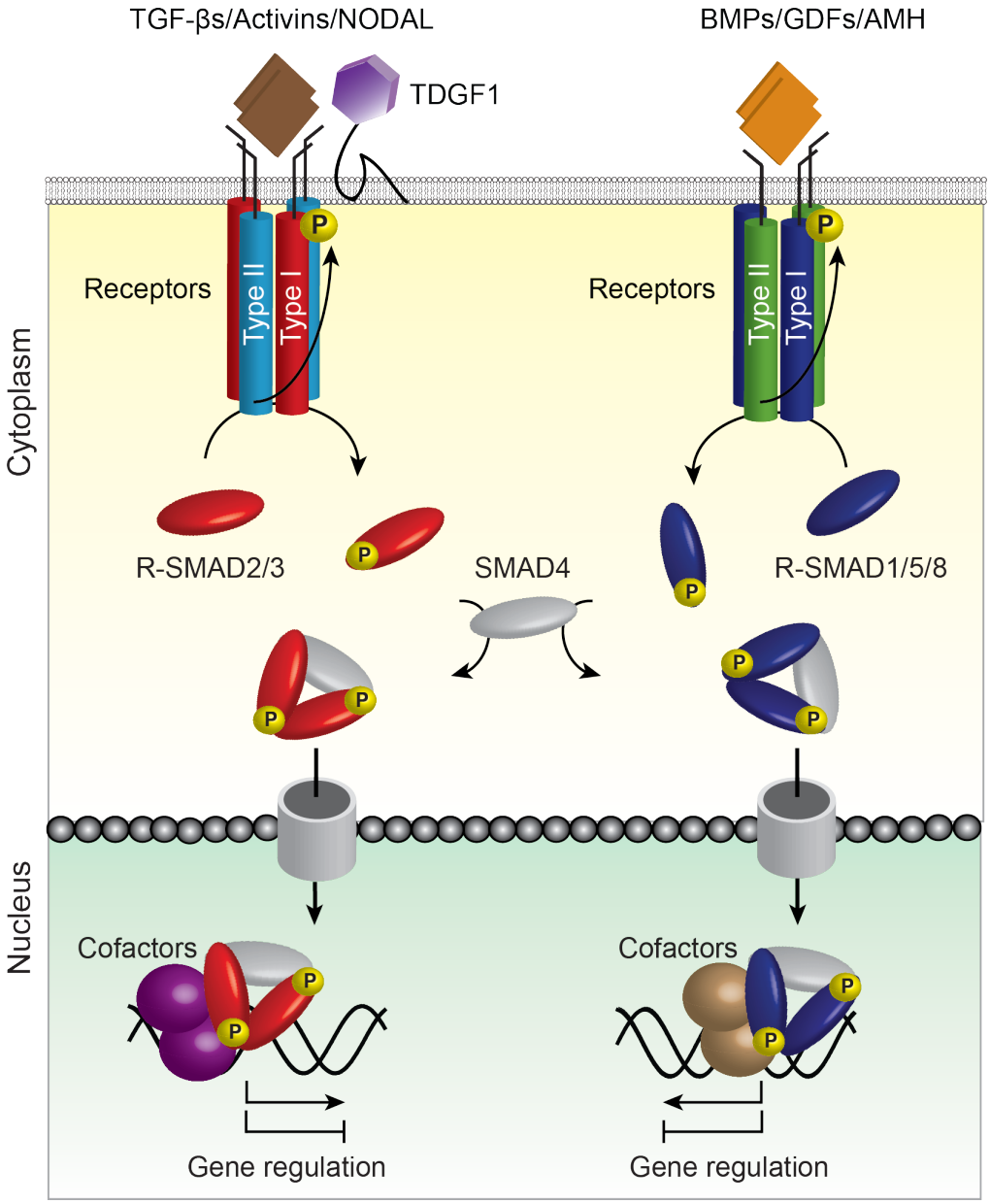


Figure 1.9. The canonical TGF-β superfamily signalling pathway.
Schematic of the signalling pathway downstream to the TGF-β superfamily members. Binding of the ligands to the type I and type II receptors causes the type II receptor to phosphorylate the type I receptor. The activated type I receptor then phosphorylates the R-SMADs, which in turn form complexes with SMAD4 and accumulate in the nucleus to regulate transcription together with cofactors. In the classical view of the pathway, the TGF-βs, Activins and NODAL signal through SMAD2/3, whilst the BMPs, GDFs and AMH activate SMAD1/5/8. NODAL additionally required the presence of the co-receptor TDGF1. Adapted from (Gaarenstroom and Hill, 2014).

1.2.2 The SMAD structure

The R-SMADs and SMAD4 have two highly conserved Mad Homology (MH) domains connected by a less conserved linker region (Figure 1.10).The N-terminal MH1 domain is critical for DNA binding and nuclear export, whilst the C-terminal MH2 domain mediates the large majority of the protein-protein interactions, including SMAD binding to the receptors, the formation of SMAD complexes and SMAD interaction with other TFs, transcriptional co-activators and repressors (Ross and Hill, 2008).The linker region is rich in proline and serine residues and acts as a substrate for several PTMs, which can in turn affect SMADs activity in various ways (for more details about PTMs, see Section 1.2.3). As mentioned earlier, different type I receptors activate distinct R-SMADs, and this specificity is determined by the unique residues present both in individual type I receptors and R-SMAD MH2 domains (Chen et al., 1998). The type I receptors phosphorylate the R-SMADs at two Serine residues in their extreme C-terminus in a SXS motif where X is Methionine or Valine, which in turn allows the R-SMADs to interact with each other and/or with SMAD4 though their MH2 domains (Shi and Massague, 2003, Chacko et al., 2004). The stoichiometry of the resulting complexes is still unclear: since in crystal structure studies the MH2 domains always form trimers, they are generally considered to occur as such with two R-SMADs and SMAD4. However, their existence as dimers has also been reported (Wu et al., 2001, Inman and Hill, 2002, Randall et al., 2004, Hill, 2016).

Once in the nucleus, SMAD4 and all the R-SMADs can directly bind to DNA via the MH1 domain, with the only exception being SMAD2. This is due to a 30 amino acids insertion in the SMAD2 MH1 domain, which prevents it contacting the DNA directly (Ross and Hill, 2008). SMAD3 and SMAD4 preferentially recognize the specific sequence GTCT (or its reverse complement AGAC), also known as the SMAD binding element (SBE) (Zawel et al., 1998). Interestingly, a splice form of SMAD2 which is expressed in early embryonic development and lacks the 30 amino acids insertion, termed SMAD2ΔExon3, also interacts with the DNA via the SBE. SMAD1/5 instead binds to GC-rich GRCGNC motifs, which are often found in conjunction with a canonical SBE separated by 5 nucleotides, which in turn likely binds SMAD4 (Gaarenstroom and Hill, 2014). Despite the existence of these consensus sites which have been confirmed *in vivo* by investigating SMAD binding

genome-wide, it is important to highlight that the affinity of SMADs for DNA is weak ($K_d \approx 1 \times 10^{-7}$ M) (Shi et al., 1998). Thus, they frequently require the interaction with additional DNA-binding proteins in order to control gene expression. This is absolutely necessary in the case of SMAD2–SMAD4 complexes, since SMAD2 can not bind DNA. The first SMAD co-factor to be characterized was FOXH1 (Chen et al., 1996), and numerous others have been identified since. Their roles in regulating SMAD transcriptional activity will be discussed later in more detail.

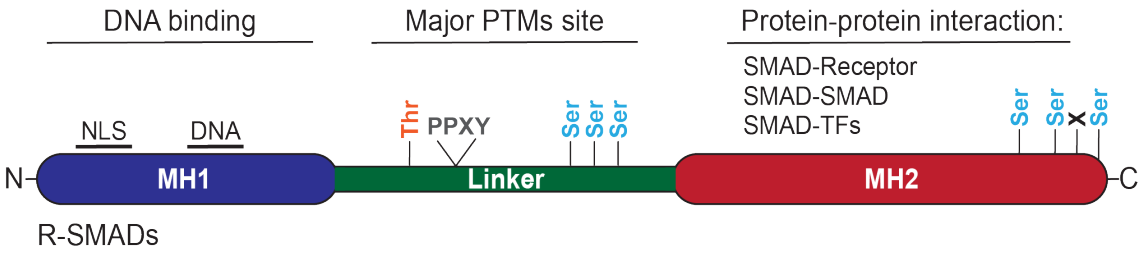


Figure 1.10. R-SMAD functional domains and major sites of post-translational modifications.
R-SMADs comprise three functional domains: the N-terminal MH1 domain (dark blue), the Linker domain (green) and the C-terminal MH2 domain (red). The MH1 is required for DNA binding (in all R-SMADs except SMAD2) and nuclear import, whilst the MH2 domain mediates protein-protein interactions and gets phosphorylated by the type I receptor at the Serine residues at the extreme C-terminus. Other post-translational modifications (PTMs) often occur in the Linker region at the Threonine and Serine residues indicated and at the PPXY motif. NLS, nucleus localization sequence.

1.2.3 General principles of TGF-β superfamily signalling regulation.

At a first glance, the TGF-β superfamily pathway as outlined so far seems quite simple and linear, in sharp contrast with the complexity and the diversity of the biological responses controlled by the different ligands. This apparent contradiction, however, begins to resolve once one considers that each of key pathway components - ligands, receptors, R-SMADs - is subject to multiple levels of regulation. In the effort of producing a full picture of the TGF-β superfamily signalling pathway, the mechanisms of ligand availability, receptors dynamics, SMAD intracellular trafficking, SMAD post-translational modifications (PTMs) and negative feedbacks must be taken into account and are briefly introduced here (Figure 1.11).

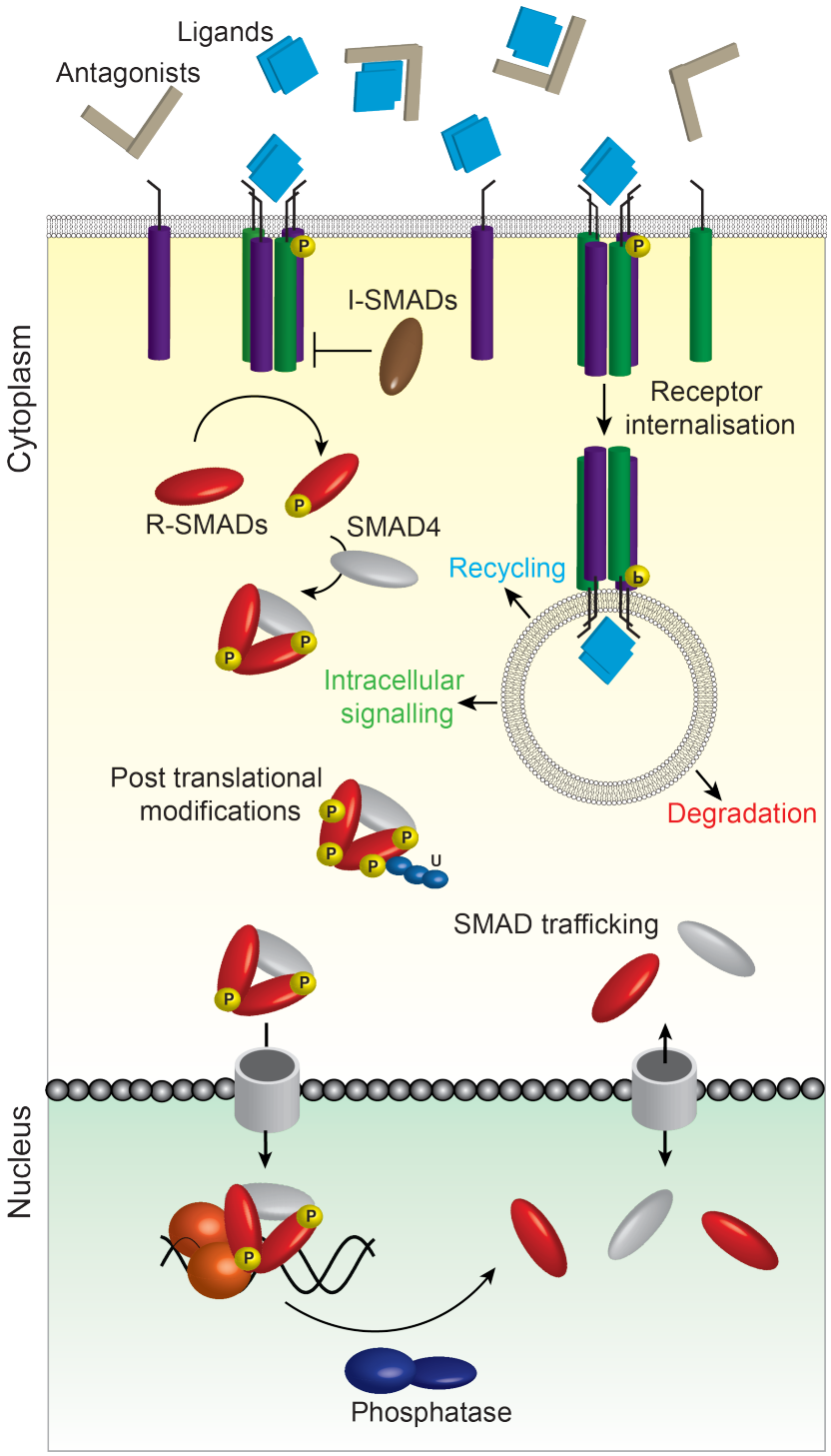


Figure 1.11. Regulatory dynamics of TGF-β superfamily signalling pathway.
The cartoon illustrates the main mechanisms regulating the activity of the TGF-β superfamily pathway components. In the extracellular space, antagonists prevent ligand binding to the receptors, whilst at the level of the cell membrane bound- and unbound receptors are constantly internalised by endocytosis. Internalised, active receptor complexes can signal from endosomes, being recycled to the cell surface or destroyed. Receptor activity is also negatively regulated by the Inhibitory SMADs (I-SMAD).
Figure 1.11 continued on next page.

All TGF-β ligands are synthesized as precursors which form dimers via the formation of disulphide linkages. The mature ligands are then cleaved from the precursors by pro-protein convertases such as Furin and get secreted as latent forms, with their propeptides not covalently attached, called the latency associated peptide (LAP) (Heldin and Moustakas, 2016). For some family members, such as the TGF-β itself, the LAP acts as an extracellular inhibitor by blocking the binding of circulating ligands to the receptors and mediating their deposition at the level of ECM, where the activity of other proteases is further required for the final activation (Moustakas and Heldin, 2009). For other ligands, the signalling activity depends instead on the presence of specific antagonists, which prevent the binding to the receptors through direct or indirect mechanisms. Examples in this sense are LEFTY1 and LEFTY2, which act as soluble antagonists of NODAL, Noggin and Chordin which inhibit BMPs and Follistatin which inhibits Activins (Wakefield and Hill, 2013).

It is now becoming clear that the levels of extracellular ligands are also constantly monitored by the signalling pathway via complex and only partially characterized mechanisms of receptor trafficking, of which the one concerning the TGF-β receptors are the best known. In the absence of signalling, TGF-β receptors are continuously internalized by endocytosis and recycled to the membrane via the small GTPase RAB11, in order to maintain their level constant at cell surface (Di Guglielmo et al., 2003, Mitchell et al., 2004). In the presence of signalling, the internalised receptors undergo different fates (Figure 1.11). Endosomes association with the anchoring protein ZFYVE9 (known also as SARA) promotes the recruitment of the R-SMADs to the activated receptors, allowing active signalling to occur in the cytoplasm as well (Tsukazaki et al., 1998). Alternatively, the internalised receptors can be recycled to the cell surface as just discussed, or be targeted for degradation upon ubiquitination by an E3 ubiquitine ligase (eg. SMURF1/2), which is recruited to

---

Figure 1.11 continued.
The pool of monomeric SMADs in the cytoplasm depends on its continuous shuttling into and out of the nucleus, and it is maintained during signalling by phosphatases dephosphorylating the R-SMAD-SMAD4 complexes in the nucleus. SMAD levels and transcriptional activity can also be affected by post translational modifications such as phosphorylation or ubiquitination mediated by enzymes responding to other signalling pathways. See main text for details.

the activated type I receptors by the inhibitory SMAD SMAD7 (Heldin and Moustakas, 2016). Due to the constant process of internalization, the amount of receptors present at the plasma membrane each moment will be a function of their rates of synthesis, degradation and recycling. Characterising how these processes are regulated for the different TGF-β superfamily ligands could be crucial to understand the differences seen in the individual responses. Recent work from our lab shows for example that TGF-β receptors are rapidly depleted from the plasma membrane following ligand treatment and only slowly re-accumulate, providing an explanation of why cells are refractory to further stimulation for a long period of time. In contrast, Activin and BMP receptor levels at the cell surface are fairly constant in the long term, such that an increased duration of exposure to these ligands leads to an increase in downstream signalling. Nevertheless, if this is due to a slower internalisation, or a faster re-accumulation at the cell surface is still unclear (Vizan et al., 2013); Miller, manuscript in preparation).

Similarly to the receptors, the R-SMADs and SMAD4 also have a highly dynamic behaviour, as they constantly shuttle between the cytoplasm and the nucleus, independently of pathway activation (Figure 1.11). In the presence of signalling, however, SMAD complexes accumulate in the nucleus, since they cannot be exported to the cytoplasm and have a higher import rate than monomeric SMADs (Schmierer et al., 2008). The R-SMADs are then dephosphorylated in the nucleus by several yet poorly characterised phosphatases, and exported back into the cytoplasm (Bruce and Sapkota, 2012, Schmierer et al., 2008). As a result, the levels of activated SMADs in the nucleus continuously reflect the levels of active receptor in the cytoplasm, in turn enabling cells to sense the dynamics of ligands over time and translate them into specific outputs. This nucleocytoplasmic shuttling, together with the lack of amplification steps in the signalling cascade, provides a functional explanation as to why many TGF-β ligands such as NODAL and the BMPs can act as morphogens during development. In the context of signalling gradients, cells must in fact be able to interpret differences in time and dose of exposure to the ligands to produce distinct responses (Schmierer and Hill, 2007); (Ashe and Briscoe, 2006).

The total levels of SMADs, their localisation in the cells and ultimately their transcriptional activity can also be affected by the presence of PTMs, which are often modulated by other signalling pathways (Figure 1.11). The most studied PTM is the phosphorylation of the conserved Ser/Thr-Pro motifs present in the linker region of

all the R-SMADs. This modification can be mediated by the MAPKs upon growth factors signals and by cyclin-dependent kinases (CDKs), whose activities directly depend on the cell cycle (Macias et al., 2015). The glycogen synthase kinase 3β (GSK3β), which in turn responds to a variety of extracellular stimuli, can also phosphorylates the R-SMAD at Ser/Thr residues in the linker region and in other domains (Macias et al., 2015, Pauklin and Vallier, 2015). The function of the linker phosphorylation is still not entirely understood. Initially, it was shown to 'prime' activated SMADs for degradation, by enabling the recruitment of E3 ubiquitin ligases, namely SMURF1 for SMAD1/5 and NEDDL4L for SMAD2/3 (Gao et al., 2009, Sapkota et al., 2007). However, when the phosphorylation occurs at the Thr 179 of SMAD3, the consequent monoubiquitination lead to SMAD3–SMAD4 complexes dissociation, rather than priming them for degradation (Tang et al., 2011). Other studies instead suggested that the linker phosphorylation could promote SMADs transcriptional activity. Phosphorylation of the Ser/Thr-Pro sites by CDK8/9, for instance, enables the binding of cofactors required for the transcriptional action, like YAP in the case of SMAD1/5 or Pin1 in the case of SMAD2/3 (Alarcon et al., 2009, Aragon et al., 2011). A possible explanation to account for the multiple functions proposed is that phosphorylation of different sites might have different effects depending on the cellular context, and the existence of a phosphoserine code coupled with a set of code 'readers' has been proposed (Aragon et al., 2011).

The last aspect in the regulation of the receptors-SMADs pathway to be considered is represented by negative feedback mechanisms. The principle negative regulators are the inhibitory SMADs, or I-SMADs, namely SMAD6 and SMAD7, whose transcription is induced in response to the ligands (Figure 1.11). In terms of structure, the I-SMAD lack the MH1 domain and therefore cannot bind the DNA (Aragon et al., 2012). They are thought to inhibit the pathway in several ways, acting both at the receptor level and in the nucleus. In the first case, they recruit E3 ubiquitin ligases or phosphatases to the receptors, and they can also compete with R-SMAD for binding to the receptors themselves. In the second case, they interfere with the transcriptional activity by recruiting repressors or disrupting SMAD-DNA binding (Itoh and ten Dijke, 2007, Moustakas and Heldin, 2009, Yan et al., 2009). In zebrafish embryos and in mouse ESC, SMAD7 activity is in turn negatively regulated by the E3 ligase RNF12, suggesting that counteracting mechanisms are necessary to fine-tune the responses to TGF-β ligands in different cellular contexts (Zhang et al., 2012).

From what has been discussed so far, it emerges the great effort put over the years to identify the different pathway components and to elucidate their structures, functions and mechanisms of regulation. However, despite the molecular framework of the TGF-β superfamily signalling pathways being much clearer now than when they were first discovered in the 1980s, the puzzle of how cells read TGF-β signal still remains. In particular, how ligands can mediate different, even opposite effects depending on cell types or conditions it is still unclear (Massague, 2012). To unravel this question, in this thesis we decided to focus on NODAL, which best exemplifies a TGF-β ligand able to control a broad range of biological processes in both physiological and pathological contexts.

1.3 The biological roles of NODAL/Activin signalling

As outlined above, NODAL signasl through the same receptors as the Activins and shares the downstream effectors SMAD2 and SMAD3. As a consequence, the individual pathways are usually regarded as undistinguishable, and I will henceforth refer to them as the NODAL/Activin signalling pathway. These ligands are also mostly considered to have similar functions, despite their patterns of tissue expression often being different (Pauklin and Vallier, 2015). Indeed, recombinant Activin A is commonly used in *in vitro* studies, such as ESC differentiation assays, to mimic NODAL activity *in vivo*. With this purpose it has also been used in this thesis, where I refer to it with the general term Activin.

NODAL is essential for chordate development, as established by genetic studies performed in frog, zebrafish, chicken and mouse models, and its conservation has been proved in basic organisms at the root of animal phylogeny such as Placozoans and sponges (Huminiecki et al., 2009, Shen, 2007). In adult tissues, NODAL is not normally produced, with the exception of the cycling endometrium and the lactating mammary gland (Quail et al., 2013, Strizzi et al., 2008, Papageorgiou et al., 2009). However, NODAL and its co-receptor CRIPTO, but not the antagonist LEFTY1, are overexpressed in many types of cancer, such as melanoma, prostate, breast and testicular tumours. In this context, reactivation of NODAL signalling has been suggested as a key step in favouring both the tumorigenic and metastatic processes (Topczewska et al., 2006, Wakefield and Hill, 2013). Growing evidence indicates that Activin may also play a crucial role in cancer.

Indeed, Activin has been proved to promote tumour progression in melanoma, skin and pancreatic cancer (Donovan et al., 2017, Lonardo et al., 2011, Togashi et al., 2015, Antsiferova et al., 2011).

Here, I will first summarize the roles of NODAL over the course of distinct developmental stages, by taking the mouse model system as a paradigm. Then, I will describe how NODAL/Activin signalling functions in ESCs at a molecular level to both control maintenance of pluripotency and cell differentiation toward different fates.

1.3.1 Functions of NODAL/Activin signalling in embryonic development

The process of embryogenesis, from the blastocyst stage to gastrulation and tissue patterning, crucially relies on NODAL/Activin signalling. In the early stages of development, NODAL is initially expressed throughout the epiblast - the symmetrical, cup-shaped cell layer of the blastocyst -where it ensures that cells maintain the full pluripotency potential necessary to gives rise to all foetal tissues (Brennan et al., 2001, Norris and Robertson, 1999). Shortly after implantation, NODAL is also a key player in setting up the embryonic axes (Figure 1.12), which is the starting point for all the subsequent steps of germ layers specification and organogenesis (Lu et al., 2001, Beddington and Robertson, 1999). Polarity in the epiblast is first established along the proximal-distal (P-D) axis by a gradient of NODAL signalling, which results from the production of the NODAL extracellular antagonists LEFTY1 and CER1 by the adjacent distal visceral endoderm (DVE) (Arnold and Robertson, 2009). In the DVE, SMAD2 also induces the expression of the Wnt antagonist DKK1, leading to the formation of a gradient of Wnt-β catenin signalling, which is required for maintenance of the P-D axis as well. Driven by NODAL signalling, the DVE cells then rapidly migrate to the anterior side of the embryo to form the anterior VE (AVE), thus generating gradients of NODAL and Wnt along the anterior-posterior (A-P) axis which eventually disrupt the radial symmetry of the epiblast (Arnold and Robertson, 2009).

At the onset of gastrulation, the high levels of NODAL signalling established at the proximal posterior side of the embryo initiate the formation of a structure called the primitive streak (Figure 1.12). The epiblast cells migrating through the primitive streak are subsequently specified into the mesoderm and endoderm germ layers by combined gradients of NODAL, BMP and Wnt signalling, with high NODAL levels

inducing endoderm and lower level inducing mesoderm (Sui et al., 2013). The third germ layer, ectoderm, arises at the end of gastrulation from the epiblast cells which fail to enter the primitive streak, and require active inhibition of NODAL signalling (Wu and Hill, 2009). Finally, at later stages of embryogenesis, NODAL controls the formation of the left-right axis, which is crucial for the asymmetrical positioning of the internal organs in the body (Shen, 2007). As anticipated, NODAL functions are highly conserved in vertebrate development, despite the fact that the processes of axis specification and tissue patterning can vary across different species (Wu and Hill, 2009). Indeed, as in mice also in *Xenopus* and zebrafish embryos localised sources of NODAL signalling control the specification of mesoderm and endoderm and, later in development, establish the left and right axis (Schier, 2009)

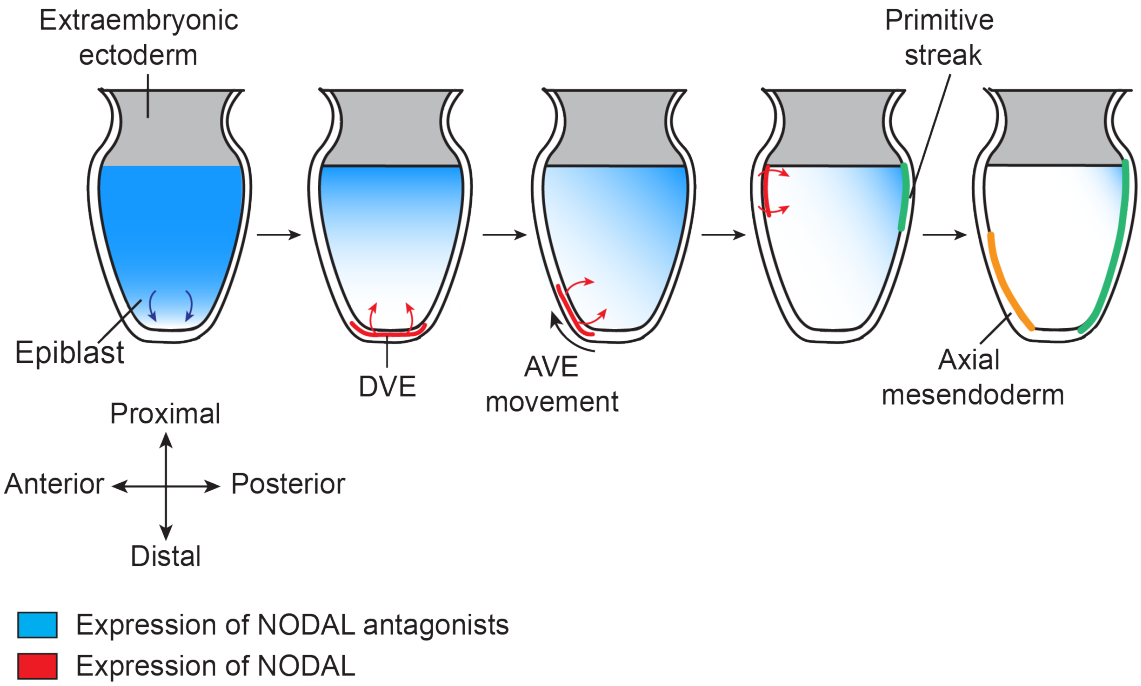


Figure 1.12. NODAL/Activin signalling in early embryonic development.
Schematic of NODAL signalling during mouse embryogenesis. Expression of NODAL in the epiblast is necessary to maintain pluripotency and to specify the formation of the proximal-distal and anterior-posterior axes along a gradient of NODAL signalling. Cells from the distal visceral endoderm (DVE), and then from anterior visceral endoderm (AVE), secrete NODAL antagonists and contribute to establishment of a NODAL gradient. NODAL is also required during gastrulation for the formation of the primitive streak and mesendoderm specification. See main text for details. Adapted from (Shen, 2007).

1.3.2 NODAL/Activin signalling in embryonic stem cells

In order to mechanistically understand how NODAL/Activin signalling acts *in vivo* to regulate the complex developmental processes just described, it is important to consider its function *in vitro* in the context of ESCs. Once established, ESCs are able to self-renew, while retaining their ability to form all three germ layers. Hence, they have been used as a model to investigate the molecular mechanisms of early embryogenesis (Itoh et al., 2014). Maintenance of pluripotency in ESCs relies on the core transcriptional circuit composed of OCT4, SOX2 and NANOG (referred to as OSN). These TFs cooperate to regulate the expression of many genes responsible for self-renewal, to retain their own expression in a transcriptional feedforward loop and to repress differentiation by keeping the developmental genes in a poised state (Young, 2011, Boyer et al., 2005). ESCs can be derived from the inner cell mass (ICM) of the blastocyst at an early embryonic stage of various species, even if most of the studies so far have been performed in mouse (m) or human (h) ESCs. Importantly, mESCs and hESCs are transcriptionally and epigenetically distinct, require unique culture conditions and do not show the same responses to external signals (Schnerch et al., 2010). It is now emerging the concept that these differences might reflect distinct stages of development, rather than being due to species-specific characteristics. Accordingly, hESCs are considered to be more 'primed' for lineage commitment with respect to the mESC, which would represent the naïve/ground state of pluripotency (Rossant, 2015, Ying and Smith, 2017). Indeed, an additional type of ESCs recently obtained at later stage in mouse development compared to mESCs has been shown to have characteristics more similar to hESCs. These cells are derived from the epiblast of post-implantation mouse embryos and are termed EpiSCs (Brons et al., 2007, Tesar et al., 2007).

In hESC and in EpiSC (Figure 1.13), maintenance of the pluripotency state depends on NODAL/Activin acting synergistically with the FGF/ERK signalling pathway (Vallier et al., 2005, Brons et al., 2007). In this context, activated SMAD2/3 drive the expression of *NANOG* and *POU5F1* (the OCT4 coding gene), directly interact with NANOG and OCT4 to regulate a plethora of self-renewal genes, block BMP-activated differentiation and prevent neuroectoderm specification (Sakaki-Yumoto et al., 2013, Vallier et al., 2009, Vallier et al., 2004). Interestingly, the functional antagonism between BMP and NODAL/Activin signalling is common

throughout embryonic development, and its regulation from a mechanistic point of view (that is, at the molecular level) it is not completely clear. One possibility is that SMAD2/3 competes with SMAD1/5 for binding to SMAD4 and titrates it out, decreasing BMP target gene expression (Avery et al., 2010). Alternatively, competition could occur at the level of the receptors, since NODAL and BMP share the type II receptors ACTR2A and ACTR2B (Massague, 2012). Moreover, NODAL has been shown to inhibit BMP signalling by inducing SMAD7, which would in turn feeds back to negatively regulate the BMP-SMAD1/5 pathway (Galvin et al., 2010). Finally, recent data from the lab suggests that SMAD2/3 may activate a yet unidentified phosphatase which specifically dephosphorylate SMAD1/5 (Ramachandran et al, manuscript in preparation).

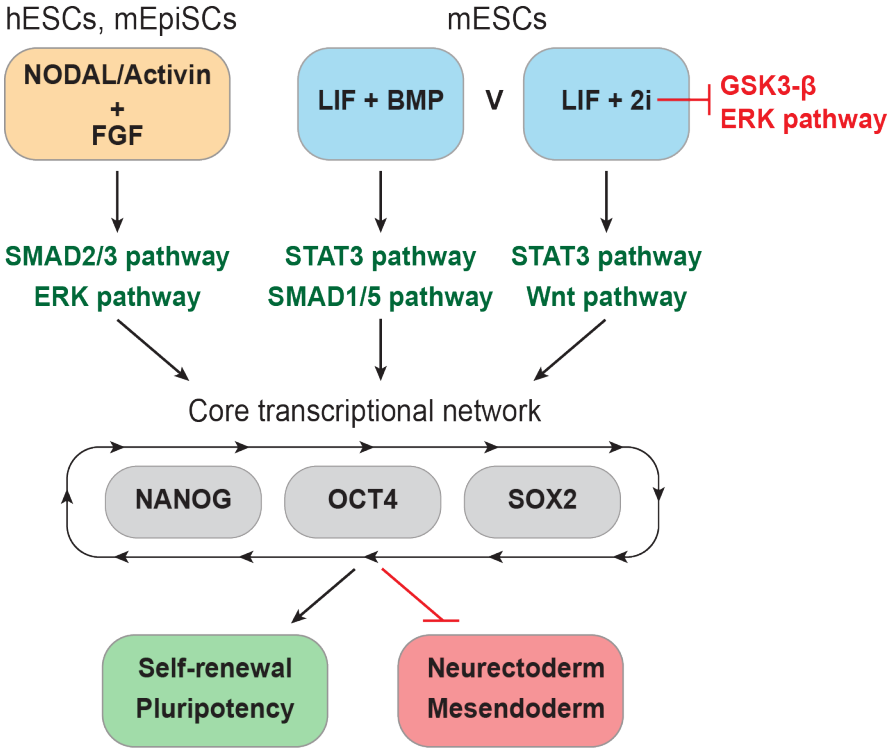


Figure 1.13. Signalling pathways regulating self-renewal of human and mouse embryonic stem cell and mouse epiblast stem cell.
Schematic of the different signalling requirements necessary to maintain the transcriptional network which controls the pluripotency state of human (h) and mouse (m) embryonic stem cells. mEpiSCs are a type of ESCs derived from the mouse epiblast with characteristic similar to hESCs. In hESCs and mEpiSCs, NODAL/Activin and FGF are required to co-regulate the core transcriptional circuit via the SMAD2/3 and ERK pathways. NANOG, OCT4 and SOX2 coordinate self-renewal and block the differentiation into neurectoderm and mesendoderm. In mESCs, pluripotency is maintained by LIF and BMP signalling through SMAD1/5 and STAT3 or by blocking the

Compared to hESCs and EpiSCs, the importance of NODAL/Activin signalling for mESCs self-renewal is less clear (Figure 1.13). Generally, maintenance of pluripotency in mESC seems to rely on fundamentally different mechanisms (Pauklin and Vallier, 2015). In the classical view, it is the result of a balance between signalling pathways suppressing lineage commitment, with the the leukemia inhibitor factor (LIF) blocking mesendoderm differentiation via the STAT3 pathway and BMP repressing the neuronal fate by SMAD1/5-mediated induction of the *Id* genes (Ying et al., 2003). Alternatively, the pluripotency properties of mESCs can also be preserved by chemically inhibiting GSK3β and the ERK kinase pathway, in the presence of LIF (culture method often referred to as 2i + LIF), and without the need for exogenous growth factors (Ying et al., 2008). Currently, there is some evidence suggesting a possible role for NODAL/Activin in self-renewing mESCs grown in the absence of serum. In this context, endogenous NODAL/Activin signalling promotes SMAD2/3 binding to the *Pou5f1* locus and it is necessary to block the differentiation of cells towards trophoectoderm (Lee et al., 2011). Nevertheless, further studies are needed to better define the relevance of NODAL/Activin signalling for the maintenance of pluripotency in mESCs.

mESCs, however, crucially rely on a gradient of NODAL/Activin signalling to differentiate along endodermal, mesodermal and neurectodermal lineages (Figure 1.14), as this is also the case for hESCs (Gaarenstroom and Hill, 2014). Acting in concert with Wnt and BMP pathways through a complex signalling network crosstalk (see Section 1.4.2), SMAD2/3 orchestrate ESC specification into mesendoderm, by controlling the transcriptional activity of genes such as *Gsc*, *Brachyury*, *Eomes* and *Mixl1* (Singh et al., 2012, Sakaki-Yumoto et al., 2013, Estaras et al., 2015, Dahle et al., 2010, Brown et al., 2011, Beyer et al., 2013). In this context, high levels of NODAL/Activin signalling further lead to definitive endoderm, whilst subsequent mesoderm derivation requires low levels of NODAL/Activin (Sui et al., 2013). Finally, inhibition of SMAD2/3 activation in the presence of FGF signalling promotes specification of ESCs into neuroectoderm (Itoh et al., 2014).

Figure 1.13 continued.
ERK pathway and activating the Wnt pathway via GSK3β inhibition in the presence of LIF.

In conclusion, NODAL/Activin signalling is equally required for either maintaining the pluripotency state and for inducing terminal differentiation, both *in vivo* and *in vitro*; but the mechanisms behind it remain to be fully elucidated. This paradoxical statement well captures the complexity of a signalling pathway able to control two opposite programmes of gene expression at the same time, and it suggests that in order to fully understand how TGF-β superfamily members elicit their biological functions it is eventually necessary to decipher how SMADs regulate transcription.

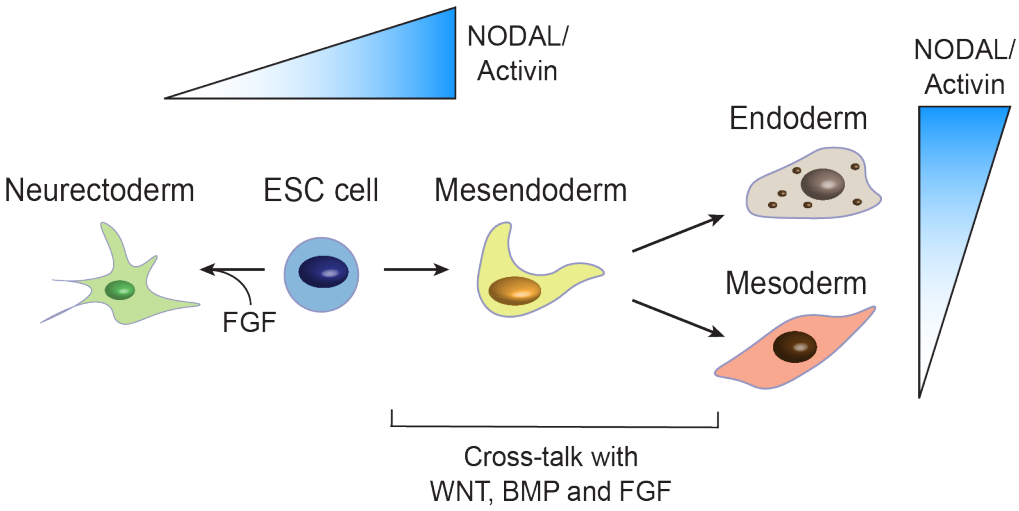


Figure 1.14. Roles of NODAL/Activin signalling in embryonic stem cell differentiation
Graded NODAL/activin signalling determines the differentiation of embryonic stem cells (ESCs) in cooperation with the Wnt, BMP and FGF pathways. In the absence of NODAL/Activin, FGF signalling drives ESC differentiation towards neurectoderm, whilst mesendoderm is induced upon high NODAL/Activin signalling. Endoderm and mesoderm are further specified along a gradient of NODAL/Activin signalling, with endoderm differentiation requiring higher levels of signalling compared to mesoderm.

1.4 How SMADs regulate transcription: the known and the unknown

Unlike the majority of transcription factors, the activated SMADs require the interaction with chromatin to recruit the general transcription machinery to target genes and induce their expression. This pioneer observation comes from early *in vitro* studies showing that, in an artificial transcription system, SMAD complexes failed to activate transcription on naked DNA, but did so efficiently in the presence of the chromatin template (Ross et al., 2006). Thus, SMAD transcriptional activity does not only rely on the interaction with other factors as anticipated earlier, but also with chromatin-modifying enzymes and with the epigenetic landscape. It is now becoming clear that the key to understanding how TGF-β signalling induces different programmes of gene expression in a cell and context-dependent manner is hidden behind the specificity of these interactions and their mutual interplay. In recent years, with the advent of Next Generation Sequencing (NGS) technologies many efforts have been made to identify SMAD cofactors, SMAD binding sites and their epigenetic features in different cell types and conditions (Trompouki et al., 2011, Morikawa et al., 2011, Mizutani et al., 2011, Fei et al., 2010, Mullen et al., 2011, Brown et al., 2011).

Here, I will briefly review the pivotal results achieved in the field, describing the different mechanisms regulating SMAD transcriptional activity identified so far and highlighting the outstanding questions which still require to be addressed (Figure 1.15).

1.4.1 Interactions with cell-type specific transcription factors

Intuitively, given the SMADs' limited DNA sequence specificity and affinity, it is reasonable to expect that the availability of partner TFs represents a major determinant in shaping the transcriptional responses to TGF-β ligands. Indeed, when comparing genome-wide SMAD binding across multiple cell types, SMAD complexes were found to co-occupy the genome at distinct target sites together with the lineage-specific TFs responsible for the different cell identities (Figure 1.15A), also known as master TFs (Chan and Kyba, 2013).

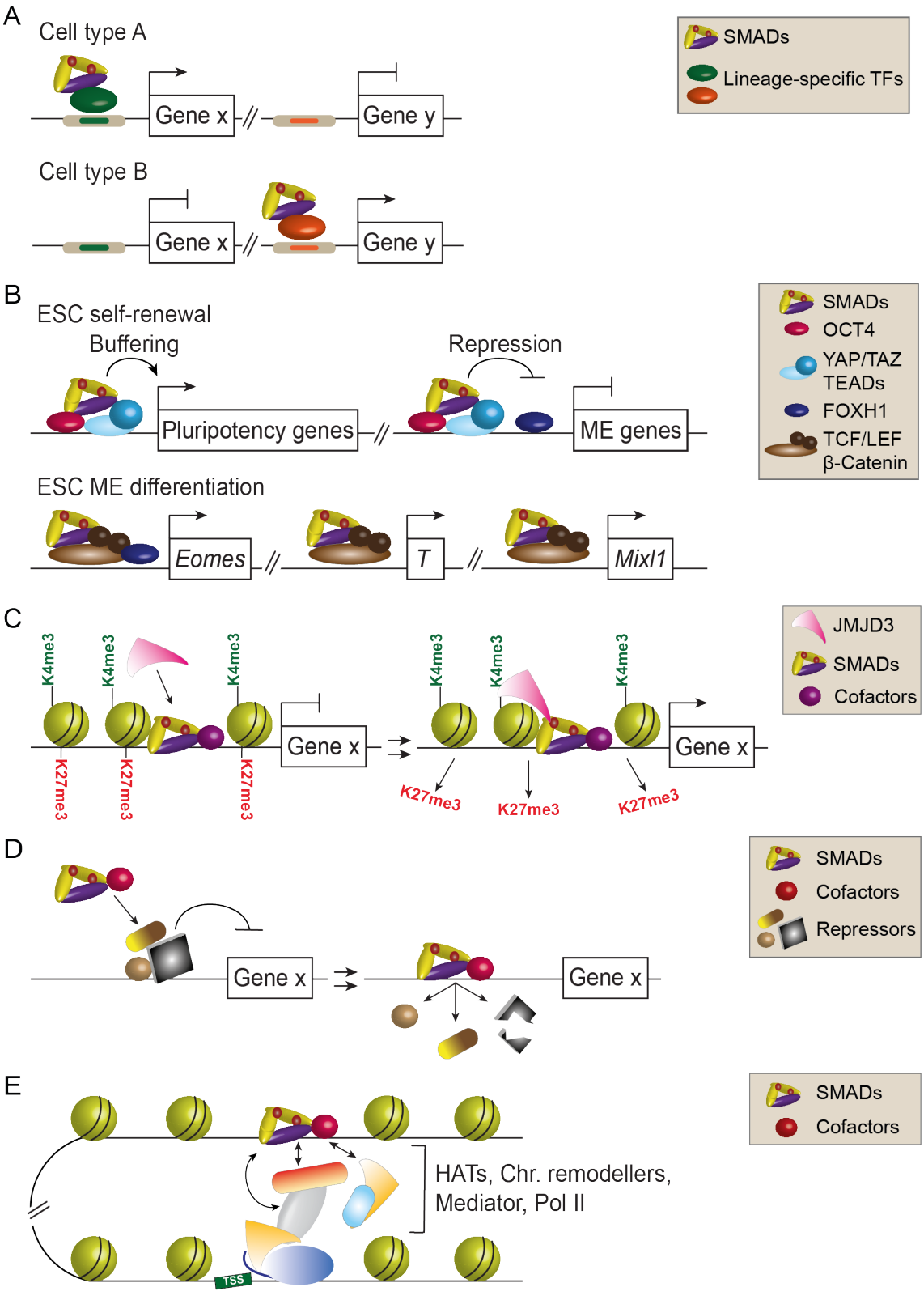


Figure 1.15. Mechanisms of SMAD regulated transcription.
(A-E) Different modes of regulating SMADs binding and transcriptional activity are depicted. **A)** Lineage-specific transcription factors (TFs) guide SMAD binding to target genes in a cell-specific manner. *Figure 1.15 continued on next page*.

For example, in erythroid and myeloid precursors BMP-activated SMAD1 co-localises with the respective lineage regulators GATA1 and C/EBPa, and the artificial expression of C/EBPa in erythroid cells is able to shift SMAD1 binding to sites newly occupied by C/EBPa (Trompouki et al., 2011). Similarly, in ESCs, myotubes and pro-B cells, SMAD3 co-occupies the genome at highly cell-type specific loci with the correspondent master TFs OCT4, MyoD and PU.1, and the formation of complexes with these TFs is necessary for the transcriptional regulation of target genes (Mullen et al., 2011). In these studies, the hierarchy of binding between SMADs and various cofactors was never fully addressed. Nevertheless, the view that emerged in the field was that the SMAD complexes are guided to cell type-specific sites by pre-bound master TFs, where they solely function as generic transcriptional regulators necessary for EP300 recruitment, without playing any instructive role in determining the different cell identities. However, the assumption that the SMADs act as pure modulators of the expression of master TFs target genes seems at odds with the clear driving functions of TGF-β ligands in many biological processes, as outlined in the previous section. It is also in sharp contradiction with what was observed by others in undifferentiated hESCs. Here the master TF FOXH1, which is necessary to mediate SMAD2/3-dependent induction of mesendoderm, is already bound to a subset of its target sites, yet it is not sufficient to recruit the SMAD2/3 complexes and to induce the expression of developmental genes (Kim et al., 2011). Finally, the actual biological relevance of the SMAD–master-TF cooperation in the different cell types has not been fully elucidated. For instance, TGF-β signalling seems to inhibit muscle differentiation by interfering with MyoD target gene activation, and the function of MyoD/SMAD3 complexes in myotubes has not been addressed (Liu et al., 2001).

Figure 1.15 continued.
B) SMAD cross-talk with other signal responsive TFs determine differential gene expression in response to changes in the extracellular environment (eg. Interaction with Hyppo and Wnt signalling pathways). **C)** Activated SMAD complexes can cooperate with components of the epigenetic machinery to modify the state of the chromatin at target sites. In the example, SMADs recruit the histone demethylase JMJD3 to poised developmental genes during ESC differentiation. JMJD3 removes the H3K27me3 repressive mark allowing gene expression. **D)** SMADs can activate transcription by replacing inhibitors bound at its target sites in the absence of signalling or **E)** by directly interacting with the general transcriptional apparatus. TSS, transcription start site.

1.4.2 Crosstalk with signal-responsive transcription factors

As exemplified by NODAL/Activin signalling, SMAD transcriptional activity varies not just across cell types, but also across conditions. During development, regulation of key target genes is achieved in collaboration with cell-specific TFs, such as NANOG, FOXH1 and EOMES, alongside with TFs responsive to other signalling pathways (Attisano and Wrana, 2013). In this context, SMAD binding sites function as 'hubs' that integrate information from the cellular and extracellular environments to ensure that different transcriptional programmes are elicited at the right time (Figure 1.15B).

Examples of such mechanisms include the interactions with the transcriptional effectors of the Wnt and Hippo pathway in ESC, which ultimately control the balance between self-renewal and differentiation (Attisano and Wrana, 2013). In hESCs, for instance, SMAD2/3 signalling converges with the components of the Hippo pathway YAP, TAZ and the TEADs alongside the pluripotency factor OCT4 to form a macromolecular complex referred to as TSO (Beyer et al., 2013). Under pluripotency conditions, the TSO provides a buffering capacity on pluripotency genes such as *NANOG* and represses the expression of mesendoderm markers (eg. *EOMES*), binding the latter in close proximity to FOXH1 occupied sites (Beyer et al., 2013). With mesemdoderm induction, SMAD2/3 complexes switch partners and enhancers, co-localizing with Wnt-activated β-catenin to induce developmental genes such as *EOMES*, *T/BRACHYURY* and *MIXL1* (Estaras et al., 2015, Beyer et al., 2013). The mechanism underlying TSO disruption upon differentiation stimuli, however, has not been addressed. Moreover, according to a recent work, the switch enhancer associated to *Eomes* is dispensable for its correct expression in early mouse embryos, which in turn requires a different SMAD-bound regulatory element (Simon et al., 2017). Thus, despite providing an attractive explanation to how one pathway can result in multiple outputs, the 'switch enhancer' hypothesis still begs a number of questions which need to be further investigated both *in vitro* and *in vivo*.

1.4.3 Regulation of chromatin modifications and control of the epigenetic machinery

In pluripotency conditions, key developmental genes are thought to be maintained in a repressed state, yet 'poised' for rapid transcriptional activation by the simultaneous presence of the active and repressive histone marks H3K4me3 and H3K27me3

(Vastenhouw and Schier, 2012, Bernstein et al., 2006). According to this hypothesis, timely gene expression upon differentiation signals would rely not only on the convergence between SMADs and other factors, but also on the existence of these bivalent domains. Indeed, growing evidences suggest that NODAL/Activin signalling is able to orchestrate ESC fate decisions by directly coordinating the molecular machineries which define the epigenetic landscape of the cells (Figure 1.15C). In self-renewing hESCs, for example, SMAD2/3 acting in concert with NANOG controls general H3K4me3 levels by recruiting the DPY30-COMPASS to both pluripotency and mesendoderm genes. Inhibition of NODAL/Activin signalling results in a decrease of H3K4me3 on these loci, which correlates with an impaired capacity of hESCs to differentiate towards mesendoderm (Bertero et al., 2015). Importantly, SMAD complexes are crucial not just to maintain the bivalent domains in pluripotency, but also to drive their resolution during differentiation processes.

For instance, during both hESCs and mESCs differentiation into endoderm, SMAD2/3 complexes transiently recruit the H3K27 demethylase JMJD3 to the promoters of mesendoderm markers such as *NODAL*, *BRACHYURY, GSC* and *EOMES*, and the loss of H3K27 methylation induced by NODAL/Activin at these genes is required for their transcriptional activation (Kim et al., 2011, Dahle et al., 2010). Interactions between JMJD3 and SMAD complexes have also been observed in different differentiation contexts in response to other TGF-β ligands. During neuronal stem cell differentiation, for instance, JMJD3 and TGF-β-activated SMAD3 co-bind to a subset of TGF-β-target genes to regulate their expression, whilst BMP-regulated SMADs require JMJD3 for the induction of Noggin during neuronal tube development. (Akizu et al., 2010, Estaras et al., 2012). Importantly, in all of these cases the timing of JMJD3 recruitment and the removal of the repressive mark has not been fully elucidated, nor has it been investigated what the involvement is of the other H3K27 demethylases such as UTX. Furthermore, the mechanisms behind the recruitment of the epigenetic machinery to only a subset and not all SMAD target genes are still unclear (Akizu et al., 2010, Kim et al., 2011, Estaras et al., 2012, Dahle et al., 2010).

1.4.4 Removal and recruitment of transcriptional repressors

In addition, to coordinate the deposition/removal of histone marks, the SMADs directly regulate the activity of transcriptional repressors (Figure 1.15D). Examples of such a mechanism can be found again in the context of ESCs. In hESCs under pluripotency conditions, SMAD2/3 together with OCT4 and NANOG inhibit the expression of the zinc finger protein Zeb-2/SIP1, whilst SOX2 promotes it, resulting in an intermediate level of transcription. Zeb-2/SIP1 then binds to mesendoderm genes such as *BRACHYURY, MIXL1* and *EOMES* keeping them inactive through the recruitment of the CtBP corepressor, and by antagonizing SMAD activity via direct interacting with the MH2 domain (Postigo, 2003, Chng et al., 2010). If NODAL/Activin signalling is inhibited, SOX2 fully activates the transcription of SIP1, which in turn blocks the expression of SMAD2/3 targets including pluripotency and developmental genes, hence favouring cell differentiation toward the neuroectoderm lineage. In contrast, during mesendoderm induction, Zeb-2/SIP1 expression is downregulated due to high levels of NODAL/Activin, and the mesendoderm genes are released from Zeb-2/SIP1 mediated repression (Chng et al., 2010).

Similarly, in the absence of TGF-β/NODAL/Activin the protein SKIL and its close family member SKI has been shown to repress SMAD target genes by binding to SBEs with SMAD4 and recruiting corepressors such as N-CoR ans/or mSin3A (Luo, 2004). In hESCs, SKIL is highly expressed and it contributes to maintain the pluripotency state through selectively repressing the mesendoderm genes (Tsuneyoshi et al., 2012). In the presence of high TGF-β/NODAL/Activin signalling, SKIL and SKI get rapidly degraded in a process involving the E3 ubiquitin ligase ARKADIA, making the SBEs accessible to activated SMAD complexes, which in turn induce gene expression (Levy et al., 2007, Tsuneyoshi et al., 2012).

Another example of SMAD-mediated displacement of transcription inhibitors comes from the interaction with TRIM33/TIF1γ. In mESCs, this E3 ubiquitin ligase forms complexes with phosphorylated SMAD2/3, which bind via the PHD-Bromo cassette of TRIM33 to H3K9me3 and H3K18ac sites close to the SBEs of mesendoderm genes such as *Gsc* and *Mixl1* (Xi et al., 2011). The interaction between TRIM33-SMAD2/3 and H3K9me3 dislodges the chromatin-compacting factor HP1γ and triggers, via yet uncharacterised mechanisms, further chromatin remodelling at the SBEs. As a result, SMAD4–SMAD2/3 are able to access their

target sites and to induce cell differentiation by activating the expression of the de-repressed mesendoderm genes (Xi et al., 2011).

Finally, it is important to highlight that SMAD complexes are able not only to activate transcription by displacing inhibitors, but also to actively obstruct it via the recruitment of repressors to target sites. Examples of such a mechanism are the interactions observed between BMP or TGF-β-activated SMADs with the transcription factors SCHNURRI and ATF3, respectively (Yao et al., 2006, Kang et al., 2003).

1.4.5 Use of the basal transcription machinery

The last aspect to consider when trying to understand how SMADs regulate gene expression is represented by the engagement of the basal transcription machinery (Figure 1.15E). So far, the recruitment at SMAD target sites of the HATs EP300 and CBP has been extensively characterised and proved to be necessary for gene activation (Ross et al., 2006, Janknecht et al., 1998, Feng et al., 1998). Multiple studies suggest that the engagement of chromatin remodelling complexes is also required to facilitate SMAD transcriptional regulation. The first of these enzymes shown to be recruited to chromatin by SMADs was the SWI/SNF nucleosome remodelling complex, with an interaction being demonstrated between the TGF-β-activated SMAD2/3 and the ATP-ase subunit BRG1 (Xi et al., 2008, Ross et al., 2006). Interestingly, BRG1-mediated chromatin remodelling seems to be crucial for the expression of a subset of TGF-β targets, but as in the case of JMJD3, what dictates gene specificity has not been addressed (Xi et al., 2008). Nor is it known if nucleosome depletion is actively induced by TGF-β superfamily signalling, or if it is a prerequisite for the SMADs to find their binding site on chromatin.

Whether the SMADs directly regulate the activity of Pol II is also poorly understood. Studies performed in *Xenopous* embryos suggest that SMAD complexes might communicate with Pol II via interactions with ARC105, a component of the Mediator complex, as knockdown or overexpression of ARC105 exactly mirrors the effects obtained by blocking or inducing NODAL signalling (Kato et al., 2002). In a more recent work performed in hESC, SMAD2/3 has also been found to boast transcription elongation by recruiting the pause-release factor PTEFb at the TSSs of mesendoderm genes during differentiation (Estaras et al., 2015).

However, if Pol II pause-release is used by the SMADs to regulate gene expression also in other cell contexts rather than hESCs, and if different modes of Pol II regulation are employed at distinct group of target genes remains unclear.

1.5 Defining new paradigms of SMAD-regulated transcription

In conclusion, the genome-wide interrogation of SMAD binding sites, transcriptional signatures and distribution of histone modifications in different cell types has greatly informed our knowledge of the mechanisms underlying SMAD transcription regulation (Mullen et al., 2011, Kim et al., 2011, Beyer et al., 2013). However, all the studies so far either only focused on autocrine chronic signalling, or inductions with ligands were performed without starting from a signal-inhibited baseline, along a time scale of days, rather than minutes or hours. Thus, the exact sequence of events occurring on chromatin from pathway activation to induction of transcriptional programmes remains obscure. How SMAD complexes find their target sites on DNA, and whether their cofactors are pre-bound or bind simultaneously with the SMADs, it is not known. To what extent SMAD binding to enhancers and promoters, and TF in general, are cause or consequence of chromatin remodelling and epigenetic modifications is also unclear. Finally, by which mechanisms the TGF-β ligands regulate Pol II activity, and how they control gene expression over time is also poorly understood.

Here, by taking a genome-wide approach I define the sequence of events that occur in response to NODAL/Activin signalling from SMAD2 binding to transcriptional activation, and I unveil the underlying mechanisms. I have used the murine P19 embryonic teratoma cell line, which show an ESC-like transcriptional response to NODAL/Activin but, in contrast to ESCs, do not differentiate when treated with the NODAL/Activin type I receptor inhibitor, SB-431542 (Vallier et al., 2005). Thus, they represent an ideal model to analyse the responses to acute and chronic NODAL/Activin signalling from a signal-inhibited baseline. In these experimental conditions, I have performed ChIP-seq for two different phosphorylation states of Pol II and various histone modifications, and I also recently characterised chromatin accessibility by taking advantage of the assay for transposase-accessible chromatin using sequencing (ATAC-seq). The integration of the data with SMAD2 ChIP-seq and RNA-seq previously performed in the lab surprisingly revealed that there is no

single mechanism of SMAD2-mediated transcription, and that SMAD2 directly orchestrates a dynamic program of gene expression downstream of NODAL/Activin signalling. ATAC-seq footprint analyses will ultimately provide the tools to identify novel SMAD2 cofactors, and to dissect the transcription regulatory network controlled by NODAL/Activin signalling.

With this work, I define new paradigms for SMADs-dependent transcription and provide a framework to understand how cells execute programmes of gene expression in response to TGF-β superfamily signalling during development, in adult organisms and in pathological contexts.

Chapter 2. Materials & Methods

The Cell Services at the Francis Crick Institute provided the following general solutions: 10 x TAE, 5 M NaCl, 0.5 M EDTA, 1 M $MgCl_2$, 1 M Tris-HCl (pH 7.5 or 8), 1 x PBS, 20% SDS, LB and sterile water.

2.1 Molecular Biology

The general procedures described below from Section 2.1.1 to Section 2.1.5 were primarily used to obtained the plasmids employed during the derivation of the CRISPR/Cas9-modified P19 cell lines, and to characterise the mutations present in these cells.

2.1.1 Bacterial transformation and plasmid preparation

The chemically-competent *Escherichia Coli* strain TG1 was used for plasmid amplification. 50 µl of competent bacteria thawed on ice were mixed with either 100 ng of plasmid DNA or 10 µl of ligation reaction and incubated on ice for further 20 minutes. Cells were then heat shocked for 40 seconds at 42°C, placed on ice for 2 minutes and, upon adding 150 µl of LB, shaken for 60 minutes at 37°C. Finally, the whole mix was plated out on LB agar plates supplemented with the appropriate antibiotic (provided by the Cell services at the Francis Crick Institute) and incubated at 37°C overnight. All plasmids used in this study were Ampicillin resistant. The day after, single colonies were picked and inoculated in 5 ml (minipreps) or 200 ml (maxipreps) of LB supplied with Ampicillin at a final concentration of 50 µg/ml, and incubated at 37°C overnight. For minipreps, plasmid DNA was extracted by the Equipment Park at the Francis Crick Institute, whilst for maxipreps it was isolated using the GenElute HD Plasmid Maxiprep Kit (Sigma Aldrich; Sigma) according to manufacturer's instructions. DNA concentration was assayed using a NanoDrop 1000 spectrometer (Thermo Fisher Scientific) by measuring the absorbance at a wavelength of 260 nm and considering an extinction coefficient of 0.02 $(ng/\mu l)^{-1}x\ cm^{-1}$. The plasmids used in this study were the pSpCas9(BB)-2A-GFP (PX458) and the pGEM-T Easy Vector Systems (Promega).

2.1.2 Cloning of CRISPR/Cas9 constructs

To clone the DNA oligos corresponding to the single guide RNA sequences listed in Section 2.5.1 into the pSpCas9(BB)-2A-GFP (PX458) plasmid (Ran et al., 2013) , the protocol from the Zhang lab available on-line at www.addgene.org was followed. The sgRNA sequences specifically targeting the genomic regions of interest were identified using the CRISPR Design Tool available at http://crispr.mit.edu. Briefly, 10 µg of plasmid backbone were digested for 1 hr at 37°C with 10 units of BbsI (NEB) in a total volume of 50 µl, containing 1x of the NEBuffer 2.1, and purified using the PCR purification kit from Qiagen according to manufacturer's instructions. For each pair of oligos, 1 µl of the single phosphorylated oligo (100 µM) was mixed with 1x T4 Ligase Buffer (NEB) in a total volume of 10 µl, and annealed in thermocycler set at 37°C for 30 minutes, followed by 95°C for 5 minutes and by a ramp down to room temperature (RT). 1 µl of a 1:200 dilution of the annealed oligo duplex was finally ligated with 50 ng of digested plasmid using 1 µl of T4 DNA Ligase (NEB) in 1x T4 Ligase Buffer (NEB), and by incubating the mix overnight at 16°C. The ligation product was then transformed as described in Section 2.1.1., colonies were picked for minipreps and the presence of the correct guide sequences was verified as described below.

2.1.3 Sequencing

For sequencing reactions, 200 ng of DNA template were mixed with 10 ng primer and 8 µl BigDye Terminator v2.0/v3.1 (Applied Biosystems), in a final volume of 20 µl. The sequencing reaction was carried out in GeneAmp® PRC System 9700 (Applied Biosystems) using the following cycling conditions:

96°C 10 seconds
50°C 5 seconds x 30 cycles
60°C 4 minutes

60°C 1 minute
10°C ∞

The Dyex 2.0 Spin kit (Qiagen) was used to remove dye terminators from sequencing reactions according to manufacturer's instructions, and the DNA was air-dried using a vacuum centrifuge. Sanger sequencing was then performed by the Equipment Park at the Francis Crick Institute. To verify the sequence of fragments amplified from genomic DNA by PCR (see section below), the PCR products were subcloned into pGEM-T Easy Vector Systems (Promega) according to manufacturer's instructions, white colonies were picked, and minipreps along with Sanger sequencing were performed by the Equipment Park at the Francis Crick Institute. In all cases, sequences were visualised using the 4Peaks 1.8 software, and aligned to the reference sequence or to each other with the SeqMan Pro or the SerialCloner 2-6-1 software.

2.1.4 Genomic DNA extraction and PCR amplification

Genomic DNA was normally extracted from 90% confluent cells grown in a 12 well tissue culture plate. After removing the media, 300 µl of QuickExtract™ DNA Extraction Solution (Epicentre) were added to the well, and the resulting cells suspension was transferred to a fresh 1.5 Eppendorf tube. Finally, samples were incubated at 65°C for 6 minutes, and then at 96°C for 2 minutes. Specific DNA fragments were amplified from genomic DNA by PCR. 3 µl of the sample obtained as just described, or 50 ng of DNA if this was extracted with other methods, were added to the Qiagen Taq mix, and the PCR was carried-out using thermal cycler (GeneAmp® PRC System 9700 (Applied Biosystems) or Veriti 96-well Thermal cycler (Applied Biosystems)) with the conditions described below. The primers used in this study for genomic PCR are reported in Section 2.5.2.

Qiagen Taq reaction mix

2.5 µl 10x PCR Buffer
0.5 µl dNTPs 10 mM
2 µl forward primer 5 µM
2 µl reverse primer 5 µM
0.125 µl Taq polymerase
50 ng DNA or 3 µl of sample as describe above
up to 25 µl final volume with water

PCR thermal cycler programme

94°C 2 minutes

94°C 30 seconds
55°C* 30 seconds x 30 cycles
72°C 1 minutes

72°C 5 minute
4°C ∞

* The annealing temperature varied depending on the primers used, most of the primers however worked at 55°C.

2.1.5 Agarose gel electrophoresis

The agarose gels (1% or 2% agarose (Invitrogen)) were prepared in 1 x TAE with 0.5 µg/ml ethidium bromide, poured into a sealed gel tray with the appropriate gel comb and allowed to set at RT. Gels were run in 1 x TAE with DNA ladder (New England Biolabs; NEB) at 120 volts for 15-45 minutes. Visualisation and photography of gels was carried-out by ultraviolet illumination using a UVP 2UV Transilluminator.

2.1.6 RNA extraction

Total RNA was extracted using Trizol (Thermo Fisher Scientific), all centrifugation steps were carried-out using a microcentrifuge (Thermo Fisher Scientific). Cells were lysed by adding 500 µl per well from a 6-well plate and mixing thoroughly. After collecting the lysates into an Eppendorf tube, 100 µl of water saturated chloroform were added and the samples vortexed for 15 seconds. Phase separation resulted from centrifugation of the samples at 12,000g for 15 minutes at 4°C. The upper aqueous phase was transferred to a fresh Eppendorf tube and mixed with 250 µl of Isopropanol. To precipitate RNA, samples were first incubated at RT for 10 minutes and then centrifuged at 12,000g for 10 minutes at 4 °C. The supernatant was removed and the pellet washed with 500 µl of 75% ethanol. After vortexing, samples were spun at 7,500 g for 5 minutes at 4°C and the ethanol subsequently discarded.

Finally, the pellet was air-dried for 10 minutes and the RNA was redissolved by adding an appropriate volume of water (usually between 40 and 60 µl). RNA concentration was assayed as described for DNA in Section 2.1.1., but considering an extinction coefficient of 0.025 $(ng/\mu l)^{-1}x\ cm^{-1}$.

2.1.7 cDNA synthesis

cDNA was synthesized from RNA by preparing the reaction mix from the AffinityScript RT-PCR kit (Agilent Technologies) reported below. The synthesis reaction was carried out using a thermal cycler (GeneAmp® PRC System 9700 (Applied Biosystems) or Veriti 96-well Thermal cycler (Applied Biosystems)) and the conditions described below.

cDNA synthesis reaction mix

2 µl 10x AffinityScript RT Buffer
2 µl DTT 100 mM
0.8 µl dNTPs 25 mM
3 µl Random 9-mer primers
1 µl AffinityScript Reverse Transcriptase
0.5 µl RNAse inhibitor (Promega)
1 µg RNA
up to 20 µl final volume with water

cDNA synthesis thermal cycler programme

25°C 5 minutes
42°C 15 minutes
55°C 15 minutes
95°C 15 minutes
4°C ∞

The cDNA was diluted 1:10 in water and used for qRT-PCR.

2.1.8 Quantitative Real-Time PCR (qPCR)

For qPCR, the following mix was pipetted into MicroAmp Fast Optical 96-well reaction plates (Applied Biosystems):

qPCR reaction mix
2 µl cDNA
5 µl Express SYBR Green Master Mix (Invitrogen)
1 µl forward primer 3 µM
1 µl reverse primer 3 µM
1 µl water

The qRT-PCR was performed using a 7500 Fast real Time PCR Machine (Applied Biosystems) set on the FAST mode. The machine carries out 40 cycles and measures the accumulation of fluorescence emitted by the SYBER Green dye during each amplification step. The threshold cycle (C_t) value, the cycle number at which the fluorescent signal is significantly above the background fluorescence, is automatically computed by the 7500 Software (Applied Biosystems). For each sample, the expression level of the gene of interest (GOI) was calculated normalising its C_t value to the C_t value measured for *Gapdh* in the same sample, according to the ΔC_t method:

(1) Relative [cDNA GOI] $=2^{-(Ct(GOI)-Ct(Gapdh))}$

To analyse the relative changes in GOI expression between treatment (tr) and control (cntrl) samples, the $\Delta\Delta C_t$ method was employed instead:

(2) *Fold change [cDNA GOI] $=2^{-(\Delta Ct(tr)-\Delta Ct(cntrl))}$

*ΔCt(tr) and ΔCt(cntrl) are the values obtained using equation (1) for the treatment and control samples, respectively.

To assess primer specificity, a melt curve was also performed. At the end of the qPCR run, the thermal cycler measures the amount of fluorescence. The

temperature of the sample is then increased incrementally as the instrument continues to measure fluorescence, which decreases when the dsDNA becomes single stranded. The graph of the negative first derivative of the melting curve is used to pinpoint the temperature of dissociation, which depends on the DNA sequence of the fragment amplified. The presence of a single peak in this graph was considered a *bona fide* indicator of primer specificity, and primers showing multiple peaks were discarded. All qPCR primers were designed using primer-BLAST (Ye et al., 2012), led to a PCR product of 70 to 250 bp and span exon-exon junction. The last parameter was not applyed if the primers were employed to amplify genomic regions, as was the case for ChIP-PCR or FAIRE-PCR experiments (see Sections 2.3.3 and 2.3.4). The full list of primers used in this study for analysis of gene expression is provided in Section 2.5.3.

2.2 Cell culture

2.2.1 General culture conditions

The cell lines used in this study were obtained from the following sources: P19 cells, Grace Gill (Harvard Medical School); C2C12 cells, Richard Treisman (Francis Crick Institute); EpH4 cells, Harmut Beug (IMP, Vienna). All three cell lines were banked by the Francis Crick Institute Cell Services, were certified negative for mycoplasma and validated as of mouse origin. All three cell lines were cultured in Dulbecco's Modified Eagle Medium containing 10% FCS (Thermo Fisher Scientific) and 1% Penicillin/Streptomicin (PenStrep) solution (Thermo Fisher Scientific). Cells were grown on plastic dishes or in tissue culture grade flasks (Corning), incubated at 37°C, 10% CO_2 and passaged using Trypsin-EDTA (Sigma).

2.2.2 Ligand and drug treatments

Listed in the table below are the ligands and drugs used in this study, as well as the solvents and the final concentrations. NODAL/Activin signalling was blocked by overnight incubation with 10 μM SB-431542 (Inman and Hill, 2002). Prior to stimulation, SB-431542 was washed out three times with PBS and replaced with 20 ng/ml Activin in full media for different times. For the SB-431542 condition, after washout, cells were incubated for 1 hr in 10 μM SB-431542 in full media. This was in order to control for a transient effect of serum stimulation in the 1 hr Activin sample.

The untreated condition represents a chronic signalling state resulting from an autocrine production of NODAL and GDF3. For the SB-431542 chase experiment, cells were treated as just described in the control samples, or 10 μM SB-431542 were added to the cells after 1 hr of Activin treatment for different times. For the Cycloheximide experiment, prior to ligand induction cells were pre-treated with the drug for 5 minutes, which was added together with the ligand again after SB-431542 washout.

Drug/Ligand	Source	Concentration	Solvent
Activin A	Peprotech	20ng/ml	PBS + 0.1% BSA
SB-431542	Tocris	10 μM	DMSO
Cycloheximide	Sigma	5 μg/ml	Water
Actinomycin D	Sigma	6 μM	DMSO

2.2.3 siRNA transfection

siRNA transfections were performed using a reverse transfection protocol. Here, a mix in OPTIMEM (Life Technologies) of Lipofectamine RNAiMAX (Thermo Fisher Scientific) and siRNA (Dharmacon) was pre-incubated for 20 minutes at RT, and then plated together with the cells. The siRNA final concentration was 20 nM for all the experiments in this study, and it has to be considered relative to the final volume in the tissue culture dish, which is the sum of mix volume and the volume of full media plated. In all cases, cells were harvested 72 hr after transfection. If the transfection was carried out in a 6 well plate, 3 μl of Lipofectamine RNAiMAX and 2.2 μl of siRNA 20 μM were mixed in 200 μl of OPTIMEM and plated using 2 ml of full media containing 120,000 cells; otherwise, quantities were scaled-up accordingly to the size of the culture dishes used. The siRNAs used in this study are listed in Section 2.5.4.

2.2.4 Generation of cell lines

P19 cells stably expressing MYC-FOXH1 were previously generated in the lab via co-transfecting a plasmid encoding MYC-tagged FOXH1 (Labbe et al., 1998) with pSUPER-retro-puro (Gronroos et al., 2012)(Gronroos et al., 2012)(Gronroos et al., 2012)(Gronroos et al., 2012), to allow for selection of puromycin-resistant stable clones. Below, a detailed description of how the CRISPR/Cas9-modified P19 lines

were obtained. To avoid problems of heterogeneity in the P19 population, I first isolated a clone of P19 cells that had the same characteristics as the P19 pool with respect to gene expression profiles in response to Activin treatment (Figure 2.1). DNA oligos corresponding to the sgRNA sequences were cloned into pSpCas9(BB)-2A-GFP (PX458) as described in Section 2.1.2, and the plasmids were transfected into P19 cells following the protocol outlined here.

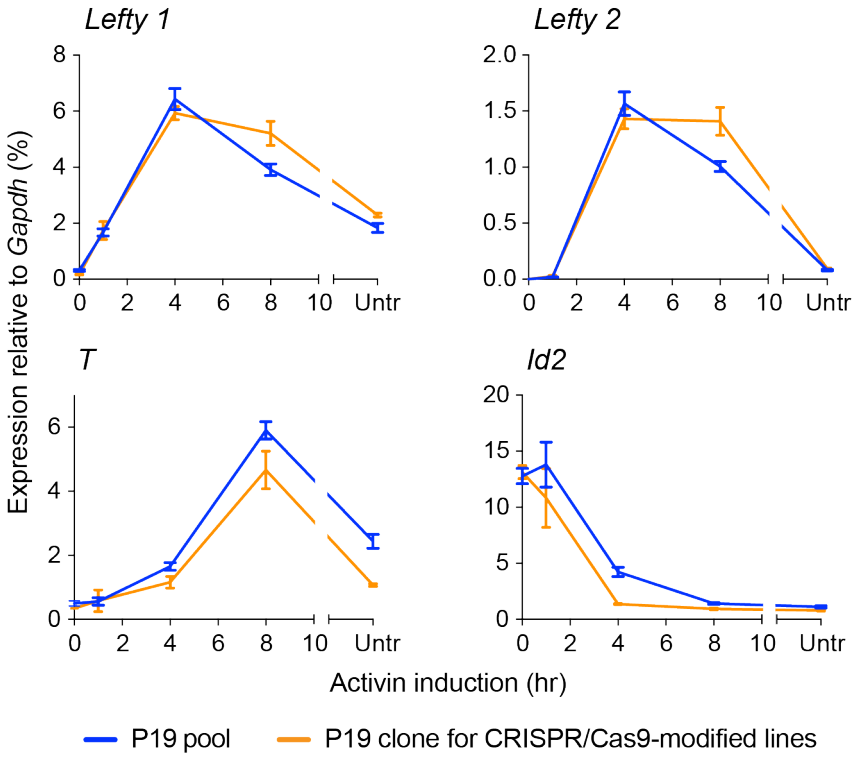


Figure 2.1. The clone of P19 cells used for generating CRISPR/Cas9-modified lines shows the same Activin-dependent transcriptional responses as the P19 pool. A time course of Activin induction was performed on the P19 pool and on the clone of P19 cells used for generating CRISPR/Cas9-modified lines. Cells were treated with 10 μM SB-431542 overnight, followed by washout and incubation with Activin for the indicated times, or SB-431542 was replaced for 1 hr. Shown are qPCR measurements of the expression of NODAL/Activin target genes for the samples described. A representative experiment (means ± SD) is shown. Untr, untreated.

2.2.4.1 Cell transfection

P19 cells ~50% confluent grown in 10 cm tissue culture dishes were transfected using Lipofectamine 2000 (Invitrogen) according to the manufacturer's instructions with minor adaptations. 30 μl of Lipofectamine 2000 were mixed with 750 μl of OPTIMEM and incubated at RT for 5 minutes. Meanwhile, 7.5 μg of each plasmid

were added to 750 µl of OPTIMEM. The two mixes were then combined together for a total volume of 1.5 ml, and left at RT for 20 minutes whilst the media in the dishes was replaced with 6.5 ml of OPTIMEM. The DNA-lipid complexes were finally added to cells, and cells were incubated for 8 hr before the OPTIMEM was removed and substituted with full media. 48 hr after transfection, cells were trypsinized and resuspended in OPTIMEM for single cell FACS sorting.

2.2.4.2 Single cell FACS sorting

Single cell FACS sorting was carried out on P19 cells transfected as just described, and was based on the detection of the GFP encoded by the CRISPR/Cas9 plasmids. Untransfected P19 were used to set the sorting parameters. The GFP-positive cells were sorted and seeded into 96 well plates using a FACSAriaIII cytometer with a 100 micron needle by the Flow Cytometry Facility at the Francis Crick Institute. Single cell clones were then grown and screened by PCR as described in Section 2.1.4.

2.3 Protein and chromatin analysis

2.3.1 Whole cell extracts, measurement of protein concentration and lysate processing

Upon washing the tissue culture plates twice with ice-cold PBS, cells were lysed on ice with D0.4 Buffer and scraped from the tissue culture dishes. Lysates were then sonicated for 10 seconds on ice and centrifuged at 13,000 rpm for 5 minutes at 4 °C to remove cell debris. The supernatant was transferred to a fresh Eppendorf tube and protein concentration was assayed using a SpectraMax Plus spectrophotometer (Molecular Devices). 1 µl of the sample was added to 200 µl of a 1:5 dilution of the Biorad Protein Assay (Biorad) in a 96 well plate, and the optical density (OD) was measured at 595 nm. Serial dilutions a 10 µg/µl BSA solution were also quantified in parallel to provide a standard curve. The protein concentration of the samples was then determined from the BSA standard curve values. Prior to loading on a SDS polyacrylamide gel, samples were diluted in 4X Laemmli Buffer and boiled for 5 minutes at 95 °C.

D0.4 Buffer

20 mM HEPES pH 7.5
10% v/v glycerol
0.4 M KCl
0.4% v/v Triton X-100
10 mM EGTA
5 mM EDTA
1X Protease Inhibitors (Roche), added fresh
25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

4X Laemmli Buffer

7.7% (w/v) SDS
0.2 M Tris-HCl (pH 6.8)
38% (v/v) glycerol
4% (v/v) 2-mercaptoethanol
0.025%(w/v) bromophenol blue

2.3.2 SDS polyacrylamide gel electrophoresis (SDS-PAGE) and Western Blotting

Normally, 40 µg of protein were separated by size using SDS-PAGE, alongside with in-house molecular weight markers (see below). 15% acrylamide gels were prepared using the recipe below, and run in 1x SDS Buffer using the electrophoresis apparatus (Cambridge Electrophoresis Ltd) for 1.5 hr at 230 Volts and 65 mAmps. Proteins were then transferred onto a PVDF 0.45 µm membrane (Millipore) using a semi-dry transfer system. The membrane was first activated in methanol and subsequently equilibrated in Transfer Buffer, together with the gel and 6 pieces of 3 mm Whatman Paper. Three pieces of Whatman Paper followed by the membrane, the gel and other three pieces of Whatman Paper were placed into a SemiPhor Transfer Unit (Hoefer), and the transfer was carried-out for 1 hr at 0.8 mA/cm^2 of gel. Ponceau S solution (Sigma) was then used to verify that proteins and markers had been successfully transferred to the membrane. To prevent non-specific antibody binding, the membrane was blocked for 1 hr using 5% (w/v) milk powder (Marvel) in PBST, or

using PBST with 5% of BSA (Sigma) and 0.08% of Sodium Azide if the primary antibody was phospho-specific. PBST consists of PBS with 0.1% (v/v) Tween-20. Primary antibodies were normally diluted 1:1000 into the respective blocking solutions and used to incubate the membrane on a rocking platform for 1 – 2 hr at RT, or overnight at 4°C. Following 5 washes with PBST, the HRP-conjugated secondary antibodies diluted 1:5000 in 5 % milk/PBST were added to the membrane for 1 hr, after that the washing step was repeated. Luminata Classico Western HRP Substrate (Millipore) or Immobilon Western HRP Substrate (Millipore) were used to detect HRP, and the luminescence was finally captured by exposing the membrane to an autoradiography film (SuperRX, Fujifilm). Blots were quantified using ImageJ. Densitometry analysis, and for any given sample the value obtained for the protein of interest were normalised to the value measured for the loading control in the same sample. The list of primary and secondary antibodies used in this study is reported in Section 2.5.5.

15% Resolving gel

5.63 ml 40% (w/v) acrylamide
0.65 ml 2% (w/v) bisacrylamide
3.75 ml 1.5 M Tris-HCl pH 8.8
Up to 15 ml with water
40 µl 20% (w/v) APS
30 µl TEMED

Stacking gel

0.63 ml 40% (w/v) acrylamide
0.33 ml 2% (w/v) bisacrylamide
0.625 ml 1 M Tris-HCl pH 6.8
Up to 5 ml with water
15 µl 20% (w/v) APS
10 µl TEMED

SDS Running Buffer

384 mM Glycine, 50 mM Tris, 0.1% (w/v) SDS

Molecular weight markers

Myosin (212 kDa), β-galactosidase (116 kDa), Phosphorylase (97.5 kDa), BSA (66.5 kDa), Catalase (57.8 kDa), Glutamate Dehydrogenase (55.5 kDa), Ovalbumin (42.7 kDa), GAPDH (35.7), Carbonic anhydrase (28.9 kDa), SBTI (20 kDa).

Transfer Buffer

150 mM Glycine
20 mM Tris
0.01% (w/v) SDS
20% (v/v) methanol

2.3.3 Generation of FOXH1 antibody

First, two peptides corresponding to amino acids 16–31 or amino acids 31–43 of mouse FOXH1 were synthetized and coupled to KLH by the Peptide Chemistry Facility the Francis Crick Institute. Two rabbits were immunized against each peptide by the Pettingill Technology Ltd company, following a protocol which involves a series of injections repeated every 10~15 days over the course of 1.5 months. For each round of immunization, samples of the animal sera were collected alongside pre-immunization sera to work as negative controls. A small aliquot from each serum was diluted 1:1000 in milk/PBST, and used to perform Western Blot analysis of MYC-FOXH1 P19 lysates. Only for the blots incubated with the sera from the two rabbits immunized against the 16–31 peptide, a band was detected at the same molecular weight at which the signal from MYC antibody was also observed. Since this specific band was better seen by the sera collected after the last round of immunization, only these samples were taken forward for further antibody purification. The terminal sera from the two rabbits immunized against the 31–43 peptide were also processed for antibody purification, altough they failed to recognize MYC-FOXH1 in Western Blot.

To purify the antibodies recognizing the two FOXH1 peptides from sera, columns of SulfoLink™ resin were coupled to the respective peptides by mean of their terminal cysteine residues, and the protocol provided by manufacturer (Thermo Fisher Scientific) was followed with minor adaptations. For each sample, 10 ml of serum were centrifuged at 3,000 rpm for 30 minutes at 4°C, and the supernatants were diluted 1:10 in 10 mM Tris HCl (pH 7.5). The resulting ~100 ml were allowed to

pass multiple times through the corresponding column by gravity-flow method, for a period of at least 12 hours and at 4°C. Columns were then washed four times with 40 ml of 10 mM Tris HCl (pH 7.5), and eluted four times with 1 ml of Glycine 0.2 M (pH 2,5). Immediately after, the eluates were brought at pH 7.5 using 1 M Tris HCl (pH 8.5). To improve the storage of the purified antibodies, a buffer exchange by dialysis was also performed. Dialysis tubing with a MWCO of 8,000 Da (Spectrum) was used, and samples were dialysed against 2 L of a solution consisting of 50% glycerol (Sigma) and 50% PBS. The presence of the purified antibodies in the dialysed samples was verified by running 20 µl alongside a positive control on a 15% acrylamide gel, followed by Coomassie staining carried out as described below. Finally, the antibodies were tested in Western Blot and ChIP experiments. The antibody recognizing the 31–43 peptide did not work for any of the two techniques, whilst the one against the 16–31 peptide was shown to recognize endogenous FOXH1 in ChIP and Myc-FOXH1 in Western Blot, as described in the result Section 5.2.9. Thus, the term in-house FOXH1 antibody refers to the antibody recognizing the amino acids 16–31 of mouse FOXH1.

2.3.3.1 Coomassie staining

The gel was incubated in staining solution for 30 minutes, then incubated in destaining solution for 1 hr with constant shaking and frequent changing of the solution. Finally, the gel was imaged using a standard computer scanner.

Coomassie stain

0.5% (w/v) Coomassie blue R250, 45% (v/v) Methanol, 10% (v/v) Acetic Acid

Destain

25% (v/v) methanol, 7% (v/v) Acetic acid

2.3.4 ChIP-PCR

P19 cells were plated on 15 cm tissue culture dishes in order to be around 45 ~ 50 x 10^6 the day when the cells were stimulated as appropriate in a volume of 20 ml and collected. For chromatin crosslinking, 550 µl of 36-37% formaldehyde (Sigma) were

added for a final concentration of 1% formaldehyde, and the dishes were incubated for 10 minutes at RT with gentle swirling. Unreacted formaldehyde was quenched by adding 1 ml of 2.5 M Glycine (final concentration 125 mM), followed by 5 minutes of incubation at RT with gentle swirling. Cells were then washed twice with cold PBS, scraped from the tissue culture dishes into 15 ml polystyrene Falcon tubes (Corning) and centrifuged at 1,000 rpm for 10 minutes at 4°C on a benchtop centrifuge (Belkman Coulter). To lyse the cells and isolate the nuclei, the pellets were thoroughly resuspended in 2.5 ml of the hypotonic Buffer A, which also contained 0.5% NP-40 (see below), and the suspension was incubated for 10 minutes on ice before being centrifuged at 2,000 rpm for 10 minutes at 4°C. Nuclei were then washed in 2.5 ml of Buffer A', and carefully resuspended in 1 ml of Buffer B, which is the sonication buffer.

Chromatin was sheared to 100-400 bp carrying out 10 30" on-off cycles of sonication at a constant temperature of 4°C. The sonication settings had been previously optimised and were kept constant across experiments. The Bioruptor Sonicator (Diagenode), 15 ml polystyrene Falcon tubes (Corning) and provided sonication probes were used for this step. After sonication, samples were moved to a 1.5 ml non-stick Eppendorf tube (Alpha Labs), and centrifuged at 13,000 rpm for 10 minutes at 4°C. Before performing the immunoprecipitation (IP), 20 µl of supernatant for each sample were set aside to use as input, and additional 20 µl were processed to check the quality of fragmentation by running the de-crosslinked, cleaned-up chromatin on a DNA agarose gel.

The amount of chromatin to use for the immunoprecipitation (IP) depended on the type of experiment, with the equivalent of 15 x 10^6 cells being required for transcription factor IPs, whilst histone modification IPs or Pol II IPs were performed respectively with 2 and 5 x 10^6 cells. In preparation of the IP step the sonicated material was first diluted with the ChIP Dilution Buffer between 5 and 10 fold (depending on the volumes used), in order to reduce the concentration of SDS, which impedes antibody efficiency. The diluted chromatin was then pre-cleared for 2 hr at 4°C on a rotating wheel with 2.5 µl of magnetic protein A/G Dynabeads (Invitrogen) per 10^6 cells. To avoid non-specific binding of proteins, beads were pre-blocked by washing them three times with cold PBS + 0.1% BSA. Following pre-clearing, each IP was performed at 4°C on a rotating wheel overnight in 2 ml non-stick Eppendorf tubes, using 0.75 – 1 µg of antibody per 10^6 cells. The day after, 6.6 µl of pre-blocked

protein A/G Dynabeads per µg of antibody used in the IP were added in the tubes and the samples were incubated at 4°C on a rotating wheel for at least 6 hr. Beads were then washed three times in ChIP Low Salt Wash Buffer, three times in CHIP High salt Wash Buffer and once in LiCl Wash Buffer, followed by the last wash in TE Buffer. Chromatin was finally eluted off the beads by resuspending them in 200 µl of Elution Buffer. In addition, 180 µl of the same buffer were added to the input samples previously collected. De-crosslinking and proteins digestion was carried out at 65°C overnight after adding 8 µl of 5 M NaCl and 5 µl of Proteinase K (Sigma) from a 10 mg/ml stock. Samples were subsequently cleaned up using the PCR purification kit (Qiagen) according to manufacturer's instructions, eluting DNA off the column in 60 µl of water. An additional 100 µl of water were added to the inputs, and a fraction was taken from each of them to make a standard pool.

From the standard pool, four 1:3 dilutions were further prepared in order to obtain a range, which was used to quantify the amount of target DNA present in the immunoprecipitated samples by qPCR. The absolute input values, which corresponds to the total amount of target DNA in the samples prior to the IP, were also computed from the input samples. Thus, the percentage of target DNA retrieved with each immunoprecipitation compared to its starting quantity was obtained dividing the ChIP value by the absolute input value. The qPCRs were performed as described in Section 2.1.8, and the primers employed are listed in the Section 2.5.6. For a list of the antibodies used for ChIP experiments in this study, see Section 2.5.5.

ChIP Buffer A

5 mM HEPES, pH 8.0

85 mM KCl

0.5% NP-40

1X Protease Inhibitors (Roche), added fresh

25 mM NaF, added fresh

25 mM β-Glycerophosphate, added fresh

ChIP Buffer A'

5 mM HEPES, pH 8.0

85 mM KCl

1X Protease Inhibitors (Roche), added fresh

25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

ChIP Buffer B
50 mM Tris-Cl, pH 8.0
1% SDS
10 mM EDTA
1X Protease Inhibitors (Roche), added fresh
25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

ChIP Dilution/IP Buffer
50 mM HEPES, pH 7.5
150 mM NaCl
1% NP-40
1X Protease Inhibitors (Roche), added fresh
25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

ChIP Low Salt Wash Buffer
50 mM HEPES, pH 7.5
150 mM NaCl
1% NP-40
0.1% sodium deoxycholate
1 mM EDTA
0.1% SDS
1X Protease Inhibitors (Roche), added fresh
25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

ChIP High Salt Wash Buffer
50 mM HEPES, pH 7.5
500 mM NaCl
1% NP-40
0.1% sodium deoxycholate

1 mM EDTA
0.1% SDS
1X Protease Inhibitors (Roche), added fresh
25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

ChIP LiCl Wash Buffer
10 mM Tris-Cl, pH 8.0
250 mM LiCl
1 mM EDTA
1% NP-40
0.5% sodium deoxycholate
1X Protease Inhibitors (Roche), added fresh
25 mM NaF, added fresh
25 mM β-Glycerophosphate, added fresh

TE Buffer
10 mM Tris-HCl, pH 8.0, 1 mM EDTA

ChIP Elution Buffer
1% SDS
0.1 M $NaHCO_3$
Prepared fresh from 10% SDS and 1 M $NaHCO_3$.

2.3.5 FAIRE-PCR

Samples for FAIRE-PCR were crosslinked, lysed and sonicated as for ChIP, followed by FAIRE as described previously, with minor adaptations (Simon et al., 2012). The amount of chromatin used for FAIRE was the equivalent of 15 x 10^6 cells, and the starting volume was 500 µl, of which 30 µl were taken out as input. Inputs where processed as described for ChIP. To isolate open chromatin, a phenol/chloroform extraction was performed. In fact, the DNA not cross-linked to nucleosome segregates to the aqueous phase, whilst nucleosome-bound regions of chromatin remain trapped in the organic phase. Briefly, 500 µl of phenol/chloroform/isoamyl alcohol (Invitrogen) were added to the samples, after which tubes were vortexed for

10 seconds and centrifuged at 12,000 g for 5 minutes at RT. The aqueous layers (top) were then transferred to fresh 1.5 ml Eppendorf tubes and a second phenol/chloroform extraction was performed as just described. After centrifugation, a one-tenth volume of 3 M sodium acetate (pH 5.2), two volumes of of 95% (v/v) ethanol and 2 µl of 10 mg/ml glycogen (Sigma) were added to each tube containing the aqueous material and mixed.

After incubation at -80^{o}C for 30 minutes or longer, samples were centrifuged at 12,000 g for 15 minutes at 4^{o}C to precipitate the DNA. Supernatants were then carefully removed, and the DNA pellets washed with 500 µl of ice-cold 70% (v/v) ethanol. Following another centrifugation at 12,000 g for 5 minutes at 4^{o}C, the ethanol was discarded and the pellets were left to air-dry for 10 – 20 minutes before being redissolved in 50 µl of 10 mM Tris HCl (pH 7.5). De-crosslinking and protein digestion was carried out at 65°C overnight after adding 2 µl of Proteinase K from a 10 mg/ml stock. Finally, FAIRE samples and inputs were cleaned up using the PCR purification kit (Qiagen) according to manufacturer's instructions, eluting DNA off the column in 60 µl of water. An additional 100 µl of water were added to the inputs, and a standard range to use in the qPCR was prepared as described for ChIP. In order to obtain the percentage of target DNA recovered with each phenol/chloroform extraction compared to its starting quantity, the qPCR data were analysed as detailed in the section above. The qPCRs were performed as described in Section 2.1.8, and the primers used in this study are listed in Section 2.5.6.

2.3.6 ATAC-seq sample preparation

Samples for ATAC-seq were obtained as described previously, with minor adaptations (Buenrostro et al., 2015). P19 cells were stimulated as appropriate, trypsinised and counted. 100,000 cells suspended in 1 ml of full media were then moved to fresh 1.5 non-stick Eppendorf tubes, and spun down at 1,200 rpm for 5 minutes at 4°C. After carefully removing the supernants, pellets were washed once with 50 µl of cold PBS, and the cell suspensions were centrifuged at 1,200 rpm for 5 minutes at 4°C. To lyse the cells, samples were resuspended in 50 µl of cold Lysis Buffer and immediately spun down at 500 g for 10 minutes at 4°C. Supernatants were then discarded and the nuclei gently mixed with 50 µl of Transposition Reaction mix. The transposition reaction was incubated at 37°C for 30 minutes, directly

followed by a column purification step carried out using the PCR purification MinElute Kit (Qiagen). Here, 10 µl of Elution Buffer (Qiagen) were used for eluting the transposed DNA off the columns. To generate the libraries for next-generation sequencing, the transposed DNA fragments were amplified by PCR using the Veriti 96-well Thermal cycler (Applied Biosystems). Reagents and conditions are outlined below. For the primer sequences, see Section 2.5.7. After the PCR, samples were cleaned up with the PCR purification MinElute Kit (Qiagen) and eluted in 20 µl of Elution Buffer (Qiagen). Finally, an additional clean up step was performed using AMPure beads (Beckman Coulter, Inc.) to remove excess primers. After being thoroughly resuspended by vortexing, 36 µl of beads were mixed with the purified libraries, and tubes were incubated at RT for 5 minutes. Beads were then washed twice with freshly prepared 80% ethanol and air dried for 10 minutes, before the DNA was eluted with 20 µl of 0.1 X TE. Since the size of the library fragments provides a good indication of the success of the transposition reaction, the High Sensitivity DNA assay was performed on all samples using the 2100 Bioanalyzer (Agilent) by the Advance Sequence Facility at the Francis Crick Institute. For each biological replicate, two libraries were generated for each treatment condition to work as technical replicates, for a total of 8 samples for individual experiment. Upon Bioanalyzer analysis, for each biological replicate one sample per condition was picked amongst the technical replicates on the basis of the fragment size distribution. A total of 8 samples (4 from each biological replicate) was then submitted for next-generation sequencing.

Lysis Buffer

10 mM Tris HCl (pH 7.5)

10 mM NaCl

3 mM $MgCl_2$

0.1% NP-40

Transposition Reaction Mix

25 µl 2x TD Buffer (Illumina), 2.5 µl Tn5 Transposases (Illumina), up to 50 µl final volume with water

PCR Reaction Mix

10 µl transposed DNA

2.5 µl Nextera PCR primer 1 Universal Ad1_noMx 25 µM

2.5 µl Nextera PCR primer 2 Barcode 25 µM

25 µl NEBNext High-Fidelity 2x PCR Master Mix (New England Labs)

up to 50 µl final volume with water

PCR thermal cycler programme

72°C 5 minutes

98°C 30 seconds

98°C 10 seconds

63°C 30 seconds x 12 cycles

72°C 1 minutes

4°C ∞

2.4 Next Generation Sequencing and Bioinformatics analyses

Bioinformatics analyses were mostly performed in collaboration with Philip East and Harshil Patel from the Bioinformatics and Biostatistics Service (BABS) of the Francis Crick Institute.

2.4.1 RNA-seq and clustering of target genes

All RNA-seq experiments were performed as biological duplicates by Tessa Gaarenstroom in the lab and analysed by the Bioinformatics and Biostatistics Service (BABS) at the Francis Crick Institute. For details about sample preparation, library sequencing and reads alignment, see (Coda et al., 2017). Here, I report the criteria used to identify the differentially-regulated genes and to cluster them accordingly to the kinetic patterns of gene expression. To determine time point-specific differential expression against the SB-431542 sample, the data were analysed using DESeq (Anders and Huber, 2010) with a FDR threshold of 0.05 and a $\log_2$FC filter of 0.7 as cut-offs. The list of genes obtained was then further manually filtered by discarding genes with < 30 reads across all treatment conditions, as these likely represented

background noise. Genes showing a significant change only in the 'Untreated' condition were also removed, since they were likely due to long-term serum depletion of that sample compared to the others. This gave a total of 747 differentially-regulated genes, which were then clustered according to their kinetic expression patterns, using the following rules:

Induced sustained: 1 hr Activin $log_2FC \geq 0.7$; 8 hr Activin $log_2FC >$ 1 hr Activin log_2FC; Untreated $log_2FC \geq 0$.
Transient induced: 1 hr Activin $log_2FC \geq 0.7$; 8 hr Activin $log_2FC <$ 1 hr Activin log_2FC and ≥ 0; Untreated $log_2FC \leq 0.7$ and ≥ -0.7
Delayed: 1 hr Activin $log_2FC \leq 0.7$; 8 hr Activin $log_2FC \geq 0.7$; Untreated $log_2FC > 0$.
Repressed: 8 hr Activin $log_2FC \leq 0.7$; Untreated $log_2FC \leq 0$.

In all cases, the log_2FC was relative to the SB-431542 sample. For classification between 'baseline on' and 'baseline off', a gene was designated as 'baseline off' if less than 30 reads were detected in the SB-431542 sample. For classification as direct or indirect (see Appendix 8.2), a gene ws defined as indirect if in the cycloheximide treated 8 hr sample the log_2FC did not reach > 0.5 (for induced sustained or delayed genes) or < -0.5 (for repressed), or if 8 hr Activin $log_2FC >$ 1 hr Activin log_2FC (for transiently induced genes).

2.4.2 ChIP-seq library preparation, reads processing and alignment

Samples for ChIP-seq were prepared as described in Section 2.3.4. Following end repair, poly-A-tailing and adapter ligation, libraries were generated using Illumina TruSeq ChIP sample preparation kits. The Illumina kit Phusion enzyme was replaced by Kapa HiFi HotStart ready mix (Kapa Biosystems). The PCR was run before gel isolation using the Invitrogen SizeSelect E-gel system (SizeSelect gel protocol, Thermo Fisher Scientific). Post PCR AMPure XP beads (AMPure bead protocol, Beckman Coulter, Inc.) were used at a 1:1 ratio to maintain size integrity. Samples were multiplexed and 51 bp single end reads were generated on an Illumina HiSeq 2500. The raw reads were aligned to the mouse mm10 genome assembly using BWA 0.6.2r126 (Li and Durbin, 2009), with a maximum mismatch threshold of 2 permissible within a seed length of 51 bp. All other parameters were kept as default.

For the Pol II and histone ChIP-seq datasets the alignments were post-processed with picard-tools 1.107 for the removal of reads that could have arisen from PCR duplication (http://sourceforge.net/projects/picard/). All ChIP-seq experiments were performed as biological duplicates.

2.4.3 SMAD2 peak calling, annotation and generation of a SMAD2 consensus peak list

SMAD2 ChIP-seq was performed and analyzed by Tessa Gaarenstroom in the lab in collaboration with Philip East from the Bioinformatics and Biostatistics Service (BABS) of the Francis Crick Institute. Since the data obtained have been extensively used in this thesis, I outline below the methods they employed. For more details, see (Coda et al., 2017). SMAD2 peaks were called using MACS (q-value 0.05) (Zhang et al., 2008). In broad terms, MACS slides a window of a fixed lengh across the genome for each individual ChIP sample and relative control (that is, the input sample), counting the number of reads in each window. If this number is significantly different between the two samples accordingly to the Poisson distribution model, that region is defined as bound. If a control sample is not available, as it is the case for the ATAC-seq experiment (see Section 2.4.12), the number of reads in each window is compared against the values measured for the neighbouring genomic regions.

SMAD2 peaks were then annotated in terms of distance to nearest gene (user-defined) using ChIPPeakAnno (Bioconductor) (Zhu et al., 2010). SMAD2-binding loci were then associated with the closest differential gene as defined by the RNA-seq and ChIP-seq for Pol II Ser5 data (max distance 100 kb from TSS/TTS).

To quantify gene-associated SMAD2 activity across the four time points, a set of consensus binding sites was first generated by collapsing all binding sites identified from all comparisons. Binding loci were merged if they shared an overlap of at least a single nucleotide. Note that the mean length of overlap between two MACS peaks was 200 bp, showing that the majority of the peaks had a ‘sizeable’ overlap. Then, the total number of mapped reads per sample were scaled to 40×10^6 and the reads mapping to SMAD2 binding loci were counted. Finally, Gene-associated SMAD2 activity was quantified by summing the normalized read counts across consensus loci that were associated with an individual regulated gene (SMAD2 footprint). Duplicate reads were removed prior to consensus peak

quantification and reads were shifted to account for the fragment effect present in the sequencing data.

2.4.4 Pol II (Ser5P and Ser2P) and H3K27Ac/H3K9Ac ChIP-seq analysis

Accounting for variations in the sampling of the underlying sequence library caused by differences in the total number of reads sequenced is not straightforward for ChIP samples. Using normalisation to total mapped reads has been shown to be unreliable, since variations in enrichment efficiency can have dramatic effects on the proportion of reads attributed to signal. A more appropriate normalisation strategy is to use only reads derived from regions with high signal and low variance across all samples. Therefore, we used DiffReps 1.55.4 (Shen et al., 2013) to calculate sample-specific normalisation factors for the Pol II (either phosphorylated on Ser2 or Ser5 of the CTD), H3K27Ac and H3K9Ac ChIP-seq datasets. Furthermore, to be able to compare the signal for the same protein or modification from different treatments, the DiffReps normalization factors generated per time point were adjusted so that the geometric mean of the SB-431542 sample was constant across all comparisons from the same IP.

DiffReps was also used to determine sites of differential enrichment of Pol II (either phosphorylated on Ser2 or Ser5 of the CTD) at each time point relative to SB-431542. For each ChIP-seq experiment, replicate samples for each time point were compared to the SB-431542 sample. The corresponding input samples were also processed with DiffReps for background noise estimation. The "--nsd" and "--frag" parameters were set to "broad" and "200", respectively, and all other parameters were kept as default. Sites of differential binding for Pol II Ser2P and Ser5P from all the comparisons were overlapped with respect to knownGene gene loci from the UCSC (Karolchik et al., 2004) with the allowance for 2 kb of flanking sequence. More specifically, for each comparison, the sum of all differential bases found to overlap a given gene was calculated, and this was reported as a ratio of the length of the gene interval. Through inspection of the dataset it was found that genes exhibiting an overlap ratio ≥ 0.09 in any one of the comparisons were the most reliable. This gene list was then manually filtered to remove microRNAs and false positives, giving a dataset of 410 genes. This dataset represents the annotated genes which clearly show differential Pol II occupancy in response to signalling.

2.4.5 Definition of the high confidence dataset of target genes and associated binding sites

Genes differentially expressed according to RNA-seq, as well as genes found to contain differential Pol II occupancy at any time point, were taken forward as SMAD2 target genes. Overlapping genes between these two lists were selected. Outliers on either side were additionally included based on the presence of SMAD2 peaks conserved between the two biological ChIP-seq replicates within 100 kb of an annotated TSS or TTS. Moreover, genes that had differential Pol II binding, but did not the pass the stringent filter used for the RNA-seq data, were included if they were found to be differentially expressed according to less stringently filtered RNA-seq analysis and by manual inspection of the data. This led to a final list of 140 genes, associated with 478 SMAD2 consensus peaks.

2.4.6 Assessing histone modification or Pol II enrichment around SMAD2 peaks and target gene TSSs using metaprofiles

To visualise histone modification or Pol II changes over time for specific peak and gene sets several metaprofiles were generated. In all cases the reads for the two biological replicates were added together to obtain more read depth. The read depth vectors for a 5 kb window across SMAD2 binding sites (the average MACS-called summit for each consensus peak) and SMAD2 target gene TSS were obtained. These vectors were then adjusted using the normalisation factors generated by DiffReps as described in Section 2.4.4, and metaprofiles were constructed by averaging per-nucleotide read depths across contributing loci. The metaprofiles were smoothened using local polynomials. In the case of TSS-centric profiles, genes in a reverse orientation were accounted for, and all plots run from 5' to 3'.

2.4.7 Hierarchical clustering of Pol II (Ser5P and Ser2P) and H3K27Ac/H3K9Ac ChIP-seq data

To quantify the $\log_2$FC of Pol II (either Ser5P or Ser2P), the total numbers of reads mapping to the entire genomic extent of a gene (±2 kb) was determined, adjusting the counts from each sample by its DiffReps normalisation factor. The mean values of these counts across the replicates was taken and compared to the mean normalized replicate value of the relevant SB-431542 sample. A similar analysis was

performed with the H3K9Ac and H3K27Ac datasets to quantify the respective log_2FC relative to the SB-431542 sample at the SMAD2 target gene TSSs (± 2.5 kb), or to determine the similarity between the histone modifications at SMAD2 binding sites (± 2.5 kb) across the time points. In all cases, hierarchical clustering was employed using euclidean distance.

2.4.8 Correlation of SMAD2 binding with overall levels of histone acetylation

From the analysis of histone modifications at SMAD2 binding sites described above, regions of high or low acetylation for H3K9 and H3K27 were defined in collaboration with Tessa Gaarenstroom selecting the nodes obtained via hierarchical clustering that show either high or low H3K9Ac/H3K27Ac (see Figure 5.8). First, 72 'low acetylation' SMAD2 peaks and 44 'high acetylation' SMAD2 peaks were identified by taking the overlap between the H3K9Ac and H3K27Ac datasets. Then, to find the peaks with low or high acetylation in the SB-431542 sample, the 100 most or least enriched acetylated peak regions in this sample were selected for each histone modification and the intersection of peaks selected. This resulted in 60 'low H3Ac SB baseline' and 63 'high H3Ac SB baseline' peaks. To define the low acetylated peaks in the SB-431542 sample which upon Activin treatment remained lowly acetylated or acquired acetylation, 120 peaks with low levels of H3K9Ac and H3K27Ac in the SB-431542 state were selected using arbitrary cut offs. Finally, these peaks were divided into the 'high H3Ac increase' group (62 peaks) or the 'no H3Ac increase' group (58 peaks) in response to signalling, using a log_2FC of ≥ 0.7 cut off for both acetylation states in the 8 hr Activin sample relative to the SB-431542 sample. In all cases, to compare SMAD2 normalised read depth across two groups of peaks unpaired t-tests were performed using the standard analysis settings within the Graphpad Prism 6 software.

2.4.9 IGV browser displays

ChIP-seq and ATAC-seq data were visualised using the IGV browser. The Pol II and histone acetylation tracks are shown normalised using the DiffReps-generated factors, read extension by 100 bp and smoothing over 10 bp windows, while the SMAD2 tracks have been extended and smoothened. The H3 tracks represent the

raw coverage obtained from the aligned reads. For details about the visualisation of ATAC-seq data see Sections 2.4.11 and 2.4.14.

2.4.10 Motif enrichment analysis

Taking all 478 consensus peaks associated with a SMAD2 target gene, the MACS-defined summits from each contributing MACS-called peak were selected (757 in total). The 500-bp sequence surrounding each summit was extracted and submitted for known and *de novo* motif discovery using the MEME-ChIP suite. Default settings were used with the exception of MEME, where the following parameters were used: site distribution – any number of occurrences; total number – 20.

2.4.11 ATAC-seq: reads processing, filtering and alignments

The ATAC-seq experiment was performed in biological duplicate. Samples were prepared as described in Section 2.3.6, and 51-bp pair end reads were generated on an Illumina HiSeq 2500. Raw reads from each sample were adapter-trimmed using cutadapt 1.9.1 with parameters "-a CTGTCTCTTATA -A CTGTCTCTTATA, minimum-length = 25 – quality-cutoff = 20". BWA 0.6.2 (Li and Durbin, 2009) with default parameters was then used to perform genome-wide mapping of the adapter-trimmed reads to the mouse mm10 genome downloaded from the UCSC (Karolchik et al., 2004). Read group addition, duplicate marking and insert size assessment was performed using, respectively, the picard tools AddOrReplaceReadGroups, MarkDuplicates and CollectMultipleMetrics 2.1.1. Finally, reads mapped to mitochondrial DNA were removed using the pairToBed command from BEDTools 2.26.0-foss-2016b (Quinlan and Hall, 2010), and additional filtering was performed to remove read pairs that were discordant, mapped to different chromosomes, ambiguously mapped, had an insert size > 2kb, and mismatch > 2 in any reads. All bam file sorting and indexing was performed with samtools 1.3.1 (Li et al., 2009). Replicate-level reproducibility across samples was assessed by counting read pairs that overlapped the union set of ATAC-seq peaks (see 2.4.12) using the Subread featureCounts tool version 1.5.0 (Liao et al., 2014) with the parameters "- O, minOverlap 1, primary, ignoreDup -p -B -C – donotsort". Read counts between replicates were then plotted on a $\log_{10}$ scale after quantile normalisation using the "normalize.quantiles" function in R version 3.3.1. When individual samples were

combined to increase the read depth, the filtered alignments from each library were merged using the picard MergeSamFiles command. Duplicate marking was re-performed on the merged alignments, and they were subsequently filtered for duplicate reads, leading to 500 – 600 million reads for merged sample.

2.4.12 ATAC-seq peak calling, annotation and generation of coverage tracks

Genome-wide ATAC-seq peaks and normalised BedGraph coverage tracks for the merged samples were obtained using MACS2 callpeak 2.1.1.20160309 (Zhang et al., 2008) with the parameters “gsize = mm, keep-dup all, nomodel, shift – 100, extsize 200 - B, SPMR, cutoff-analysis – broad”. The annotate Peaks.pl program from HOMER 4.8 (Heinz et al., 2010) was used to annotate ATAC-seq peaks relative to mm10 RefSeq features downloaded from UCSC. BedGraph coverage tracks generated by MACS2 were converted to BigWig using the bdg2bw utility available in the kent tools package from UCSC 20161115 (Kent et al., 2010), and visualised using the IGV genome browser. Finally, in order to obtain a union set of intervals the ATAC-seq peaks from all samples were merged.

2.4.13 Overlapping of BED files and comparison of lists of elements

To intersect or subtract different BED files, BEDTools or the Table browser found within the UCSC Genome browser were generally used (Karolchik et al., 2004). To identify common elements within two, three of four different lists, the Venny tool was used http://bioinfogp.cnb.csic.es/tools/venny/.

2.4.14 DiffReps analysis and hierarchical clustering

The individual replicate samples were employed to call differential ATAC-seq sites between conditions. Intervals of differential ATAC signal were obtained using Diffreps 1.55.4 (Shen et al., 2013) with the parameters “window 200, step 20, nsd broad, frag 0, noanno, nohs”. To identify changes in chromatin accessibility at SMAD2 binding sites, the resulting files were intersected with the consensus ATAC-seq and SMAD2 peaks using the BEDTools. To increase the stringency of the data, I decided to consider only the diffReps interval which had a padj lower then 0.1 and a $\log_2$FC greater than 0.5 as an absolute value. Hierarchical clustering using

euclidean distance was employed to group the different SMAD2 binding sites on the base of the diffReps interval log_2FC values.

2.4.15 Footprint and genome-wide motif analyses

Footprint analysis was performed across the union set of ATAC-seq peaks on the unshifted, merged alignments using the "wellington_footprints.py" command in the pyDNAse package 0.2.5 (Piper et al., 2013) with parameters "-fp 6,30,2 -fdr 0.05 -A". Strand-specific coverage tracks representing Tn5 transposase cut sites were then generated using the pyDNAse command "dnase_wig_tracks.py" with the "-A" parameter, and visualised using the IGV genome browser. Finally, the individual footprints from all samples were merged in order to obtain a union set of footprint intervals.

To retrieve the genome-wide locations of all known motifs in the HOMER database, the "scanMotifGenomeWide.pl" command was used with parameters "-bed -int -keepAll". The resulting files were then intersected with the merged footprint intervals present in both ATAC-seq and SMAD2 peaks, using the BEDTools command intersectBed and setting the percentage of overlap to 1E-9 (i.e. 1bp). This analysis allowed calculation of motif footprint frequency (see Figure 6.18), and the generation of different kinds of footprint heatmaps and plots for each of the known motifs in the HOMER database. First, motif centric heatmaps were obtained using the pyDNAse "dnase_to_javatreeview.py" command with parameters "--window_size 100 -A -n". Then, in order to identify changes in motif accessibility over time, differential heatmaps were generated by calculating the absolute value of the difference between each signalling condition compared to the SB-431542 sample. Additional differential heatmaps were also produced to conserve the strand information. This was achieved by extracting the absolute difference between each signalling condition and the SB-431542 sample, and then weighting it by +/- 1 depending on whether a change in sign was observed. Finally, the average profile plots around motif sites were generated using the pyDNAse "dnase_average_profile.py" command with parameters "--window_size 100 -A".

2.4.16 Statistical Analysis

Experiments were performed at least twice independently (biological duplicates) and the majority were performed at least three times. Within each qPCR experiment, technical duplicates were run. Statistical analyses were performed using an unpaired t-test unless otherwise specified. $p < 0.01$ was considered statistically significant.

2.5 List of reagents used in this study

2.5.1 CRISPR/Cas9 guide oligos

Target gene	Guide oligo (listed 5' to 3')
Lefty1 SBS 5' guide fwd	CACCGGTTGTGTGGCCTACGACTA
Lefty1 SBS 5' guide rev	AAACTAGTCGTAGGCCACACAACC
Lefty1 SBS 3' guide fwd	CACCGAGTGTCAAACGACAATATG
Lefty1 SBS 3' guide rev	AAACCATATTGTCGTTTGACACTC
Lefty2 SBS 5' guide fwd	CACCGTCGCAGCATAGAAGCTCCGA
Lefty2 SBS 5' guide rev	AAACTCGGAGCTTCTATGCTGCGAC
Lefty2 SBS 3' guide fwd	CACCGTATCAGTCCCGTGTGACCC
Lefty2 SBS 3' guide rev	AAACGGGTCACACGGGACTGATAC

2.5.2 PCR primers for screening of CRISPR/Cas9-mediated deletions

Genomic region	Sequence (listed 5' to 3')
Lefty1 fwd	ACTATTTTGGAGGTCGCCCA
Lefty1 rev	GTCAGGAGCTCAGTCGTGTG
Lefty2 fwd	GAAGCACGGAGGTGTTTTGC
Lefty2 rev	GGAATACCTGGGAGTCGCTT

2.5.3 Primers for gene expression analysis

Target gene	Sequence (listed 5' to 3')
Eomes fwd	ACCAAAACACGGATATCACCCAGC
Eomes rev	AGGGGAATCCGTGGGAGATGGA
Fgfr2 fwd	CCCTGCGGAGACAGGTAACA
Fgfr2 fwd	TTGCCCAGCGTCAGCTTATC

Foxh1 fwd	TCGGTGCTTCCATAAGGTGCCC
Foxh1 rev	CGACGGCACAGGGCAGTGTT
Gapdh fwd	TCTTGTGCAGTCCCAGCCT
Gapdh rev	CAATATGGCCAAATCCGTTCA
Hes1 fwd	CATGGAGAAGAGGCGAAGGG
Hes1 rev	ATGCCGGGAGCTATCTTTCT
Id2 fwd	GCAGATCGCCCTGGACTCGC
Id2 rev	CAGATGCCTGCAAGGACAGGATGC
Lefty1 fwd	GCTCGATCAACCGCCAGTCCTG
Lefty1 rev	GCCACCTCTCGAAGGTTCTGGCT
Nanog fwd	CTGATTCAGAAGGGCTCAGCACCA
Nanog rev	AGTCTGGCTGCCCCACATGGA
Nodal fwd	CACCGTCCCCTCTGGCGTAC
Nodal rev	TCCGGTCACGTCCACATCTTGC
Pitx2 fwd	TCTGTCACCATCCCCAGGCGT
Pitx2 rev	TGGCCCTTATCTTTCTCTGCGACTT
Pmepa1 fwd	AGCCCGCTCCTTCATCAGCC
Pmepa1 rev	CATAGACCTGTGGCTCCGGCA
Pou5f1 fwd	CTGTAGGGAGGGCTTCGGGCACTT
Pou5f1 rev	CTGAGGGCCAGGCAGGAGCACGAG
Smad7 fwd	CCCCCGGCTGAGAGGCTCAT
Smad7 rev	CACCTGCTGCCAGTCTGCCC
Smarca4 fwd	TGAGCTCCCATCCTGGATCA
Smarca4 rev	TAGCCTTCAGGGTCTTGAGC
T fwd	TGTGACCAAGAACGGCAGGAGG
T rev	CAGCGGTGGTTGTCAGCCGT
Tdgf1 fwd	ACGAGCCGTCGAAGATGGGGT
Tdgf1 rev	TCTCTCCCAGCAACGGGTCCA
Trh fwd	CAGGTCCGGCCAGAACGTCG
Trh rev	TCCAGGAATCTAAGGCAGCACCAA

2.5.4 siRNAs

Target gene	Catalogue number	Type
Foxh1	D-042546-01 D-042546-02 D-042546-03 D-042546-04	siGenome individual duplexes as a pool

Smarca4	D-041135-01 D-041135-03	siGenome individual duplexes as a pool
Non targeting control	D-001810-01-20	ON-TARGETplus™ control siRNA

2.5.5 Antibodies and their applications

Target protein	Catalogue number	Application
FOXH1	Home made (see methods)	ChIP
H3K27Ac	Abcam Ab4729	ChIP
H3K27Me3	Active Motif 39155	ChIP
H3K4Me1	Abcam Ab8895	ChIP
H3K9Ac	Abcam Ab4441	ChIP
H3K9Me3	Abcam Ab8898	ChIP
H3K18Ac	Abcam Ab1191	ChIP
H3K23Ac	Millipore 07-355	ChIP
Histone H3	Abcam Ab1791	ChIP
MCM6	Santa Cruz Biotechnology SC-9843	WB
MYC (9E10)	In-house, clone 9E10	WB
MYC (ChIP)	Millipore 05-724	ChIP
pSMAD2	CST 3108	WB
pSMAD2 (A5S)	Millipore 04-953	WB
Pol II (Ser2P)	Diagenode C15200005	ChIP
Pol II (Ser5P)	Diagenode C15200007	ChIP
SMAD2	CST 5339	ChIP
SMAD2/3	BD 610843	WB
TDGF1	CST 2818	WB
TUBULIN	Abcam Ab6160	WB

The secondary antibodies used for Western Blot were the following:

Goat anti-Rabbit HRP conjugated (DAKO)

Goat anti-Mouse HRP conjugated (DAKO)

Rabbit anti-Rat HRP conjugated (DAKO)

2.5.6 Primers for ChIP-PCR and FAIRE-PCR

Genomic region	Sequence (listed 5' to 3')
Gapdh coding region fwd	CACTCCCCTTCCCAGTTTCC
Gapdh coding region rev	GCATCTTCTTGTGCAGTGCC
Lefty1 5' fwd	ACTATTTTGGAGGTCGCCCA
Lefty1 5' rev	TCTGCCCCTACCTCCTTGTT
Lefty1 SBS fwd	GGTGGTGAGACACATCGCAGCA
Lefty1 SBS rev	ACTTCCCCCTGGCTGATTGGTCA
Lefty1 3' fwd	CAGATGCAAAGGGGGACTGT
Lefty1 3' rev	TTGGCTGGCCTTGCTACATT
Lefty1 intergenic fwd	CAGGAGGCCTGAGGTTCATC
Lefty1 intergenic rev	TGAGTCGGGTGGCAATTGAG
Lefty1 TSS fwd	AAGCTGTTCCGTACCGTACCATTC
Lefty1 TSS rev	AAGAGGAGCCTTGGATGTGTGTGT
Lefty1 TTS fwd	TCCCTAATGGGCAATCCCTGTGTGT
Lefty1 TTS rev	AGACCACTGGGTCAGAGCGCC
Lefty2 5' fwd	CACGGAGGTGTTTTGCATGG
Lefty2 5' rev	CTGTTCGCTCCATCCTCTGG
Lefty2 SBS fwd	GGAGAGGCCTAGCTTTTGCAT
Lefty2 SBS rev	CCCAAAGAAGGAAGCAGATGTG
Lefty2 3' fwd	TGGGGCGTTCCCTAAATTGG
Lefty2 3' rev	CACTCCTGGGGTGACCTCTA
Negative control fwd	CTACAGTCCTCCCCGTACCA
Negative control rev	CCATGCCCTTGACAATCCCT
Negative control 2 fwd	CATCTCCTTTCAGGGTCCAA
Negative control 2 rev	ATAGCTCTGTCTGGCCAAGG
Nodal SBS fwd	TGTCCTCTGGGGCCAGACGG
Nodal SBS rev	TCCGCCTCCAGGTCGTGAGG
Pitx2 5' fwd	TTCCCAACGCAAAGAGAAGGA
Pitx2 5' rev	TCCCGGCTCCTTTATCAACC
Pitx2 SBS fwd	AGCTGCTCTCTGGGGCGACT
Pitx2 SBS rev	ATTTGCGGGAGGCTGGCCG
Pitx2 3' fwd	TTTTGGACCCCCAAAGAGGG
Pitx2 3' rev	TAGCTCTAGAGGCCTCACCC
Pmepa1 5' fwd	AAGGACAGCACACAGACACC
Pmepa1 5' rev	TGTACCCCAGTCGCATCAAC
Pmepa1 SBS fwd	TCCGCTTTATCTGCGGGAAG
Pmepa1 SBS rev	CAGGTTTTGGGGGTCAGACA

Pmepa1 3' fwd	TGCAGATGCTACCGTGTGTT
Pmepa1 3' rev	TCATCACCTTGCAACCTCCC
Pou5f1 5' fwd	GGCTAGGGCACATCTGTTTCA
Pou5f1 5' rev	CCCTAAGCGTGCCTAGAGTA
Pou5f1 SBS fwd	GGAGGTTGAGAGTTCTGGGC
Pou5f1 SBS rev	AGGAAGGGCTAGGACGAGAG
Pou5f1 3' fwd	TAGAGCCACTGACCCTAGCC
Pou5f1 3' rev	ACAGCCTCAAAAAGCCCAGA
Smad7 5' fwd	GTGCTGAGACCCTTCGAGAC
Smad7 5' rev	CTTGGTGTTTTGCAGACCCG
Smad7 SBS fwd	CCTAGGCTCCGCAAGGTTAG
Smad7 SBS rev	AACCCGGTGGCATACAGATG
Smad7 3' fwd	TGATATGCATTCCCAAAGGGGT
Smad7 3' rev	AGACTCTAGAAGTCGGTCCCA
Smad7 intergenic fwd	CGTCCCATCCAGACAAGCTG
Smad7 intergenic rev	ACGTGAGTGGTGCTAATCCC
Smad7 TSS fwd	TCCTGGCCGGTGTAAATGTC
Smad7 TSS rev	CCGGTTAGTGGCCCGATTTA
Tdgf1 SBS fwd	ACCCTCCCTCACCTTCGCCC
Tdgf1 SBS rev	TTCCCAGGCCTGTCGGGGTC
Trh SBS 1 5' fwd	TACCAGGGTGTTTCTCCCCA
Trh SBS 1 5' rev	CCACTTGCAATCGTGGCTTC
Trh SBS 1 fwd	GACAGACTGCCAGTGAACCA
Trh SBS 1 rev	GTAATCCACCCTCCCTCCCT
Trh SBS 1 3’ fwd	ATGTCCTCTTGACCAACTGGC
Trh SBS 1 3’ rev	CGTGGGGTCAGGACTCATAA
Trh SBS 2 fwd	TGTCCCTCCCCCAGATTTCA
Trh SBS 2 rev	CCACGTTTTTCCCAGGAGGA
Trh SBS2 3' fwd	CGCAGAAATCTTCAGCCTGC
Trh SBS2 3' rev	CCTGGTGTCACTGGGAACTC

2.5.7 Nextera primers for indexing

Adaptor	**Sequence (listed 5’ to 3’)**
Ad 1_noMX	AATGATACGGCGACCACCGAGATCTACACTCG TCGGCAGCGTCAGATGTG
Ad 2.1	TAAGGCGACAAGCAGAAGACGGCATACGAGAT TCGCCTTAGTCTCGTGGGCTCGGAGATGT

Ad 2.2	CGTACTAGCAAGCAGAAGACGGCATACGAGAT CTAGTACGGTCTCGTGGGCTCGGAGATGT
Ad 2.3	AGGCAGAACAAGCAGAAGACGGCATACGAGAT TTCTGCCTGTCTCGTGGGCTCGGAGATGT
Ad 2.4	TCCTGAGCCAAGCAGAAGACGGCATACGAGAT GCTCAGGAGTCTCGTGGGCTCGGAGATGT
Ad 2.5	GGACTCCTCAAGCAGAAGACGGCATACGAGAT AGGAGTCCGTCTCGTGGGCTCGGAGATGT
Ad 2.6	TAGGCATGCAAGCAGAAGACGGCATACGAGAT CATGCCTAGTCTCGTGGGCTCGGAGATGT
Ad 2.7	CTCTCTACCAAGCAGAAGACGGCATACGAGAT GTAGAGAGGTCTCGTGGGCTCGGAGATGT
Ad 2.8	CAGAGAGGCAAGCAGAAGACGGCATACGAGAT CCTCTCTGGTCTCGTGGGCTCGGAGATGT

Chapter 3. NODAL/Activin signalling induces multiple patterns of gene expression

3.1 Introduction

During early embryonic development, NODAL/Activin signalling induces programmes of gene expression by activating SMAD2/3 via heterotetrameric type I–type II receptor complexes (Pauklin and Vallier, 2015, Wu and Hill, 2009). It is now becoming clear that, in this context, cells respond differently to different times of exposure to the ligands (van Boxtel et al., 2015, Hagos and Dougan, 2007). So far, NODAL/Activin transcriptional targets and SMAD2/3 genome binding events have been characterised only in the autocrine chronic signalling condition, or inductions with ligands were performed starting from a non-signalling-inhibited baseline (Beyer et al., 2013, Brown et al., 2011, Estaras et al., 2015, Kim et al., 2011, Mullen et al., 2011). As a consequence, how NODAL/Activin signalling regulates gene expression over time is poorly understood. The sequence of events occurring on chromatin from SMAD binding to transcription activation is also unknown.

In this thesis, I have sought to address these questions using the mouse embryonic teratoma cell line P19, which provides an excellent tool to study NODAL/Activin signalling. In the untreated state, these cells secrete active NODAL and they can be acutely or chronically stimulated using recombinant Activin, which is commonly employed *in vitro* to mimic NODAL activity (Kumar et al., 2001, Labbe et al., 1998). In addition, treatment of P19s with the NODAL/Activin type I receptor inhibitor SB-431542 completely blocks the pathway without affecting the cells' phenotype (Ross et al., 2006). In contrast to ESCs (Vallier et al., 2005), P19s do not differentiate when cultured in the presence of SB-431542, allowing us to investigate SMAD-dependent transcription from a signal-inhibited baseline. Furthermore, in response to NODAL/Activin signalling P19s upregulate both pluripotency and mesendoderm genes, hence mimicking the activity of the pathway in early vertebrate development (Nakaya et al., 2008, Marikawa et al., 2011). Thus, the results obtained in P19s are biologically relevant and can be highly informative to understand how cells interpret NODAL/Activin signals *in vivo* during pluripotency and differentiation.

In this chapter, I first describe the dynamics of SMAD2 activation and gene expression in response to NODAL/Activin in P19 cells, focusing on the four signalling states – inhibited, acute, prolonged and chronic – which constitute the framework of the subsequent analyses. By performing a series of qPCR experiments for representative NODAL/Activin target genes, I also investigate the requirement of continuous SMAD signalling and/or protein synthesis for the modulation of transcription in the long term. I then introduce the genome-wide datasets obtained by me and others in the lab to characterize over time: 1. The transcriptional responses to NODAL/Activin signalling; 2. SMAD2 chromatin binding; 3. Enrichment of two different forms of RNA Polymerase II (Pol II). I outline the criteria used to integrate these data and define a high confidence set of NODAL/Activin target genes and associated SMAD2 peaks, which is relevant to address several questions throughout the thesis. Finally, I validate the method employed to associate SMAD2 peaks to target genes by carrying out CRISPR/Cas9-mediated deletions of the SMAD2 binding sites upstream of two Activin-regulated genes, and verifying the effects of these mutations on gene expression in response NODAL/Activin signalling.

3.2 Results

3.2.1 The dynamics of NODAL/Activin signalling in P19 cells

I first set out to characterise SMAD pathway activation in P19 cells over time in response to NODAL/Activin signalling. Because NODAL and Activin induces phosphorylation of both SMAD2 and SMAD3, I initially checked by Western blot their protein levels in P19 cells alongside with two other mouse cell lines routinely used in the lab to study this pathway, C2C12s and EpH4s. The membrane was probed with an antibody that recognizes SMAD2 and SMAD3 equally (Mike Howell, unpublished), and SMAD3 was found to be undetectable in P19s, but readily detectable in the other two lines. The levels of SMAD2 were instead comparable in all three lines (Figure 3.1). I concluded that in P19 cells SMAD2 is the predominant receptor-regulated SMAD downstream of NODAL/Activin, and therefore phosphorylated SMAD2 (pSMAD2) was used as a readout for monitoring the pathway activity over time.

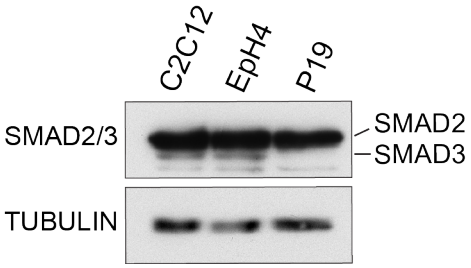


Figure 3.1. SMAD2 is the predominant receptor-regulated SMAD downstream of NODAL/Activin in P19 cells.
Whole cells lysates were extracted from P19 cells and two other mouse cell lines, C2C12 and EpH4. Levels of SMAD2/3 and TUBULIN (as a loading control) were assayed by Western blot. Compared to the other cell lines, P19 cells have undetectable levels of SMAD3, but similar levels of SMAD2.

It was known from the work from others in the lab that P19 cells have a basal level of pSMAD2 (see lane 1, Figure 3.2), due to an autocrine production of NODAL (Ross et al., 2006). Since I wanted to investigate the dynamics of SMAD pathway activation starting from a signal-inhibited baseline for the reasons outline above, cells were stimulated with a saturating dose of Activin (20 ng/ml) in full media after an overnight treatment with SB-431542, followed by washout. By harvesting cells at different time points, this experimental set-up allowed us to characterise the response to both acute and prolonged signalling conditions. Note that in the control condition, SB-431542, the inhibitor was replaced after the washout together with fresh media for 1 hr in order to control for the addition of media containing serum.

Growth factors contained in the serum have been reported to rapidly induce transcription of target genes via the MRTF-SRF and TCF-SRF signalling pathways (Esnault et al., 2014). Thus, replacing the media in the SB-431542 condition ensures that the differential transcriptional responses observed upon 1 hr of Activin are due to SMAD pathway activation, rather than being a consequence of serum stimulation. Western Blot analysis of whole cell lysates over a time course of Activin treatment revealed that phosphorylation of SMAD2 was induced within 30 minutes, peaked at 60 – 90 minutes, and then attenuated to lower levels at later time points, whilst levels of total SMAD2 did not change (Figure 3.2). The untreated state was characterised by low levels of pSMAD2 (lane 1, Figure 3.2), which are due to an autocrine production of NODAL and GDF3 in a NODAL/Activin dependent manner (Coda et al., 2017). These ligands signal via ACVR1B, ACVR2A/B and TDGF1, the last of which is under transcriptional control of NODAL/Activin signalling (Schier, 2009).

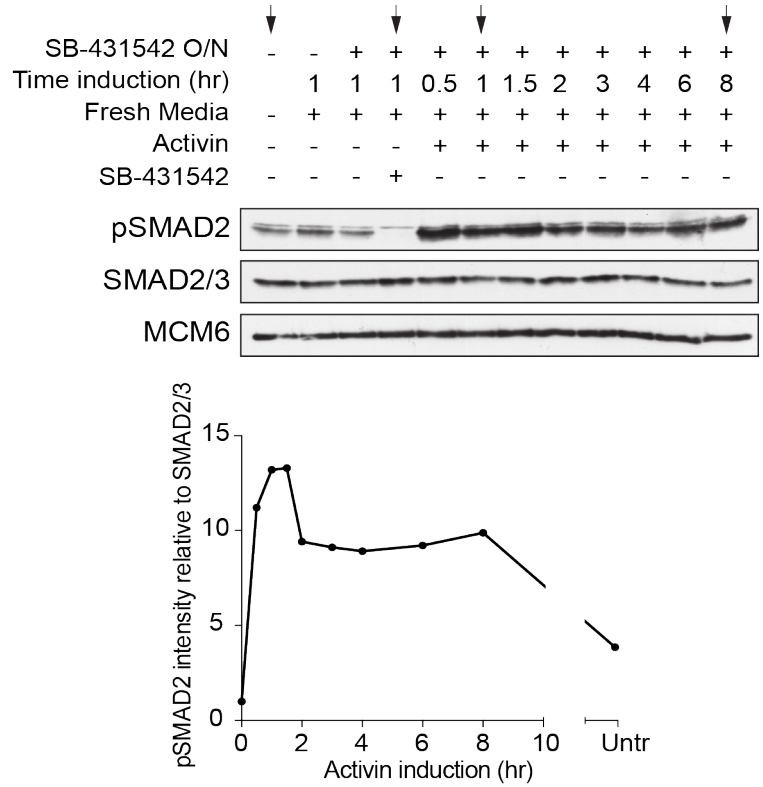


Figure 3.2. SMAD2 phosphorylation kinetics in response to NODAL/Activin signalling.
P19 cells were treated with 10 μM SB-431542 overnight or not; after which it was washed out and cells were incubated with fresh media alone or with 10 μM SB-431542 or 20 ng/ml Activin for the times indicated. *Figure 3.2 continued on next page*.

Indeed, TDGF1 was lost in the presence of SB-431542 and required up to 8 hr to be detected by Western Blotting in conditions where SB-431542 had been washed out and replaced with media only (Figure 3.3). As a consequence, SMAD2 phosphorylation slowly accumulated over time upon SB-431542 washout. This delay is due to the fact that receptors, whose transcription has been kept repressed by SB-431542, needed to be re-expressed upon secretion of NODAL and GDF3 to trigger the downstream pathway. In contrast, Activin does not require the presence of TDGF1 to induce SMAD2 phosphorylation via ACVR1B and ACVR2A/B complexes, enabling us to acutely stimulate P19 cells with this ligand starting from the SB-431542-treated state (Kumar et al., 2001). Since SMAD pathway activation was routinely achieved in the lab using a saturating dose of Activin (20 ng/ml), I then investigated if pSMAD2 induction was dependent on Activin dose. The Western Blot analysis in Figure 3.3 revealed that the acute phosphorylation of SMAD2 after 1 hr of treatment was obtained also with a ten-fold lower dose of Activin (2 ng/ml), but it was lost if cells were exposed to a minimal dose of Activin (0.5 ng/ml). In this case, pSMAD2 signal accumulated over time following the dynamics observed in response to media alone, possibly due to the initial limiting ligand concentration (Figure 3.3).

Overall, these time course experiments show that pSMAD2 in P19 cells exhibits a rapid peak-like pattern in response to Activin treatment, which modestly attenuates with prolonged signalling. They also suggest that the modulation of pSMAD2 at later time points is mediated by the synthesis of ligands and receptors which are under control of the signalling pathway itself.

Figure 3.2 continued.
Immunoblots of cell extracts were probed with antibodies against pSMAD2, total SMAD2/3 and MCM6 as loading control. The four conditions used in the RNA-seq experiment (see Figure 3.5) are indicated by the arrows. Below, the quantification of densitometry measurements from the same experiment, normalised to total SMAD2/3. SB, SB-431542; Untr, untreated.

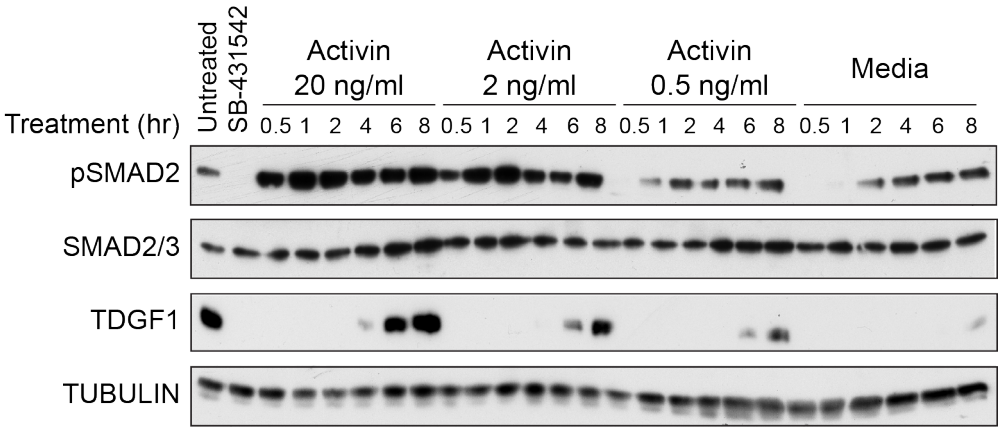


Figure 3.3. pSMAD2 kinetics in response to different concentrations of Activin. P19 cells were treated with 10 μM SB-431542 overnight, followed by washout and incubation with the concentrations of Activin shown, or with fresh media alone, or SB-431542 was replaced for 1 hr. pSMAD2, SMAD2/3, TDGF1 and TUBULIN as a loading control were assayed by Western blot

3.2.2 Activin stimulation directly induces/represses transcription

In order to characterise the transcriptional responses to both acute, sustained and chronic NODAL/Activin signalling, the same samples of Figure 3.3 were assayed for the expression levels of representative transcripts known to be targets of the SMAD pathway (Beyer et al., 2013, Kim et al., 2011, Levy and Hill, 2005). The qPCR results show that genes were induced with distinct kinetics in response to the signalling: the expression of genes like *Lefty1*, *Pmepa1* and *Nodal* rapidly increased after Activin treatment, others like *T* required a longer time to be induced, whilst some genes like *ID2* were instead repressed in response to the signalling (Figure 3.4). Interestingly, although the different concentrations of Activin partly affected the gene expression profiles in term of absolute levels, the kinetics of induction were conserved across Activin doses. This observation suggested that time of ligand exposure rather than dose is responsible for the transcriptional responses elicited by the signalling pathway.

Before I joined the Hill lab, another student (Tessa Gaarenstroom) extensively investigated the responses to NODAL/Activin signalling in P19 cells on a genome-wide scale in the same experimental setting. She performed RNA-seq (biological duplicates) in four different conditions representative of the distinct signalling states – inhibited, acute, prolonged and chronic – already discussed.

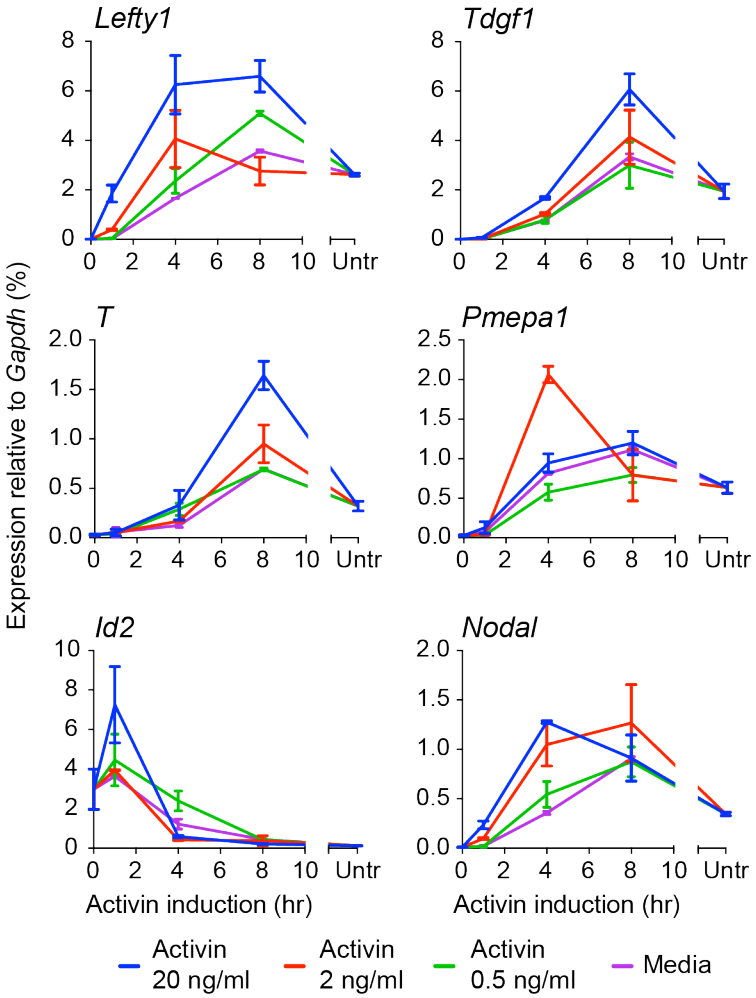


Figure 3.4. Gene expression kinetics in response to different concentrations of Activin.
qPCR measurements of the expression of NODAL/Activin target genes in the same samples described in Figure 3.3. P19 cells were treated with 10 μM SB-431542 overnight, followed by washout and incubation with the concentrations of Activin shown, or with fresh media alone, or SB-431542 was replaced for 1 hr. A representative experiment (means ± SD) is shown. Untr, untreated.

To achieve these signalling states, cells were either treated with SB-431542, or treated with Activin 20 ng/ml for 1 hr or 8 hr, or left untreated, respectively. This analysis allowed her to identify 747 differentially-regulated genes, amongst which there are pluripotency genes like *Nanog* or *Pou5f1* and mesendoderm genes like *T* or *Eomes*, consistent with the ESC-like nature of P19 cells (van der Heyden and Defize, 2003, Kumar et al., 2001, Nakaya et al., 2008). She then went on to define four classes of responsive genes which have distinct expression kinetics: 'induced sustained' (acutely upregulated genes whose transcription persists over time e.g. *Lefty1, Tdgf1*); 'delayed' (genes only significantly induced after prolonged signalling

e.g. *Trh*, *Fst*); 'transient induced' (acutely upregulated genes whose transcription subsequently declines *Smad7, Hes1*); 'repressed' (genes that are actively inhibited in response to signalling e.g. *Id1, Id2, Id3*) (Figure 3.5). For details about cluster generation, see Section 2.4.1.

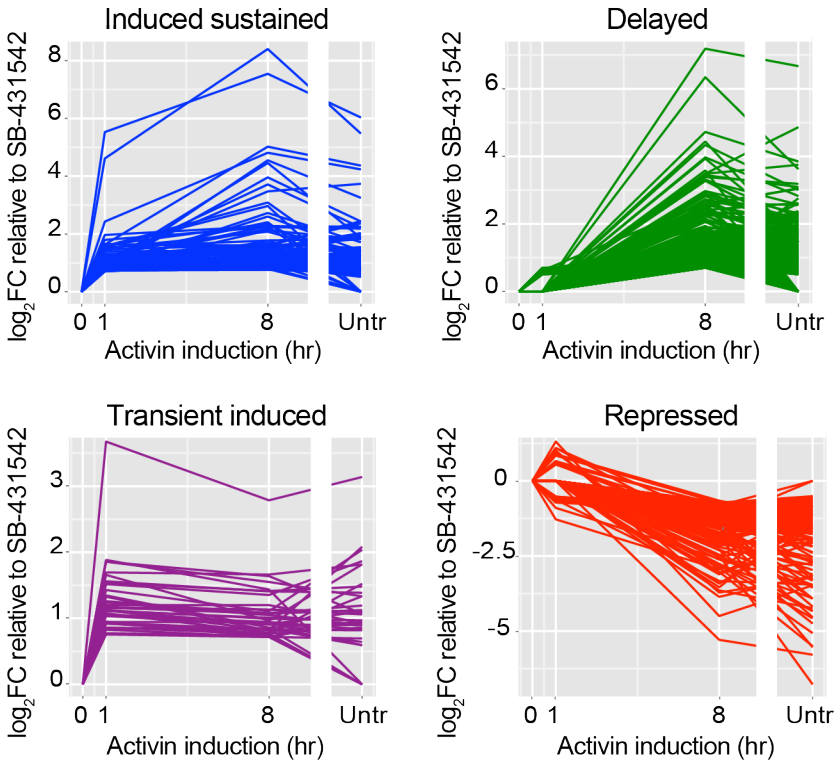


Figure 3.5. Characterisation of the transcriptional responses to NODAL/Activin signalling.
RNA-seq on P19 cells treated as in Figure 3.2 (see arrows) was performed and analysed by Tessa Gaarenstroom. The target genes were divided in four different clusters based on their kinetics of expression. For each gene inside a cluster, the log_2FC values relative to SB-431542 at each time point has been plotted. Untr, untreated.

It is important to note that the target gene expression profiles identified could also be the result of different mRNA stabilities rather than transcriptional effects per se. To address this, I treated P19 cells with the transcription inhibitor Actinomycin D, and measured by qPCR the mRNA half-life for a number of representative NODAL/Activin targets belonging to different RNA-seq categories (Koba and Konopa, 2005). Interestingly, for the small number of genes analysed, I observed no relationship between message stability and the gene category (Figure 3.6). For example, *T* and *Tdgf1* were both very stable transcripts, yet *T* is a delayed gene and *Tdgf1* is in the 'induced sustained' category. Similarly, *Lefty1* and *Hes1* had relatively short half-lives, and are in the 'induced sustained' and 'transient induced' categories, respectively. On the other hand, genes which have similar expression kinetics showed different message stability, as is the case of *Lefty1* and *Tdgf1* from the induced sustained category or *T* and *Fst* from the 'delayed' category. Finally, genes which are repressed, either immediately or after a transient induction, such as *Smad7*, *Hes1* or *Id2*, all had unstable transcripts, as expected from the observation that their mRNA levels decrease within a few hours of Activin treatment (Figure 3.6). To sum up these results, it is obvious that mRNA stability has an impact on shaping the RNA-seq profiles, since it affects the quantity of transcripts that accumulate over time. However, from the limited number of representative genes considered in the Actynomicin D experiment it happears that the distinct kinetics profile identified might not be the result of mRNA stability. In particular, mRNA half-life does not seem to be responsible for the differences seen for example between 'induced sustained' and 'transient induced'. This last finding also suggests that in order to be expressed at later time points genes with short half-life as such as *Lefty1* would require on-going signalling.

Figure 3.6 on next page.

Figure 3.6. The mRNA stability of NODAL/Activin target genes.
P19 cells were treated with the RNA synthesis inhibitor Actinomycin D (6 μM) for the indicated times and samples were prepared for qPCR analysis. Plotted in green are the corresponding levels of mRNA for target genes representative of the four kinetics categories described in Figure 3.5. The half-life of each transcript was determined using a one-phase decay model, and the curve fitting the experimental data is displayed (grey line). Note that mRNAs with similar kinetics of induction have different half-lives. Shown are the means and SEM of three independent experiments performed in duplicate.

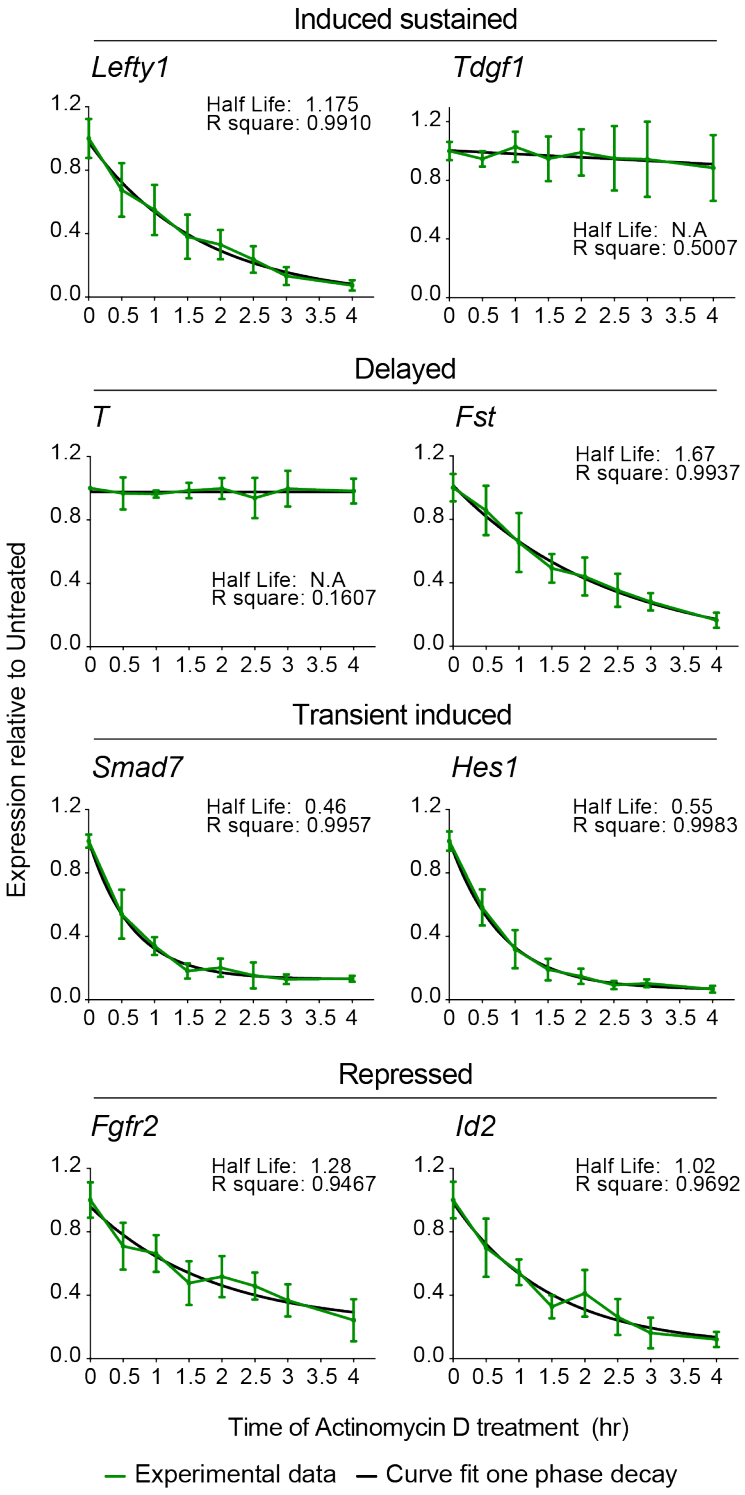
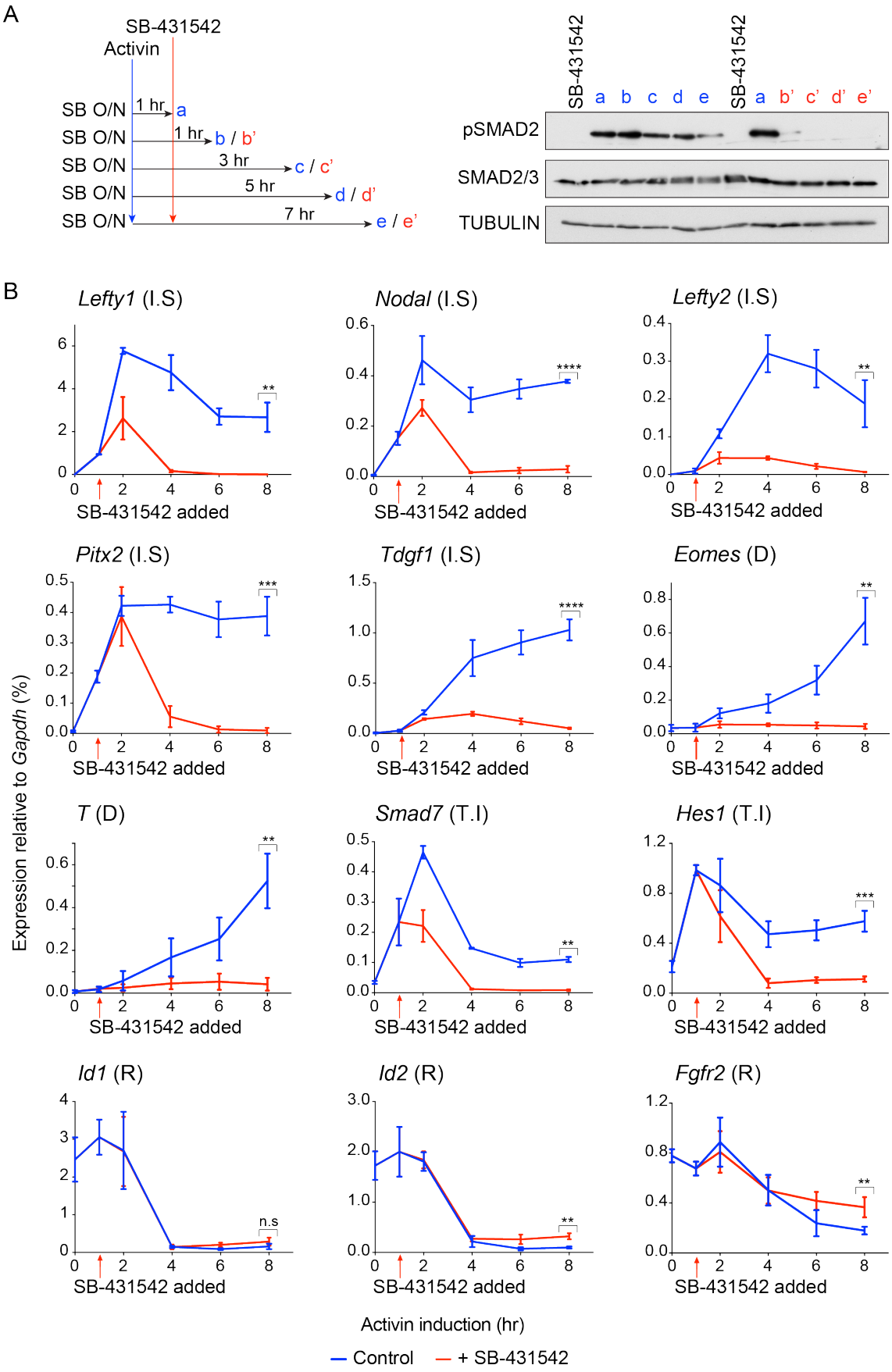


Figure 3.6. See previous page for legend.

Taking this question further, I decided to investigate to what extent the modulation of all long-term responses depends on continuous NODAL/Activin signalling. To this end, I acutely stimulated P19 cells with Activin and then blocked the pathway with SB-431542, measuring by qPCR the mRNA levels of a subset of Activin-regulated genes at different time points (Figure 3.7). As shown in Figure 3.7A, pSMAD2 was rapidly lost within 60' minutes after adding SB-431542 and it was still inhibited after 7 hr, whilst total SMAD2/3 levels did not change. The expression profiles of the representative SMAD target genes were significantly impaired when the pathway was blocked (Figure 3.7B). The induction of transcripts which in the control condition were upregulated after 1 hr of Activin treatment and kept being expressed at 8 hr, such as *Lefty1* and *Nodal,* was not maintained at later time points in the SB-431542 chase condition. Similarly, genes which were only induced by SMAD signalling at 8 hr such as *T* and *Trh* were no longer transcribed after the SB-431542 chase.

Conversely, the mRNA of genes which in the control were repressed at later time points, such as *Fgfr2*, was significantly higher in the SB-chase condition. This was less the case for the repressed genes *Id1* and *Id2*, whose expression profiles were largely unaffected by blocking the pathway after 1 hr of Activin treatment. Over an Activin time course, SMAD2 chromatin binding is detected around these genes following acute stimulation, but it is lost with prolonged signalling (data not shown). Thus, it is likely that *Id1* and *Id2* repression at later time points is independent of the SMAD pathway, which is required to mediate only the initial response to NODAL/Activin signalling. Finally, for transiently induced genes like *Smad7* and *Hes1* I found that when signalling was terminated after 1 hr their levels fell rapidly to baseline, whereas in the context of continuous signalling these genes were repressed in a more gradual fashion. Considering the results of the experiment altogether, it is clear that the SMAD pathway directly elicits the transcriptional responses downstream of NODAL/Activin signalling, both in acute and chronic treatment conditions.


A
SB-431542
Activin
SB O/N
1 hr
a
1 hr
b / b'
3 hr
c / c'
5 hr
d / d'
7 hr
e / e'
SB-431542
a b c d e
SB-431542
a b' c' d' e'
pSMAD2
SMAD2/3
TUBULIN
B
Lefty1 (I.S)
Nodal (I.S)
Lefty2 (I.S)
Pitx2 (I.S)
Tdgf1 (I.S)
Eomes (D)
T (D)
Smad7 (T.I)
Hes1 (T.I)
Id1 (R)
Id2 (R)
Fgfr2 (R)
SB-431542 added
Expression relative to Gapdh (%)
Activin induction (hr)
Control
+ SB-431542

This requirement for continuously-activated SMADs does not mean, however, that NODAL/Activin signalling alone is sufficient to modulate the long-term transcriptional effects. Indeed, when Tessa Gaarenstroom repeated the RNA-seq experiment introduced previously in the presence of protein synthesis inhibitors, the expression pattern of many target genes was significantly altered. This was particularly evident for many 'repressed' and 'delayed' genes (Coda et al., 2017). Here, I validated these results by qPCR for some representative targets of the pathway. The expression of the 'induced sustained' genes *Lefty1* and *Nodal* was not affected by the presence of Cycloheximide, confirming that they are directly regulated by NODAL/Activin signalling. Indeed, the transcriptional control of these genes at later time points did not require the new synthesis of any other components. In contrast, genes which in the control condition were expressed with delayed kinetics, such as *Trh* and *Eomes*, were no longer induced when protein synthesis was blocked. Similarly, genes whose mRNA levels dropped at later time points such as *Smad7* or *Fgfr2* were no longer repressed, suggesting the requirement for other factors to be synthesised in order to shape the transcriptional program downstream of NODAL/Activin (Figure 3.8).

Overall, the results presented in this section indicate that NODAL/Activin signalling directly induces/represses transcription and it also plays an instructive role in determining the transcriptional responses observed at later time points.

Figure 3.7 on previous page.
Figure 3.7. The transcripitional responses downstream of NODAL/Activin require on-going signalling.
(A) Schematic of the experimental set-up is shown (left panel): cells were treated overnight with SB-431542, washed out, then either treated with Activin for the indicated times (blue letters), or SB-431542 was added to the Activin-containing media after 1 hr of Activin treatment (red letters). Levels of pSMAD2, SMAD2/3 and TUBULIN as a loading control were assayed by Western blot (right panel). **(B)** Samples from cells treated as in (A) were prepared for qPCR and the mRNA quantities of target genes representative of the four distinct transcriptional profiles were measured. SB, SB-431542. Plotted are the means and SEM of three independent experiments performed in duplicate. **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01; n.s., not significant.

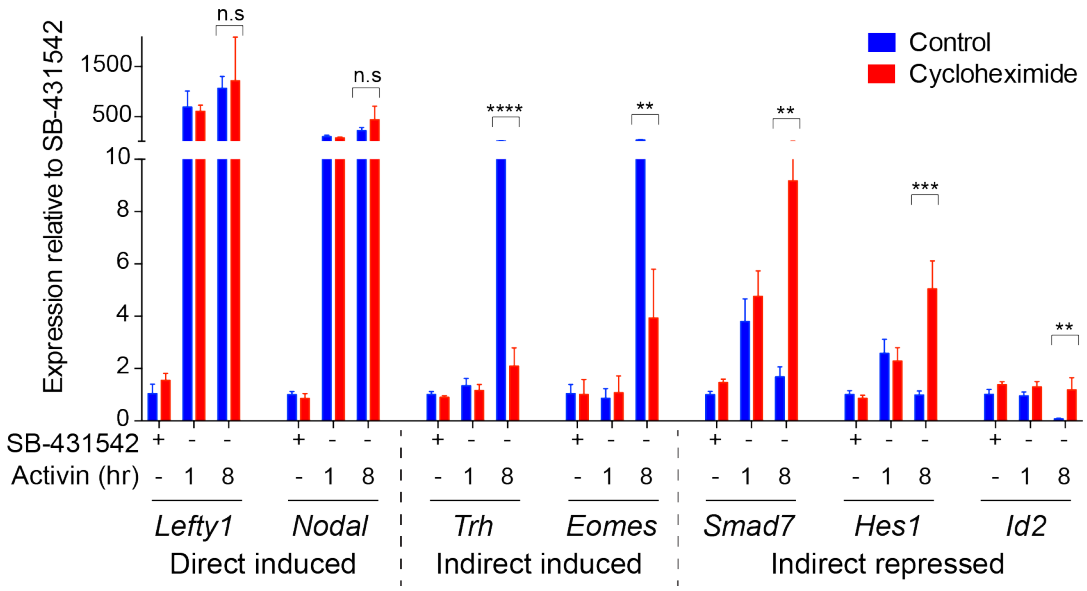


Figure 3.8. Protein synthesis inhibition alters the expression profiles of many NODAL/Activin target genes.
An Activin time course was performed on P19 cells with or without of the protein synthesis inhibitor Cycloheximide. In the control conditions, cells were treated overnight with SB-431542, washed out, then either incubated with Activin for 1 hr and 8 hr or SB-431542 was replaced. The same treatments were also carried out in the presence of Cycloheximide, which was added to the media at the same time as Activin or SB-431542. Levels of induction or repression of target genes representative of the different kinetics profiles were assayed by qPCR. Shown are the means and SEM of two independent experiments performed in duplicate. n.s., not significant. **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01.

3.2.3 Relating SMAD2 chromatin binding to gene expression

To understand how NODAL/Activin pathway activation elicits such a complex programme of gene expression we carried out a series of ChIP-seq experiments at the same time points as were used for RNA-seq. By performing ChIP-seq for SMAD2 alongside ChIP-seq for different forms of Pol II, we sought to identify SMAD2 binding sites (SBSs) and Activin-regulated genes on a genome-wide level, and also to investigate the mechanisms whereby SMAD complexes regulate Pol II activity. For the library preparation and sequencing we took advantage of the Francis Crick Institute Advance Sequencing Facility, whilst for the bioinformatics analyses downstream we collaborated with Philip East and Harshil Patel from the Francis Crick Institute Bioinformatics and Biostatistics Service (BABS).

First, to define SMAD2 binding genome-wide in the four different conditions, Tessa Gaarenstroom performed ChIP-seq for SMAD2 and used MACS (Zhang et al., 2008) to call peaks of SMAD2 enrichment compared to input chromatin. She observed no SMAD2 peaks in the SB-431542 condition, with SMAD2 binding to chromatin being induced by NODAL/Activin signalling in all the other states. Interestingly, SMAD2 binding also appeared to be dynamic, with only a proportion of sites being constantly occupied among the different NODAL/Activin treatment conditions (Coda et al., 2017).

Secondly, I performed ChIP-seq (biological replicates) for two different forms of Pol II: Serine-5-phoshorylated Pol II CTD (Ser5P), which is characteristic of paused Pol II; and Serine-2-phoshorylated Pol II CTD (Ser2P), which is associated with elongating Pol II (Levine, 2011). The experiments were carried out for the four signalling conditions introduced above, and an additional time point, 4 hr after Activin treatment, was also included to better characterise the responses to an intermediate level of signalling. The ChIP-seq results were first validated by qPCR at *Lefty1* transcription start site (TSS) and transcription termination site (TTS). As expected, Ser5P was enriched at the TSS and TTS of *Lefty1*, but not in the gene-desert region used as a negative control. Ser2P binding was also detected at the TSS and TTS of *Lefty1*, with a bigger enrichment at the TTS then at the TSS (Figure 3.9), as typically observed for Pol II Ser2P (Descostes et al., 2014). It is worth noting here that the use of these two markers provided me with the opportunity to accomplish two goals. The first goal was to investigate how SMAD signalling mechanistically regulates Pol II by looking at the differences in signal between Ser5P and Ser2P in the same or across treatment conditions (see Chapter 4). The second goal was to pinpoint genes which exhibited differential Ser5P/Ser2P enrichments in the stimulated states compared to the SB-431542 state. Combining these data with the SMAD2 ChIP-seq data and the RNA-seq data would allow us to better identify the genes directly regulated by the pathway.

The IGV display of the *Lefty1/Lefty2* genomic locus in Figure 3.10 provides an example of the data obtained by combining the alignment tracks from the SMAD2 ChIP-seq and the Pol II ChIP-seq experiments. *Lefty1* and *Lefty2* are known Activin target genes, and indeed SMAD2 chromatin binding was detected around these genes in response to NODAL/Activin signalling. Ser5P enrichment across the genes bodies was also induced in a signalling dependent manner, confirming that these

genes were not expressed in the SB-431542 state. In summary, the results obtained at *Lefty1/Lefty2* genomic locus worked as a positive control to validate the quality of the ChIP-seq datasets, and gave us confidence in using the data for downstream analysis.

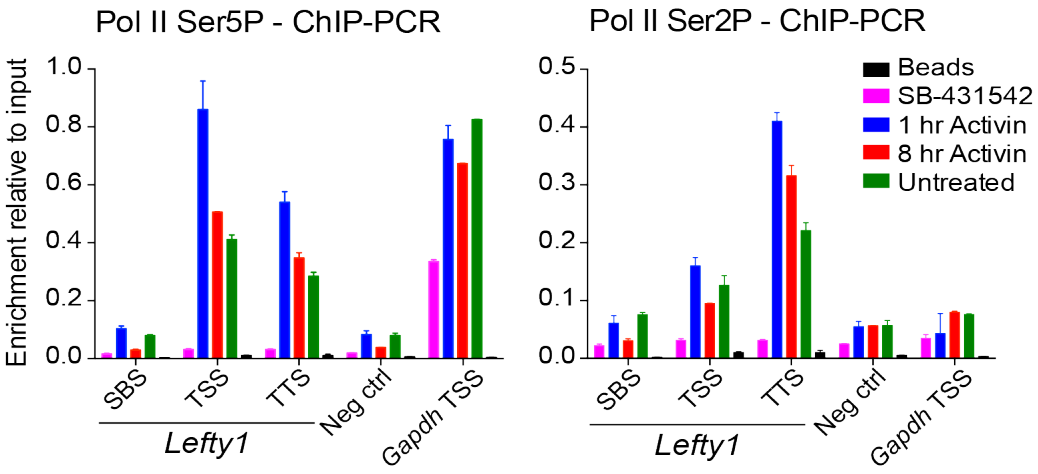


Figure 3.9. Pol II Ser5P and Pol II Ser2P bind to the *Lefty1* gene in response to NODAL/Activin signalling.
ChIP qPCR for Pol II Ser5P or Ser2P over the *Lefty1* region in P19 cells treated with SB-431542 or with Activin for 1 hr and 8 hr, or left untreated as previously described (see Figure 3.2, arrows). As negative control for the IPs, chromatin from each sample has been pooled together and incubated with beads alone (Beads). Plotted are the enrichment for Ser5P or Ser2P at the SBS, TSS and TTS sites relative to a negative control (Neg ctrl) region and the *Gapdh* TSS. A representative experiment is shown (mean ± SD). Note that this experiment provides a validation for the ChIP-seq datasets obtained with the same antibodies in the same experimental setting (see Figure 3.10).

Figure 3.10 also illustrates well the challenge faced in connecting the SMAD2 peaks to Activin-regulated genes. In fact, considering the screenshot of the *Lefty1/Lefty2* locus, it is not clear which peak drives the expression of which gene. Since it is known that enhancers can interact with their target regions over distances in the range of hundreds of kb (Long et al., 2016), it could be possible that *Lefty1* expression is not regulated by the closest SMAD2 enhancer peak (10 kb away), but instead by the enhancer peak upstream of *Lefty2* (around 50 kb away), or by both enhancers together. To test this hypothesis, I decided to take advantage of the CRISPR/Cas9 technology to delete, separately, the two major SMAD2 binding sites (SBSs) upstream of *Lefty1* and *Lefty2*.

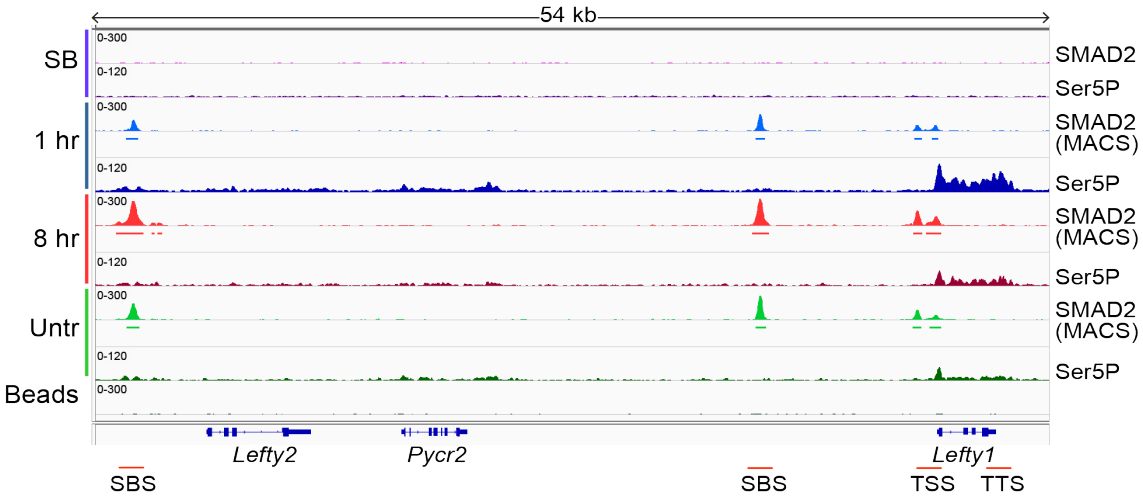
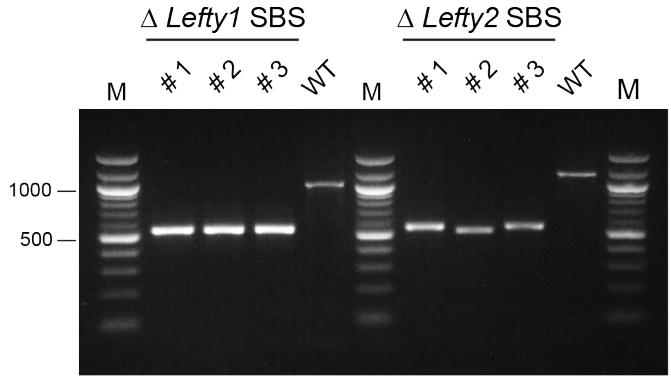


Figure 3.10. NODAL/Activin signalling induces changes in SMAD2 and Pol II binding genome-wide.
IGV browser visualisation over the *Lefty2/Lefty1* genomic locus of ChIP-seq data for SMAD2 and Pol II Ser5P (Ser5P) in P19 cells treated as described in Figure 3.9. For the SMAD2 ChIP-seq the MACS-called peaks are also shown. For future reference, regions for which ChIP-PCR primers were designed are highlighted in red below the figure. SB, SB-431542; 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

In both cases, I was able to obtain three individual clones which lacked a 500-bp region encompassing the respective SBSs on both alleles, as demonstrated by PCR using primers flanking the CRISPR guides and subsequently confirmed by sequencing (Figure 3.11A and B). I then performed in these cells a time course of Activin treatment and measured by qPCR the induction of target genes. Interestingly, *Lefty1* induction was severely impaired in the clones lacking *Lefty1* SBS, whilst *Lefty2* expression was not affected (Figure 3.12A). Similarly, when *Lefty2* SBS was deleted, *Lefty2* induction was lost, but *Lefty1* was unchanged compared to the control (Figure 3.12B). The induction of the others Activin-target genes I looked at as a further control were also not affected in both scenarios (Figure 3.12C, D). I therefore concluded that, at least in this case, there is a one-to-one relationship between the SBSs and the genes they regulate. Taking on board this result, we decided to associated the SMAD2 peaks with the closest Activin-regulated gene using a cut-off of 100 kb from an annotated TSS or TTS. The dataset of Activin-regulated genes was defined as described below.

A



B

Δ *Lefty1* SBS, DNA deletions

# 1	chr1: 180924718 - 180925179; chr1: 180924715 - 180925181;
# 2	chr1: 180924718 - 180925179; chr1: 180924716 - 180925177;
# 3	chr1: 180924718 - 180925179; chr1: 180924716 - 180925177;

Δ *Lefty2* SBS, DNA deletions

# 1	chr1: 180888621 - 180889200;
# 2	chr1: 180888590 - 180889200;
# 3	chr1: 180888621 - 180889200;

Figure 3.11. Derivation of P19 lines with deletions of the *Lefty1* or the *Lefty2* upstream SBS using CRISPR/Cas9 technology.
For the identification of the deletions generated by each pair of guide RNAs targeting either the *Lefty1* or the *Lefty2* upstream SBS, PCR on genomic DNA was performed using primers upstream and downstream of the selected regions. The agarose gel in **(A)** shows the PCR results for three individual clones with deletions of *Lefty1* SBS or *Lefty2* SBS on both alleles. The same PCRs were performed on wild type genomic DNA as a control. M, 100 bp DNA ladder. **(B)** To further characterise the deletions, the PCR products described in **(A)** were cloned into pGEM-T vector and sequenced. In the case of *Lefty1*, all the clones are compound heterozygotes. Note that one deletion is recurrent in all the clones, but the sequence of the amplified fragment varied across the individual clones. In the case of *Lefty2*, each clone has the same deletion on both alleles, but the flanking sequences are different.

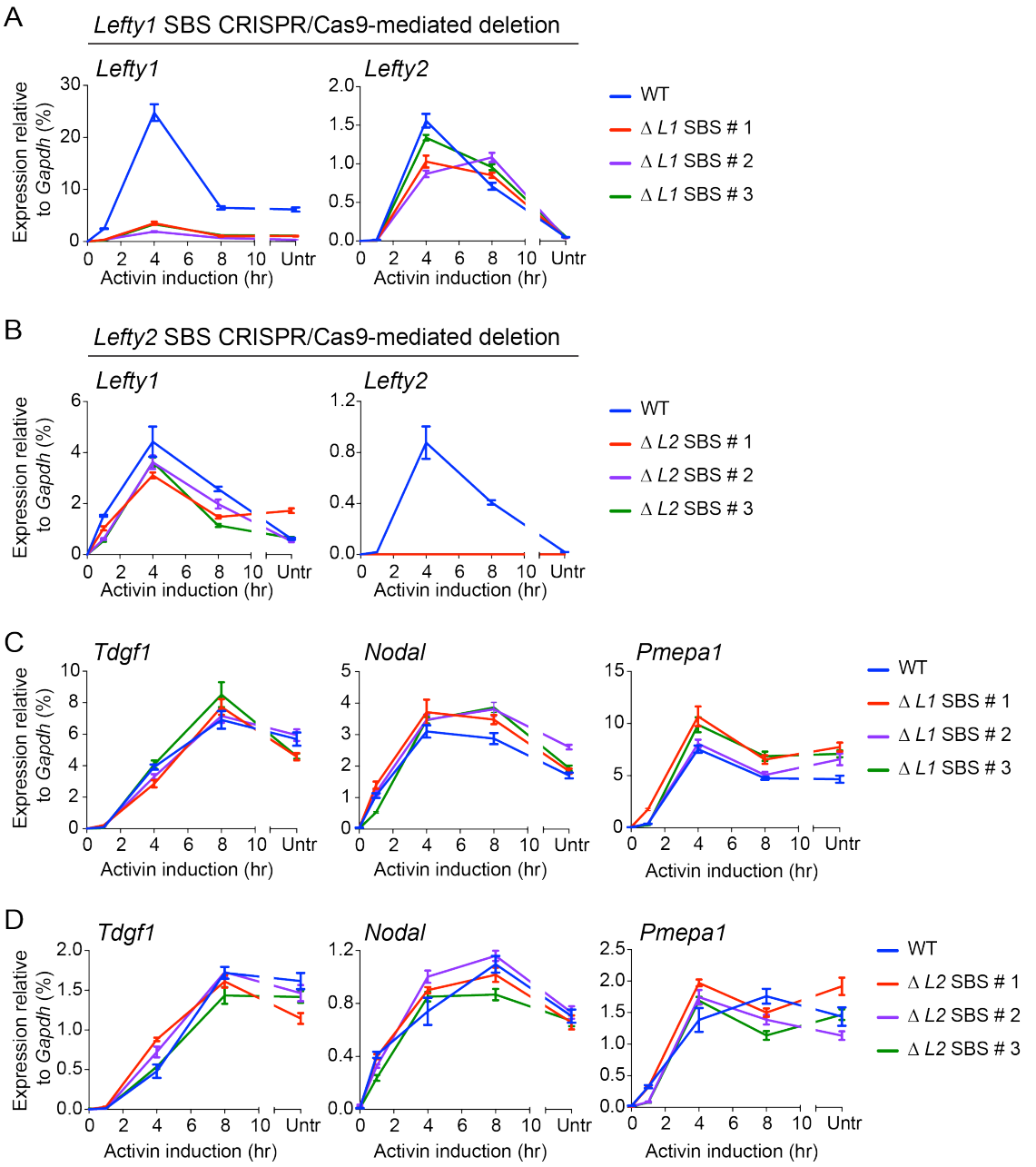


Figure 3.12. Effects of deleting the *Lefty1* or *Lefty2* upstream SBSs on Activin-mediated induction of *Lefty1* and *Lefty2*.
(A and **B)** A time course of Activin induction was performed for the indicated times on the three independent clones with deletions for the *Lefty1* (*L1*) upstream SBS described in Figure 3.10, and on WT P19 cells as a control. Transcript levels of *Lefty1* and *Lefty2* were measured by qPCR relative to *Gapdh*. Compared to the control, *Lefty1* induction is impaired in all the three clones, whilst *Lefty2* induction is not affected (A). The same experiment was performed on the the three independent clones with deletions for the *Lefty2* (*L2*) upstream SBS described in Figure 3.10. qPCR results show that *Lefty2* induction is impaired in all the three clones, whilst *Lefty1* induction is not affected (B). **(C** and **D)** The transcriptional profiles of other Activin-regulated genes are not affected in any of the clones with deletions for *Lefty1* (*L1*) upstream SBS (C) or with deletion for *Lefty2* (*L2*) upstream SBS (D).

3.2.4 Defining a high confidence dataset of Activin-regulated target genes and associated SMAD2 peaks

The RNA-seq experiment had initially identified 747 genes differentially regulated in response to NODAL/Activin signalling with respect to the SB-431542 state. Importantly, this list was potentially confounded by the fact that the amount of mRNA measured for any given transcript does not directly reflect the transcriptional activity observed at the gene level. To avoid inclusion of noise, we decided to integrate the RNA-seq data with the ChIP-seq for SMAD2 and Pol II Ser5P/Ser2P. With this approach, we aimed to define a high confidence set of Activin-regulated genes and associated SMAD2 peaks to use for downstream analyses.

As a first step, we determined sites of differential enrichment of Pol II Ser5P/Ser2P at each time point relative to the SB-431542 condition using the DiffReps package (Shen et al., 2013). This analysis identified 410 genes which clearly showed changes in Pol II Ser5P/Ser2P enrichment in at least one signalling condition compared to the SB-431542 state (see Section 2.4.4 for further details). We then intersected this list with the RNA-seq and the SMAD2 ChIP-seq data, and took forward only those genes that had a SMAD2 peak within 100 kb, retrieving a set of 99 genes. Finally, we manually curated the list obtained to include genes that had surrounding SMAD2 peaks and whose induction/repression in response to NODAL/Activin signalling had been previously verified by qPCR. Genes that were found to have differential enrichment of Pol II Ser5P/Ser2P, a SMAD2 peak within 100 kb and a $\log_2$FC by RNA-seq greater than 0.5 in at least one signalling condition were also added to the final list.

In conclusion, this analysis allowed us to define a high confidence set of Activin-regulated target genes composed of 140 transcriptional targets associated to 478 SMAD2 peaks (for the full list, see Appendix 8.1, 8.2). We then divided the 140 genes accordingly to the four gene expression categories described in Section 3.2.2, identifying 28 ‘Induced sustained’, 50 ‘Delayed’, 10 ‘Transient induced’ and 52 ‘Repressed’ genes (Figure 3.13). The 140 Activin-regulated target genes were also separated into ‘baseline on’ (111 genes) or ‘baseline off’ (29 genes) on the basis of their level of expression in the SB-431542 state (see Section 2.4.1 for details). Included in the first group are all the genes which were already transcribed in the

absence of signalling, whilst the second group contains those which were activated from an undetectable baseline in response to NODAL/Activin (Figure 3.13).

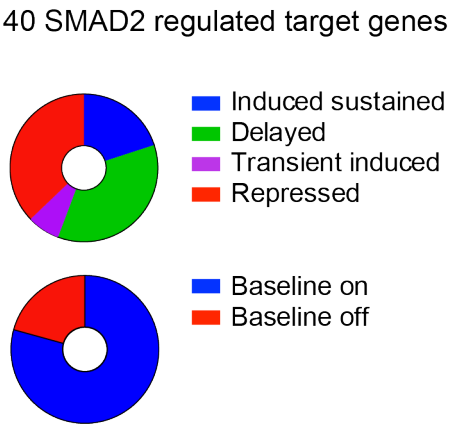


Figure 3.13. Characterisation of the high confidence dataset of SMAD2-regulated target genes.
The 140 SMAD2-regulated target genes (see main text for dataset construction criteria) divided accordingly to the four gene expression categories (top plot), or to the level of expression in the SB-431542 state (bottom plot).

3.3 Discussion

3.3.1 Summary of main findings

- In P19 cells, over an Activin time course from a SB-431542 signalling-inhibited baseline, pSMAD2 levels peak at ~ 60 minutes, and then attenuate to lower levels at prolonged signalling states. In these cells, a basal level of SMAD2 phosphorylation due to autocrine production of NODAL and GDF3 is also observed.
- Four distinct categories of genes can be identified based on their kinetics of expression in response to NODAL/Activin signalling by RNA-seq. These profiles are not a result of different mRNA half-lives.
- The correct transcriptional responses downstream of NODAL/Activin require on-going signalling, and, in many cases, also new protein synthesis.
- ChIP-seq for SMAD2 identifies its chromatin binding events genome-wide in response to NODAL/Activin over time.
- CRISPR/Cas9-mediated deletion of *Lefty1* and *Lefty2* upstream SBSs serves a proof of principle for the criteria used to associate SMAD2 peaks to target genes.
- Integrating RNA-seq with ChIP-seq for SMAD2 and Pol II Ser5P/Ser2P defines a high confidence SMAD2 target gene set made of 140 genes associated with 478 SMAD2 peaks. Out of the 140 genes, 29 are induced from RNA baseline off.

3.3.2 NODAL/Activin signalling in P19 as a model system to study complex programmes of gene expression

In this chapter I first characterised in P19 cells the dynamics of SMAD pathway activation in response to NODAL/Activin signalling. Moreover, I introduced the genome-wide datasets I will use in the rest of the thesis to delineate the sequence of events from SMAD2 chromatin binding to transcription regulation over time. Here, I will discuss the reasons why the use of P19 as a model system, coupled with time courses of ligand induction, can provide a framework not just to better understand the biological role of NODAL/Activin signalling, but also to unravel how extracellular signals in general control gene expression over time.

Over the last several years, a number of groups have used Next Generation Sequencing (NGS) technologies to identify SMADs binding sites and target genes in

different cellular contexts (Beyer et al., 2013, Brown et al., 2011, Estaras et al., 2015, Kim et al., 2011, Mullen et al., 2011). In one of these works, the authors only focused on autocrine chronic signalling, which without the comparison to signal-inhibited conditions is not informative enough to establish a causal link between SMAD chromatin binding and regulation of gene expression (Mullen et al., 2011). In other studies, inductions with ligands were performed on a time scale of days over complex differentiation processes, as in the case of human and mouse ESCs (Brown et al., 2011, Estaras et al., 2015, Kim et al., 2011). In these experiments, it is difficult to untangle to what extent the observed transcriptional responses are directly downstream of NODAL/Activin signalling, or are the result of the activity of TFs induced by the pathway itself. The advantage of the P19 cells is that time courses of Activin treatment can be performed from a signalling-inhibited baseline over minutes or hours without affecting the cells' phenotype. P19s have also an autocrine production of NODAL, which enabled us to study the pathway in chronic signalling conditions. In addition, the transcriptional responses observed in P19 cells downstream of NODAL/Activin recapitulate the biological complexity of the signalling pathway. In fact, I showed that classic target genes such as *Lefty1*, *Pmepa1* and *Smad7* are induced in response to Activin alongside with markers of mesendoderm differentiation such as *T* and *Eomes* or pluripotency genes such as *Nanog* and *Pou5f1*.

Another example of the biological relevance of using P19 as a model system come from the observation that the transcriptional outputs seem to be more a function of time of exposure to the ligand then ligand dose itself. Indeed, the duration of NODAL signalling, rather than its amplitude, has been shown to control cell fate specification in zebrafish embryos (Hagos and Dougan, 2007). In this context, cells that have been exposed to NODAL signalling for longer times give rise to endoderm, whilst cells which see the ligand for shorter times specify mesoderm ((van Boxtel et al., 2015); van Boxtel et al., in revision)). My experiments also demonstrate that dynamics of pathway activation as monitored by pSMAD2 levels do not predict the transcriptional activity downstream of NODAL/Activin signalling. Whilst SMAD2 phosphorylation peaks approximately 1hr after Activin stimulation, many genes are in fact only induced at later time points, as observed for *Eomes* or *T*. In these cases, the delay is possibly due to the fact that the target gene expression likely requires the presence of factors which are synthesised in response to NODAL/Activin

signalling itself. Nevertheless, the analysis of pSMAD2 patterns over time can still provide interesting insight into the biology of NODAL/Activin.

Indeed, the signalling dynamics of these ligands are substantially different to those observed in response to canonical TGF-β. First, when considering a time course of TGF-β treatment, pSMAD2 induction rapidly decreases to very low level, in contrast to what is observed with NODAL/Activin, where the signal only modestly attenuates in prolonged and chronic conditions (Vizan et al., 2013). Secondly, TGF-β-mediated SMAD2 phosphorylation is entirely induced within 5 minutes of signalling, and a prolonged exposure to the ligand does not result in higher pSMAD2 levels (Vizan et al., 2013). This is not the case for NODAL/Activin, as SMAD2 phosphorylation progressively increases over the first hour of signalling, suggesting that the duration of exposure to NODAL/Activin determines the pSMAD2 response (Miller et al., manuscript in preparation). Third, TGF-β exposure triggers the rapid depletion of receptors from the cell surface, which fully re-accumulate only 12 – 24 hr after the removal of the ligand. As a consequence, continuous TGF-β signalling results in desensitised cells that are refractory to further acute stimulation (Vizan et al., 2013). Such a *phenomenon* is not observed in response to prolonged Activin treatment, and cells can be efficiently stimulated in acute manner when the ligand is re-applied (Miller et al., manuscript in preparation).

These differences are likely to be due to distinct receptors dynamics, since Activin receptors do not depleted from the cell surface over time in contrast to TGF-β receptors ((Vizan et al., 2013); (Miller et al., manuscript in preparation)). The fact that P19s secrete NODAL and GDF3 in response to pathway activation could also provide a possible explanation for the induction of SMAD2 phosphorylation with prolonged and chronic signalling. Nevertheless, what determines the attenuation in pSMAD2 level observed at later times in response to Activin, it is still unclear. Negative feedback regulators of the pathway induced in response to the signalling such as SMAD7 or receptors recycling dynamics are plausible mechanisms which could be interesting to investigate. The distinct characteristics of SMAD2 pathway activation in response to NODAL/Activin or TGF-β may also reflect the different roles played by these ligands *in vivo*. During embryonic development, NODAL is known to act as morphogen, whilst TGF-β has never been shown to form gradients (Schier, 2009). It can be speculated that this is due to the fact that only in the context of

NODAL/Activin, but not TGF-β, cells are able to monitor the duration of exposure to ligands and adapt their programmes of gene expression accordingly. For example, during zebrafish development, long term NODAL signalling induces a greater pSMAD2 response and give rise to endoderm ((van Boxtel et al., 2015); van Boxtel et al., in revision)). In contrast, TGF-β is likely to work *in vivo* for short period of time, by being released from latency complexes and activated locally in a precise spatiotemporal manner (Tada et al., 2001, Robertson and Rifkin, 2016).

The RNA-seq experiment performed in P19 cells revealed that a complex programme of gene expression is elicited downstream of NODAL/Activin. Importantly, I showed that the distinct kinetic profiles identified are not the consequences of different mRNA half-lives. Moreover, I demonstrated that not just acute, but also long term transcriptional responses require on-going NODAL/Activin signalling, indicating that the factors which modulate gene expression at later time points likely interact with chromatin-bound SMAD2 (For further discussion, see Section 4.2.3). These results, coupled with the observation that many genes need protein synthesis for their correct induction/repression, strongly suggest that NODAL/Activin signalling directly orchestrates a transcriptional network that regulate gene expression over time.

In conclusion, I have provided here evidence supporting the use of P19 cells as an ideal model system to study the NODAL/Activin–SMAD pathway. The data generated in P19s can be employed not just to investigate how transcription is mechanistically regulated in response to extracellular signals, but also as a valid resource to dissect the biological complexity of NODAL/Activin signalling.

3.3.3 Defining a high confidence set of SMAD2 target genes and associated SMAD2 peaks: challenges and alternative approaches

When considering the genome-wide datasets introduced in this chapter with the results obtained by other studies investigating NODAL/Activin signalling in different systems, differences in terms of ‘sizes’ can be noted. The numbers of target genes and SMAD2 peaks previously reported are in fact one or two orders magnitude bigger than what is seen in P19s (Brown et al., 2011, Estaras et al., 2015, Kim et al., 2011, Mullen et al., 2011). This is particularly evident when considering that the high confidence set of SMAD2 target genes we obtained contains just 140 genes.

Technical reasons, the use of different model systems and analysis methods could easily explain the discrepancies. Moreover, the decision to integrate RNA-seq data with the ChIP-seq data also impacted on the size of the final dataset. However, we deliberately chose to work with stringent cut-offs to get a high confidence dataset, as opposed to take forward a larger number of SMAD2 peaks and target genes, which would have resulted in a bigger but less reliable dataset. The RNA-seq alone could have been potentially misleading to identify robust Activin-regulated genes, since changes measured over time do not necessarily reflect the transcriptional activity going on at the gene level. If a transcript is very stable, for instance, a modest increase in gene expression can lead to a great mRNA accumulation over time. Refining the RNA-seq list in light of the Pol II ChIP-seq results allowed us to exclude this scenario, and it ensured that only the genes showing robust changes in Pol II Ser5P/Ser2P in response to NODAL/Activin signalling were included in the final dataset. The fact that only the 20 percent of the genes identified as differentially expressed by RNA-seq made the refined dataset could also reflect the differences in sensitivity between the two techniques used, with ChIP-seq for Pol II Ser5P/Ser2P being the limiting one. Nevertheless, the high confidence dataset maintains the complexity of the transcriptional responses identified by RNA-seq, with all the four gene categories been represented. As a result, the high confidence dataset provides a unique opportunity to study how NODAL/Activin signalling regulates Pol II with respect to the different groups of genes, which will be the subject of the next chapter.

Another issue commonly faced in genome-wide studies concerns the criteria used to assign the TF peaks detected by ChIP-seq to the genes they regulate. In the absence of further data, this task can become challenging especially if the binding events occurs in intergenic regions far from annotated promoters, as was the case for many SMAD2 peaks. It has been shown in fact that enhancers can regulate target genes over long distances, with three-dimensional higher order chromatin structures dictating enhancer-promoter associations via chromatin looping (Long et al., 2016). To account for this, we decided to associate each SMAD2 peak to the closest Activin-regulated gene within a distance of 100 kb, greater than what normally used in other studies. An alternative approach would be to perform chromosome conformation capture experiments in P19 in the difference signalling conditions and comparing it to the SMAD2 ChIP-seq. Hi-C has successfully been used to predict *bona fide* enhancer-promoter pairs, although the resolution of this technique is still a matter of

debate (Schmitt et al., 2016). Moreover, since the majority of these long-range chromatin interactions are conserved across cell types, reliable results could also be obtained by simply overlapping the ChIP-seq data with maps of chromatin interactions obtained in other cell lines or in other experimental conditions, as it has been done in other studies (Gualdrini et al., 2016, Zanconato et al., 2015).

Chapter 4. NODAL/Activin signalling regulates Pol II via *de novo* recruitment

4.1 Introduction

The first step in understanding how NODAL/Activin orchestrates the complex transcriptional programme described in Chapter 3 is to define how SMAD signalling regulates Pol II. From a mechanistic point of view, distinct scenarios are possible ((Levine, 2011); Figure 4.1)). Typically, gene activation can result from the release of 'paused' Pol II from promoter-proximal locations or from Pol II *de novo* recruitment to the DNA template. In the first case, Pol II is stalled downstream of the transcription start site of inactive genes after producing a short nascent transcript, as a result of a yet not fully characterized mechanisms involving different protein complexes and DNA sequences (Liu et al., 2015, Levine, 2011, Fuda and Lis, 2013, Adelman and Lis, 2012). This form of Pol II is generally characterised by the phosphorylation of Serine 5 (Ser5P) at the carboxyl terminal domain (CTD) of its largest subunit. For Pol II to escape into the gene body, the positive transcription elongation factor b kinase (PTEF-b kinase) needs to be recruited by transcription factors and to phosphorylate the pausing proteins and the Pol II CTD at Serine 2 (Ser2P) (Adelman and Lis, 2012). The resulting actively transcribing Pol II is therefore identified by the presence of both Ser5P and Ser2P on its CTD (Figure 4.1).

Since Pol II pause-release is often employed to rapidly switch on genes in response to external stimuli like heat shock or growth factors, this mechanism for has been commonly associated with signal-induced transcription (Liu et al., 2015). An alternative to the pause-release mechanism is transcription control via Pol II recruitment. In this case, Pol II does not bind to inactive genes prior to expression, but it is recruited *de novo* to promoters and rapidly converted from initiating Pol II (Ser5P) into elongating Pol II (Ser5P and Ser2P) to fully execute its transcriptional activity ((Fuda and Lis, 2013); Figure 4.1). With respect to NODAL/Activin signalling, nothing is currently known about how the signalling pathway regulates Pol II activity to control the expression of the different target genes.

Here, I answer this question using the data obtained from the ChIP-seq experiments for Pol II Ser5P and Pol II Ser2P introduced in the previous chapter.

Once I had defined Pol II binding dynamics over time, I then set out to explore the relationship between Pol II occupancy at target genes and the amount of transcript as measured by mRNA-seq. Finally, I describe how SMAD2 binding to chromatin correlates with transcription over time.

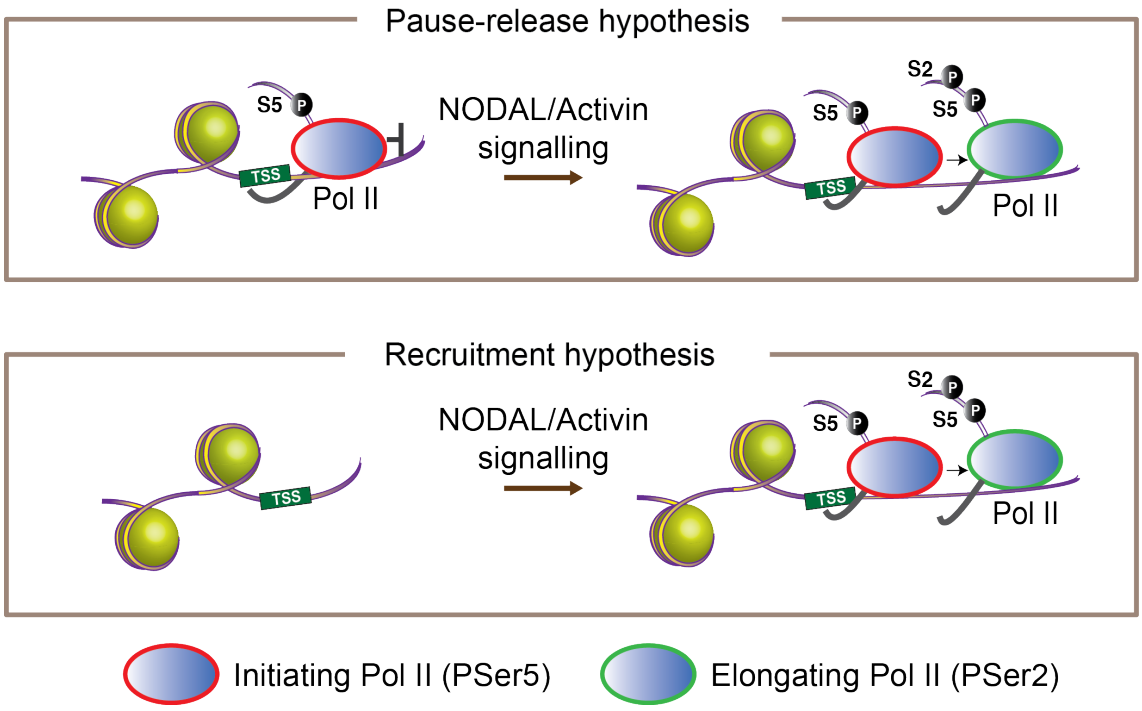


Figure 4.1. Hypothetical modes of regulating Pol II by NODAL/Activin signalling. A cartoon to illustrate how NODAL/Activin signalling could regulate Pol II transcriptional activity in the context of a model of Pol II pause-release (top panel) or Pol II recruitment (bottom panel).

4.2 Results

4.2.1 NODAL/Activin signalling regulates Pol II via *de novo* recruitment

As mentioned in Chapter 3, to understand directly how SMAD signalling regulate transcription I performed ChIP-seq experiments (biological duplicates) with antibodies against Pol II Ser5P and Ser2P in the same conditions used for the RNA-seq (see Section 3.2.2). An additional time point at 4 hr of Activin treatment was also included to more precisely characterise the sequence of events. The data obtained were then analysed by working together with Harshil Patel from BABS. First, to validate on a genome-wide scale the reliability of the datasets both qualitatively and quantitatively, metaprofiles of the averaged signals for Pol II Ser5P or Pol II Ser2P across all genes were generated for all time points. Here, only the 1hr Activin plots are presented as an example (Figure 4.2). As expected, the signals for the initiating Pol II (Ser5P) and the elongating Pol II (Ser2P) showed the characteristic patterns, with Ser5P peaking at the TSS and Ser2P at the TTS (Descostes et al., 2014). The metaprofiles also confirmed the robustness of the method used to normalise the data (see Section 2.4.4 for details), since the two replicates appeared indeed very similar to each other in both the ChIP-seq experiments (Figure 4.2).

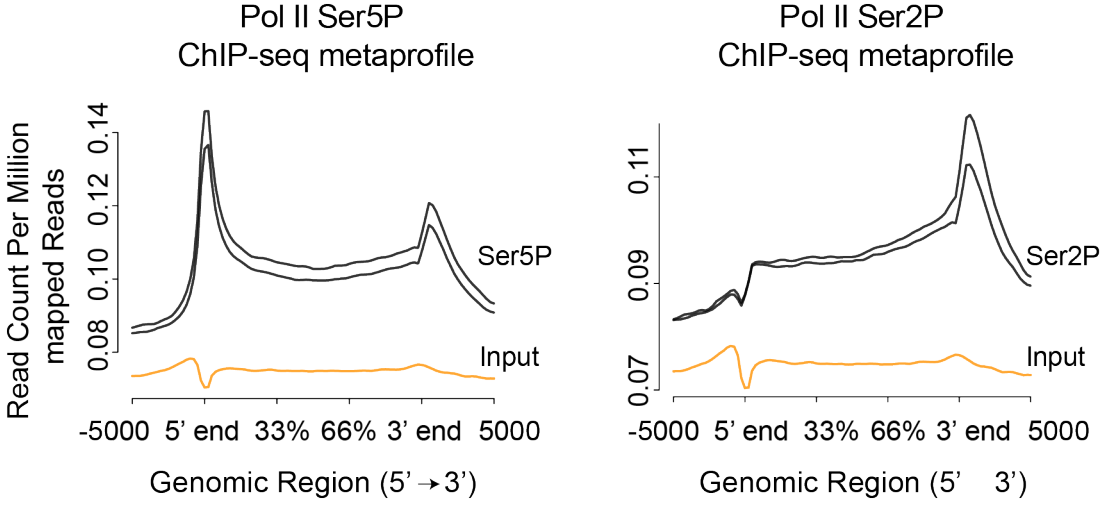


Figure 4.2. Pol II Ser5P and Pol II Ser2P have distinct profiles genome-wide. Metaprofiles of mean coverage across all genes for Pol II Ser5P or Pol II Ser2P ChIP-seq. Displayed are the normalised read counts for two ChIP-seq replicates from the 1 hr Activin sample (black lines) and the corresponding inputs (orange lines).

In order to understand how Activin signalling regulates Pol II activity, I then focused on the SMAD2-regulated genes which showed differential Pol II Ser5P or Pol II Ser2P binding and plotted the log_2 Fold Change (log_2FC) for the two marks relative to the SB-431542 condition over time, alongside the RNA-seq data (Figure 4.3A). It appeared that Activin induced changes in Ser5P and Ser2P occupancies with dynamics that were overall consistent with those observed for the RNA seq data. In fact, for the large majority of genes, the signal for both isoforms of Pol II tended to peak at 1hr and then to attenuate over time, in line with the idea of sustained signalling resulting in either repression or dampening of transcription discussed in Chapter 3. Interestingly, when comparing the binding patterns of the two marks it was also evident that Pol II Ser2P generally mirrored Pol II Ser5P behaviour over time. The plots in Figure 4.3B further confirmed that Ser5P and Ser2P values correlated very well at all the time points. Overall, this last result suggested a mechanism of Pol II recruitment dependent on Activin signalling, rather than a pause-release mechanism. I reasoned that if the SMADs induced transcription by releasing paused Pol II, I would expect only the signal for Pol II Ser2P to be differentially enriched compared to SB-431542, but not the signal for Pol II Ser5P. Clearly, this is at odds with what I observed.

To better dissect this aspect, I decided to analyse the occupancy of the two isoforms of Pol II over time separately for each of the target genes groups previously identified, focusing first on ‘baseline off’ and ‘baseline on’ genes (see Figure 3.13). By definition, ‘baseline off’ genes were not transcribed in the absence of NODAL/Activin, making them the ideal candidates to look at for the presence of paused Pol II. Since signals detected along the gene bodies were generally low and the differential between the two marks was small, Harshil Patel generated metaprofiles for Pol II Ser5P and Pol II Ser2P centred around the TSSs and the TTSs of target genes, where the differentials were much bigger. Strikingly, when considering the ‘baseline off’ genes, Pol II Ser5P was not enriched at the TSSs in the SB-431542-treated state, but was readily detected in all other conditions, thus excluding the presence of paused Pol II. In line with this finding, the same temporal patterns were also observed for Pol II Ser2P at the TSS and at the TTS (Figure 4.4).

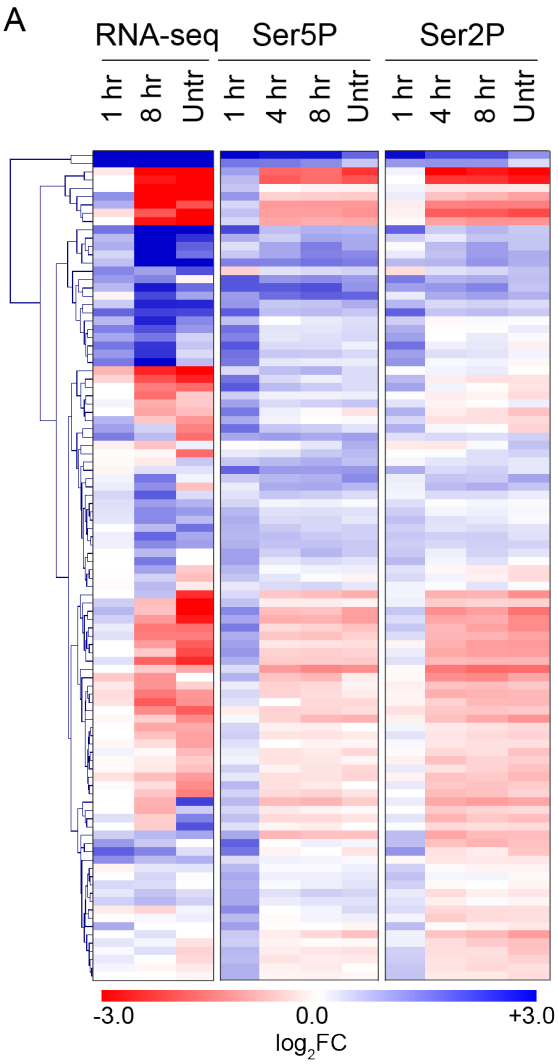
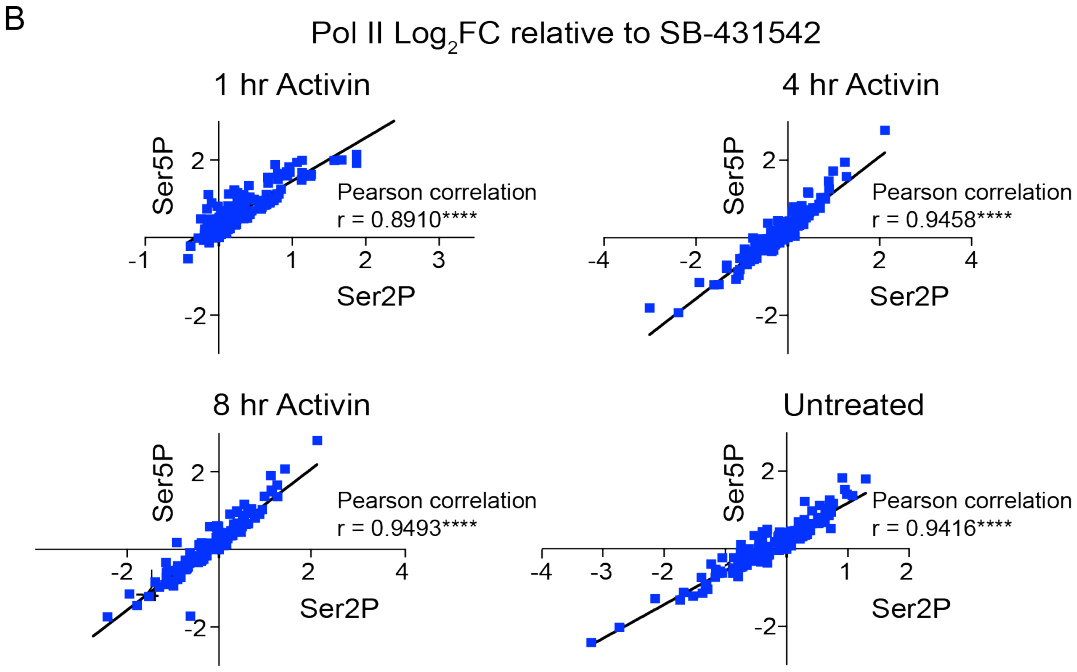


Figure 4.3. Pol II Ser5P and Pol II Ser2P binding dynamics follow similar temporal patterns.
(A) For the genes in the high confidence dataset the log_2FC relative to SB-431542 of mean normalised read depth for Ser5P or Ser2P isoforms of Pol II were calculated. The data were displayed as a heatmap alongside the log_2FC relative to SB-431542 for the RNA-seq and hierarchically clustered. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated. **(B)** Correlation plots comparing at each time point the log_2FC for Pol II Ser5P and Pol II Ser2P obtained as described in **(A)**.



Pol II Ser5P and Pol II Ser2P were instead seen at the TSS and TTS in the SB-431542 condition for the 'baseline on' genes, as expected from the fact they are transcribed in the absence of NODAL/Activin. Upon acute pathway activation, however, Pol II was further recruited at these genes, as suggested by the fact that the signal for both marks increased within 1 hr of Activin treatment. Importantly, I observed that the amplitude of the curves changed upon signalling, and not their shape. This is the opposite of what would be expected in the case of 'stalled' Pol II (Figure 4.4).

A similar result was also obtained when the same type plots were generated with respect to the gene expression categories previously defined (see Section 3.2.2). Note that metaprofiles for the 'transiently induced' genes were not displayed, since their relevance was hard to interpret due to the small number of genes included in this group. Overall, the amplitude of Pol II Ser5P and Pol II Ser2P profiles mirrored the different dynamics of gene expression over time, whilst their shape stayed constant both at the TSS and at the TTS (Figure 4.5). A closer look to the 'induced sustained' plots also revealed that for these genes Pol II occupancy decreased at later time, providing a plausible explanation for the attenuation in the rate of transcript accumulation detected by RNA-seq (Figure 4.3A, Figure 3.5). Interestingly for the 'delayed' genes, both isoforms of Pol II were seen at 1 hr and remained bound at 8 hr (Figure 4.5).

Summarising this section, all the data described provide no evidence for NODAL/Activin signalling regulating transcription via a pause-release Pol II mechanism, rather Pol II is *de novo* recruited by SMAD signalling to induce target gene expression.

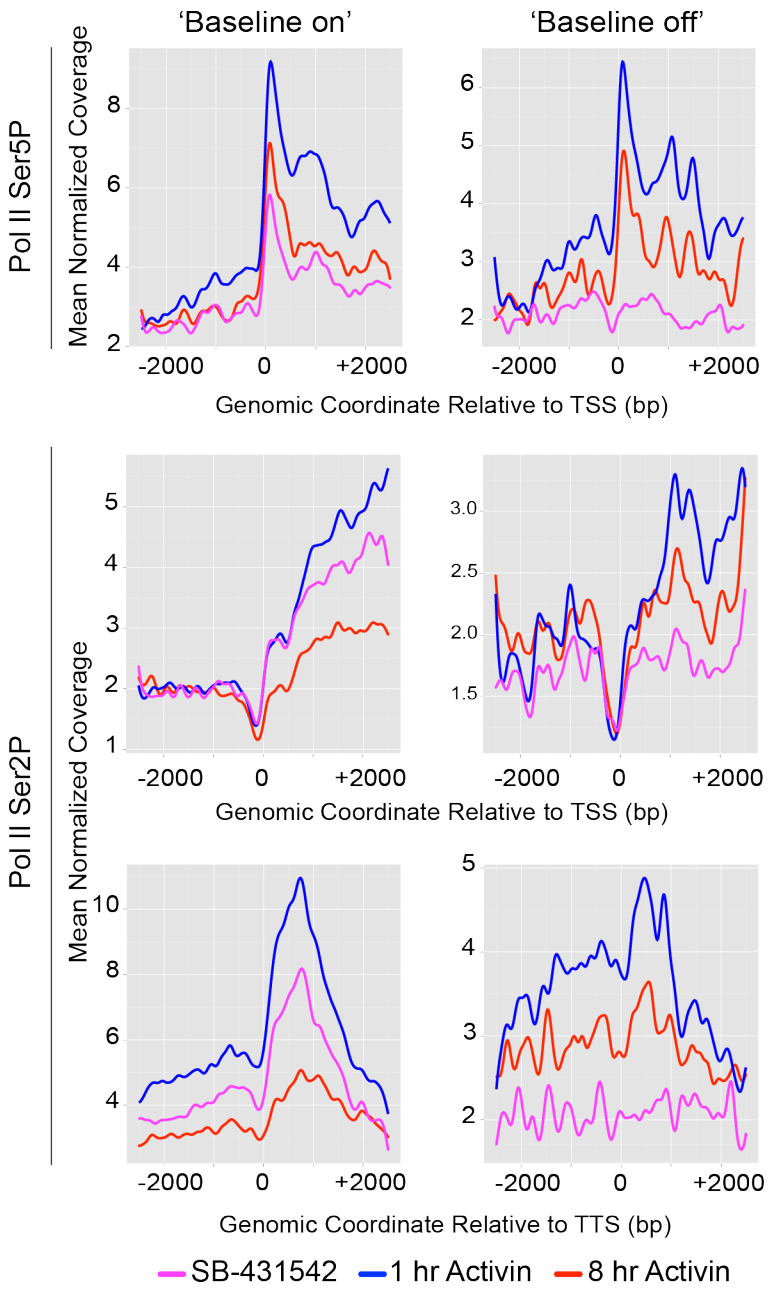


Figure 4.4. Pol II is regulated by SMAD signalling via *de novo* recruitment.
For 'baseline on' or 'baseline off' genes, the averaged signals of Pol II Ser5P or Pol II Ser2P ChIP-seq at each condition were computed and plotted as coloured lines. The metaprofiles span a 5 kb window centred on the TSS or TTS of the genes in the two groups.

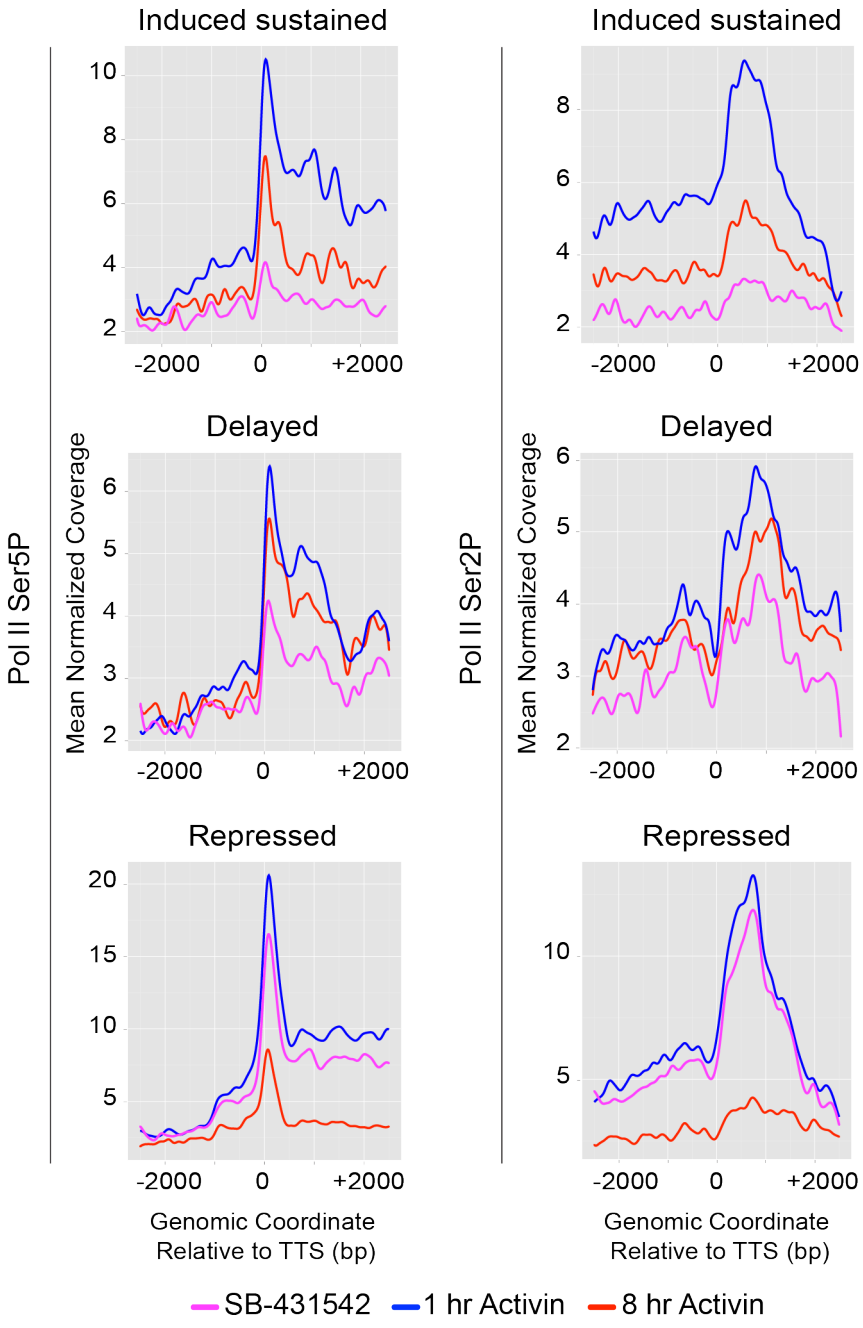
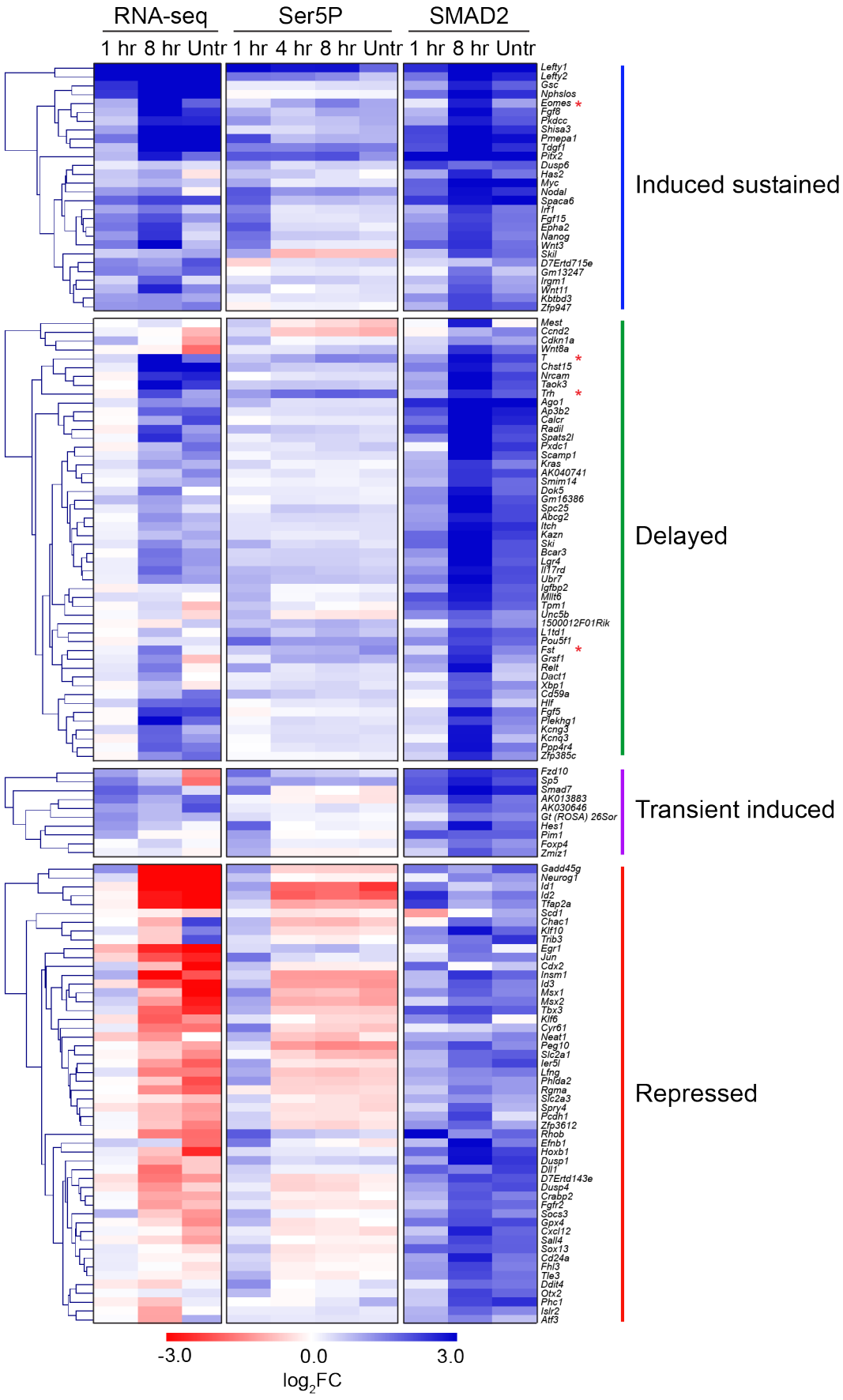


Figure 4.5. Activin-mediated Pol II recruitment and the dynamics of gene expression.
For the indicated groups of target genes, metaprofiles of averaged Pol II Ser5P and Pol II Ser2P signals were obtained as described in the legend to Figure 4.4.

4.2.2 SMAD2 chromatin binding does not directly correlate with transcription over time

Since the heatmap in Figure 4.3A demonstrated that the RNA-seq profiles follow the dynamics of Pol II enrichment over time, which in turn I have shown to be directly regulated by NODAL/Activin signalling via a recruitment mechanism, I wanted to ask how SMAD2 binding on chromatin correlates with these two datasets for each time point. Intuitively, one would expect SMAD2 occupancy to decrease or disappear for all those genes where transcription (measured by RNA-seq and Pol II ChIP-seq) is either dampened or repressed at later time points. Indeed, a general assumption in the transcription field has been that transcriptional termination results from loss of TFs from the chromatin template, and a direct correlation between the amount of SMAD complexes in the nucleous and the levels of transcriptional activity has been proposed elsewhere (Lee and Young, 2013, Warmflash et al., 2012). Yet this scenario is at odds with the previous observation that on-going signalling is necessary for correctly modulating the expression of target genes over time (see Section 3.2.2, Figure 3.7). To address this question Tessa Gaarenstroom and Philip East from BABS quantified the SMAD2 binding for each gene at each time point by defining a SMAD2 footprint which accounted for both the number of peaks associated to the gene and their intensity (For details, see Section 2.4.3). The footprint values for the 140 SMAD2 targets where then plotted alongside the RNA-seq and Pol II Ser5P ChIP-seq data, grouping them according to the four categories of gene expression. The heatmap in Figure 4.6 revealed that overall SMAD2 binding to chromatin increased or did not change with prolonged signalling, in sharp contrast to the pattern of SMAD2 phosphorylation induced by NODAL/Activin, which instead peaked at 1 hr and then attenuated over time (see Section 3.2.1, Figure 3.2). More importantly, the SMAD2 footprint did not generally mirror the RNA-seq or Pol II dynamics. The absence of correlation between SMAD2 footprint and transcription was particularly evident for the ‘transient induced’ and ‘repressed’ genes (Figure 4.6). As well exemplified by *Smad7* and *Tbx3*, the levels of SMAD2 binding at 8 hr after Activin induction were in fact comparable to those at 1 hr, whereas Pol II Ser5P signals were back to baseline at later time points (Figure 4.7).


RNA-seq
Ser5P
SMAD2
1 hr 8 hr Untr 1 hr 4 hr 8 hr Untr 1 hr 8 hr Untr
Lefty1
Lefty2
Gsc
Nphslos
Eomes *
Fgf8
Pkdcc
Shisa3
Pmepa1
Tdgf1
Pitx2
Dusp6
Has2
Myc
Nodal
Spaca6
Irf1
Fgf15
Epha2
Nanog
Wnt3
Skil
D7Ertd715e
Gm13247
Irgm1
Wnt11
Kbtbd3
Zfp947
Induced sustained
Mest
Cond2
Cdkn1a
Wnt8a
T *
Chst15
Nrcam
Taok3
Trh *
Ago1
Ap3b2
Calcr
Radil
Spats2l
Pxdc1
Scamp1
Kras
AK040741
Smim14
Dok5
Gm16386
Spc25
Abcg2
Itch
Kazn
Ski
Bcar3
Lgr4
Il17rd
Ubr7
Igfbp2
Mllt6
Tpm1
Unc5b
1500012F01Rik
L1td1
Pou5f1
Fst *
Grsf1
Relt
Dact1
Xbp1
Cd59a
Hlf
Fgf5
Plekhg1
Kcng3
Kcnq3
Ppp4r4
Zfp385c
Delayed
Fzd10
Sp5
Smad7
AK013883
AK030646
Gt (ROSA) 26Sor
Hes1
Pim1
Foxp4
Zmiz1
Transient induced
Gadd45g
Neurog1
Id1
Id2
Tfap2a
Scd1
Chac1
Klf10
Trib3
Egr1
Jun
Cdx2
Insm1
Id3
Msx1
Msx2
Tbx3
Klf6
Cyr61
Neat1
Peg10
Slc2a1
Ier5l
Lfng
Phlda2
Rgma
Slc2a3
Spry4
Pcdh1
Zfp3612
Rhob
Efnb1
Hoxb1
Dusp1
Dll1
D7Ertd143e
Dusp4
Crabp2
Fgfr2
Socs3
Gpx4
Cxcl12
Sall4
Sox13
Cd24a
Fhl3
Tle3
Ddit4
Otx2
Phc1
Islr2
Atf3
Repressed
-3.0
0.0
3.0
log2FC

This pattern was seen as well when the ChIP-seq data for SMAD2 and Pol II Ser5P were visualised on the IGV browser (Figure 4.8A, C). With respect to the plots shown in Figure 4.7, also note that for *Smad7* the mRNA levels (deduced from the RNA-seq data) decreased more slowly over time than the Pol II enrichment, reflecting the time necessary to degrade the mRNA accumulated after the initial induction. Overall, I concluded that for these groups of genes the downregulation of transcription observed with prolonged signalling was not a consequence of loss of SMAD2 binding. Rather, the data suggested that the block to transcription was likely mediated by repressors recruited by chromatin-bound SMAD2, consistent with the finding that modulation of gene expression at later time points required continuous signalling and protein synthesis. Regarding the other two groups of genes, the SMAD2 footprint better correlated with transcription dynamics. For most of the genes that are induced with sustained or delayed kinetics, such as *Lefty1* and *Eomes*, SMAD2 binding mirrored the RNA-seq profiles (Figure 4.6, 4.7). However, when Pol II peaked at 1 hr as was the case for many 'induced sustained' genes like *Lefty1*, SMAD2 enrichment increased or stayed constant over time, suggesting once more that SMAD2 could play a crucial role not only in the acute induction of transcription, but also in its control over time (Figure 4.6 and Figure 3.10). Intriguingly, I also noticed that some of the genes which are not induced at 1 hr, but induced at 8 hr seem to have more Pol II bound at 1 hr, with the important exceptions of *T*, *Trh, Fst* and *Eomes.* (Figure 4.6 and Figure 4.8B).

Taking all the observations in this section together it is clear that dynamics of SMAD2 binding to chromatin do not reflect the amount of pSMAD2 detected at each time point and do not linearly equate with transcription over prolonged signalling conditions.

Figure 4.6 on previous page.
Figure 4.6. SMAD2 chromatin binding does not correlate with transcription.
The genes in the high confidence dataset were divided according to the four kinetics categories. For each group of genes, the heatmaps display the log_2FC relative to SB-431542 for RNA-seq (left) and for Pol II Ser5P mean normalised read depth (middle), alongside with SMAD2 footprint (right). The order of genes in each group reflects the hierarchical clustering of the data presented on the left. The red asterisks indicate the positions on the heatmap of *Eomes, T*, *Trh* and *Fst*. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

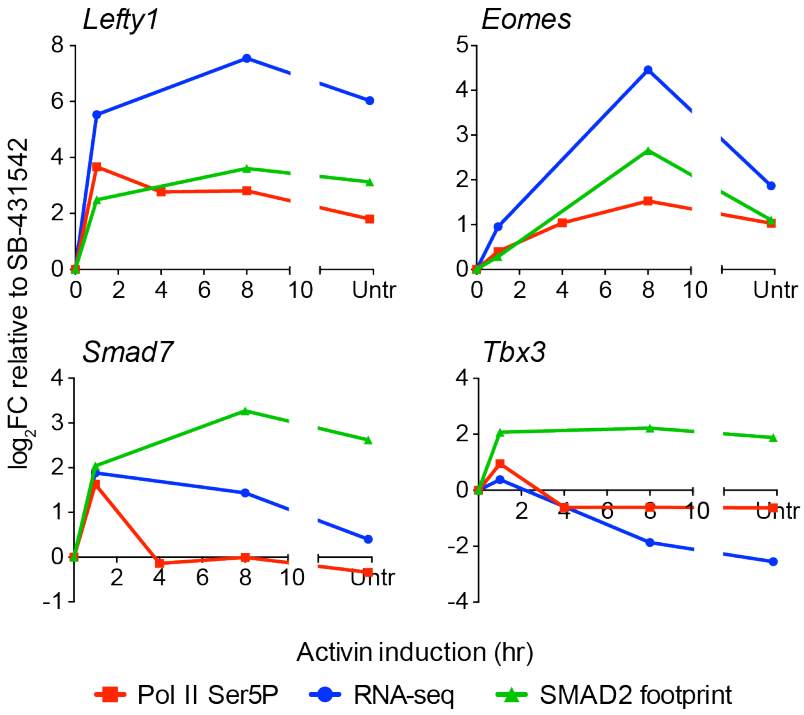


Figure 4.7. SMAD2 chromatin binding and the correlation with transcription: some representative examples.
Each plot displays the values of RNA-seq, Pol II Ser5P and SMAD2 footprint from Figure 4.6 for four genes (*Lefty1*, *Eomes*, *Smad7*, *Tbx3*) chosen as representative of the different transcriptional and SMAD2 binding dynamics observed. Untr, Untreated.

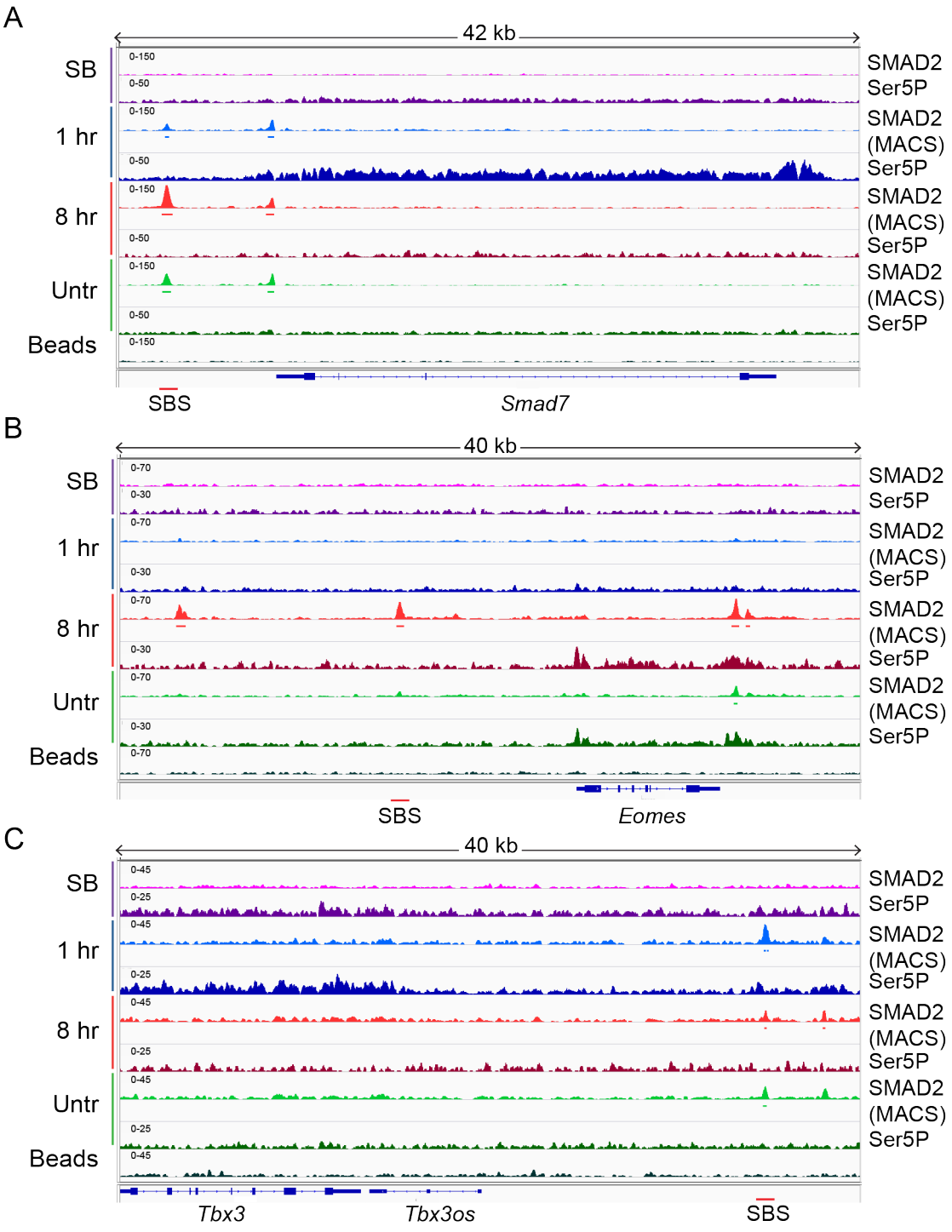


Figure 4.8. Different genes show distinct Pol II and SMAD2 binding dynamics. **(A-C)** IGV browser visualisation of ChIP-seq for SMAD2 and Pol II Ser5P over *Smad7* **(A)**, *Eomes* **(B)** or Tbx3 **(C)** genomic loci. For the SMAD2 ChIP-seq the MACS-called peaks are also shown. SBS, SMAD2 binding site; SB, SB-431542; 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

4.3 Discussion

4.3.1 Summary of main findings

- The ChIP-seq signals for Pol II Ser5P and Pol II Ser2P show the expected profiles of the initiating and elongating forms of Pol II on a genome-wide scale.
- NODAL/Activin signalling induces changes in Pol Ser5P and Pol II Ser2P occupancy at SMAD2 target genes and their binding dynamics follow the same pattern over time.
- There is no evidence for NODAL/Activin regulating transcription via a pause-release Pol II mechanism, instead target gene expression is achieved by SMAD signalling inducing Pol II *de novo* recruitment.
- Overall, Pol II binding dynamics are in agreement with the four distinct kinetics profiles defined by RNA-seq.
- SMAD2 chromatin binding over time does not linearly equate with levels of pSMAD2, nor with the transcription profiles downstream of pathway activation.

4.3.2 NODAL/Activin signalling induces transcription via Pol II recruitment

By comparing the ChIP-seq signals for Pol II Ser5P and Pol II Ser2P I addressed in this chapter by what mechanism SMAD signalling induces transcription of target genes differing in terms of biological function and kinetics of induction. Interestingly, I found no evidence for Activin-dependent Pol II pause-release in P19 cells, rather Pol II is recruited *de novo* by SMAD signalling regardless of the genes categories analysed. This result is particularly striking when focusing on genes induced from an off baseline as deduced by RNA-seq. Included in this group are both classic NODAL/Activin targets (*Lefty1*, *Lefty2*, *Pmepa1*, *Tdgf1*) and markers of mesendoderm differentiation such as *Gsc*. Moreover, some transcripts like that of *Lefty1* are readily detectable within 1 hr of Activin treatment, whereas others like those of *Tdgf1* or *Gsc* are induced with delayed kinetics. Nevertheless, the transcription of these genes is regulated via the same mechanism, which is Pol II *de novo* recruitment. This finding is rather unexpected for a number of different reasons.

First, Pol II pause-release is employed in other systems to rapidly switch on genes in response to external stimuli like heat shock or growth factors (Liu et al., 2015). Therefore, the acute induction of Activin target genes via such a mechanism

seemed plausible. Secondly, in the context of ESCs, pause-release is crucial to achieve coordinated differentiation and many developmental genes in this context have been associated with paused Pol II, including classic Activin target genes such as *Eomes* and *Gsc* (Adelman and Lis, 2012, Estaras et al., 2015). However, in the case of P19 cells Pol II is recruited *de novo* by SMAD signalling at these two genes. To account for this discrepancy, it may be possible that the same genes are regulated differently with respect to Pol II in different cell contexts. In hESC, for example, paused Pol II is recruited to the promoters of *Eomes* and *Gsc* by Wnt signalling, and Activin-SMAD2/3 signalling is then required to release Pol II and to promote Pol II elongation (Estaras et al., 2015). It could be speculated that the first Wnt-dependent step is not required in P19s for the induction of these two genes. Moreover, it is interesting to note that in ESCs developmental genes are bivalently marked with both activating and repressive histone modifications (Bernstein et al., 2006, Buecker and Wysocka, 2012), whereas in P19s I observed very little H3K27me3 at these sites in unstimulated conditions. Whether distinct chromatin signatures are mechanistically related to the presence of paused Pol II has not been addressed. To this end, it could be tested if SMAD signalling regulates Pol II via *de novo* recruitment also in other cell lines with an epigenetic landscape comparable to that of P19s.

Similarly, it would be relevant to explore if Pol II pause-release is at all used in P19 cells to control gene expression in response to any other extracellular signal. Indeed, by visual inspection I was not able to find evidences for paused Pol II on classic growth factors targets like *Myc*, *Fos* and *Jun* (Liu et al., 2015). Of course, the experimental setting used for studying NODAL/Activin signalling is not ideal for addressing this question and a time course of serum induction at times shorter then 1 hr would be more appropriate.

4.3.3 Transcription dynamics over prolonged NODAL/Activin signalling and correlation with SMAD2 chromatin binding

Since changes in transcription and SMAD2 genome binding in response to NODAL/Activin were measured over different times, was possible not just to study how expression of target genes is initiated, but also how is maintained or terminated in prolonged signalling conditions. As already pointed out, the relationship between SMAD binding and transcription has been previously investigated by others.

However, in these studies the authors either only focused on autocrine chronic signalling or inductions with ligands were performed on a time scale of days, rather than minutes or hours (Beyer et al., 2013, Brown et al., 2011, Estaras et al., 2015, Kim et al., 2011, Mullen et al., 2011). How SMAD binding correlates with transcription of target genes has remained unclear.

Before considering this question in respect to SMAD2, it is important to discuss the correlation observed between Pol II occupancy over time and mRNA levels of target genes. Intuitively, transcript levels should mirror Pol II binding over time. This assumption however does not take into account that the amount of mRNA measured at each time point is the result of degradation rate and transcription rate. As pointed out in the Results section, for most of the genes, Pol II enrichment peaks at 1 hr, whereas transcript levels increase over time. Nevertheless, if considering the rate of change of transcript levels it can be noted that the rate at which mRNAs accumulate for the 'induced sustained' and 'transiently induced' genes is maximum at 1 hr. Assuming that the mRNA degradation rate stays constant with prolonged signalling, this could be explained with a decrease in the transcription rate, which in turn would account for the reduction of Pol II binding observed on these genes at later time points. However, establishing a direct link between the amount of Pol II enrichment measured on a gene and its transcription rate in the absence of further data, is at least premature. To shed some light on this, with help from Harshil Patel, we estimated the Pol II travelling ratio over time across different groups of genes. For each SMAD2 target gene, the travelling ratio was defined as the ratio of Pol Ser5P enrichment at the TSS and Pol Ser5P enrichment over the gene body, accordingly to what described by Rahl et al. (Rahl et al., 2010). Rather confusedly, we found that the travelling ratio was constant over time within the same gene category. The picture was further complicated by the fact that genes from the same exhibited different travelling ratios, with genes with short mRNA half-lives having high travelling ratios and high Pol II occupancy, compared with more stable genes. Clearly, this preliminary analysis highlights the complexity of establishing a direct relationship between the amount of Pol II present on a gene and the amount of mRNA as measured by RNA-seq. GRO-seq represents the gold standard approach to measure on-going transcription rates (Jonkers et al., 2014), and it will be important in the future to use this technique to investigate how Pol II enrichment relates to transcription rate.

Having said that, it is clear that for both ‘induced sustained’ and ‘transiently induced’ genes transcription is modulated or dampened at later time points, whilst SMAD2 binding often increases in peak numbers/intensity or stays constant. A similar scenario is also observed for some of the repressed genes, which are rapidly transcribed in response to Activin before being actively switched off. Since I demonstrated in the previous chapter that on-going NODAL/Activin signalling is crucial for long term modulation of gene expression, it is evident that retained SMAD2 binding to chromatin has a functional role in orchestrating the transcriptional responses described. Importantly, this observation opposes the assumption that TFs binding equates with gene expression and that transcription termination result from loss of TFs binding. However, how mechanistically SMAD2 regulates secondary repression or secondary dampening, is still to be addressed. A plausible hypothesis is that over prolonged signalling factors acting as repressors replace transcriptional activators at SMAD bound complexes or are recruited to the additional SMAD2 peaks detected at later time points. Indeed, repressors such as TGIF, SKI, SKIL, ZEB1/2 and EVI-1 have been shown to interact with SMAD2/3 and it would be interesting to verify their relevance for NODAL/Activin signalling in P19 cells (Wotton et al., 1999, Tsuneyoshi et al., 2012, Postigo, 2003, Kurokawa et al., 1998, Deheuninck and Luo, 2009). Additionally, it is known that in other model systems SMAD complexes can interact with distinct TFs at different target genes or switch binding sites and/or partners over time (Beyer et al., 2013, Mullen et al., 2011). Since work in the lab has shown that many of the delayed and repressed genes require protein synthesis for their correct expression profiles, one could also test if newly synthesized TFs are responsible for delayed SMAD2 binding occurrences (eg. *Eomes* and *Trh*) and secondary transcriptional events.

To conclude, the experimental evidence strongly suggests the existence of a NODAL/Activin-depended transcriptional network responsible to orchestrate the expression of target genes over prolonged signalling conditions. How mechanistically this transcriptional programme is set in place and which are the key players involved needs to be addressed in future work.

Chapter 5. SMAD2 induces changes in the chromatin landscape and the mechanisms underlying it

5.1 Introduction

In the last several years, a number of studies have used Next Generation Sequencing (NGS) technologies to investigate TF binding and epigenetic changes in different cellular contexts (Buecker and Wysocka, 2012, Rada-Iglesias et al., 2011, Spitz and Furlong, 2012). However, much of this mapping has been carried out in steady state conditions, and it remains unclear whether TF occupancy induces chromatin modifications, or if the chromatin landscape dictates TF recruitment. The best way to shed light on this question is to study TFs that dynamically bind chromatin and regulate transcription in response to extracellular stimuli, provided that time courses of ligand induction are performed from a signal-inhibited baseline. Since SMAD2 acts as a ligand-inducible transcription factor, in this chapter I use NODAL/Activin signalling as a model system to address the sequences of events occurring on chromatin downstream of pathway activation. My other major aim is to understand how SMAD2 finds its binding sites in chromatin, and to determine the molecular mechanisms involved.

From previous work, it was known that activated SMAD complexes have low affinity and low specificity for DNA, and thus frequently require other TFs to bind the chromatin substrate (Gaarenstroom and Hill, 2014). Indeed, activated SMADs have been shown to interact with distinct TFs in different cell types, amongst which are FOXH1, PU.1, MyoD1, MIXER, POU5F1/OCT4, SOX2, NANOG, EOMES and TEAD (Beyer et al., 2013, Brown et al., 2011, Chen et al., 1996, Faial et al., 2015, Germain et al., 2000, Kunwar et al., 2003, Mullen et al., 2011). Since many of these TFs are responsible for maintaining the respective cell identities and are thought to act as pioneer-factors (Brown et al., 2011, Mullen et al., 2011), the assumption emerged in the field that SMADs are recruited to already accessible chromatin by pre-bound TFs. However, in these studies the binding of such pioneer factors has not been investigated in the absence of SMAD pathway activation.

Unlike the majority of TFs, the activated SMADs are not able to directly recruit the general transcription apparatus to naked DNA, but instead require the chromatin

substrate to regulate transcription (Ross et al., 2006). Interactions of SMAD complexes with chromatin modifying proteins such as the histone acetyltransferase EP300, the ATP-dependent helicase SMARCA4 and the H3K27 demethylase JMJD3 have been reported (Dahle et al., 2010, Feng et al., 1998, Ross et al., 2006, Xi et al., 2008). Nevertheless, the mechanistic role played by these enzymes in relationship to SMADs chromatin binding is still unclear.

In this chapter I first characterise by ChIP-seq the levels of H3K9ac, H3K27ac and total H3 at SMAD2 target sites and over regulated genes in the different NODAL/Activin signalling conditions. By coupling genome-wide analyses with rigorous validation of the data at representative regions I identify distinct chromatin remodelling scenarios with respect to SMAD2 binding events. Using an siRNA approach, I then further dissect the role of FOXH1 in recruiting SMAD2 to a subset of target loci, and address the mechanisms underlying the chromatin changes observed at these sites.

5.2 Results

5.2.1 NODAL/Activin signalling induces chromatin changes

To understand how NODAL/Activin signalling leads to changes in chromatin landscape I have performed ChIP-seq (biological replicates) for different histone modifications in the same conditions used for the Pol II ChIP-seq experiment described in the previous chapter. I decided to look at two different histone modifications characteristic of active chromatin (H3K27ac and H3K9ac) and at two different histone modifications characteristic of repressive chromatin (H3K27me3 and H3K9me3) (Calo and Wysocka, 2013, Zentner and Scacheri, 2012, Zentner et al., 2011). Moreover, I performed ChIP-seq for total histone H3 to identify signalling-dependent changes in nucleosome occupancy. As for the Pol II analyses, Harshil Patel in the Francis Crick Institute BABS facility processed the sequencing data for alignment and visualization.

Upon displaying the ChIP-seq signals using the IGV genome browser, it readily appeared that H3K27ac and H3K9ac signals changed at the SBSs and over many genes known to be involved in NODAL/Activin response. However, H3K27me3 and H3K9me3 signals were generally very low over SBSs and SMAD2 target genes and mostly did not change upon Activin treatment, but were indeed enriched at other loci (data not shown). As a result, I decided to only focus on the acetylation datasets. Overall, two distinct scenarios seemed to emerge in absence of signalling. For genes like *Lefty1* or *Pmepa1*, H3 acetylation was completely absent at the level of the SBSs and over the gene bodies and nucleosome occupancy was uniform across these regions (Figure 5.1). Conversely, a basal H3 acetylation was observed for loci like *Pou5f1* or *Trh* (Figure 5.2) both at the level of nucleosome-depleted SBSs and at promoters/TSSs. It is important to note that *Pou5f1* and *Trh* were already expressed in the SB-431542 condition and belonged to the 'baseline on' group of genes, whilst *Lefty1* and *Pmepa1* were 'baseline off' genes. In response to NODAL/Activin signalling, loss of H3 was observed at the enhancer-like SBSs of *Lefty1* and *Pmepa1* and H3 acetylation appeared at these loci and over the TSSs. Interestingly, H3 acetylation was very localised when occurring at the side of these SMAD2 enhancer peaks, where it did not spread further then 500 pb. Upon Activin treatment, H3K27ac and H3K9ac also increased over the *Pou5f1* and *Trh* genomic regions in a similar

manner (Figure 5.1, Figure 5.2). Considering this preliminary overview of the data, I reasoned that NODAL/Activin signalling could induce changes in the chromatin landscape and that SMAD2 could equally bind to loci with opposite characteristics. I therefore decided to address this hypothesis by performing the analyses presented in the next sections.

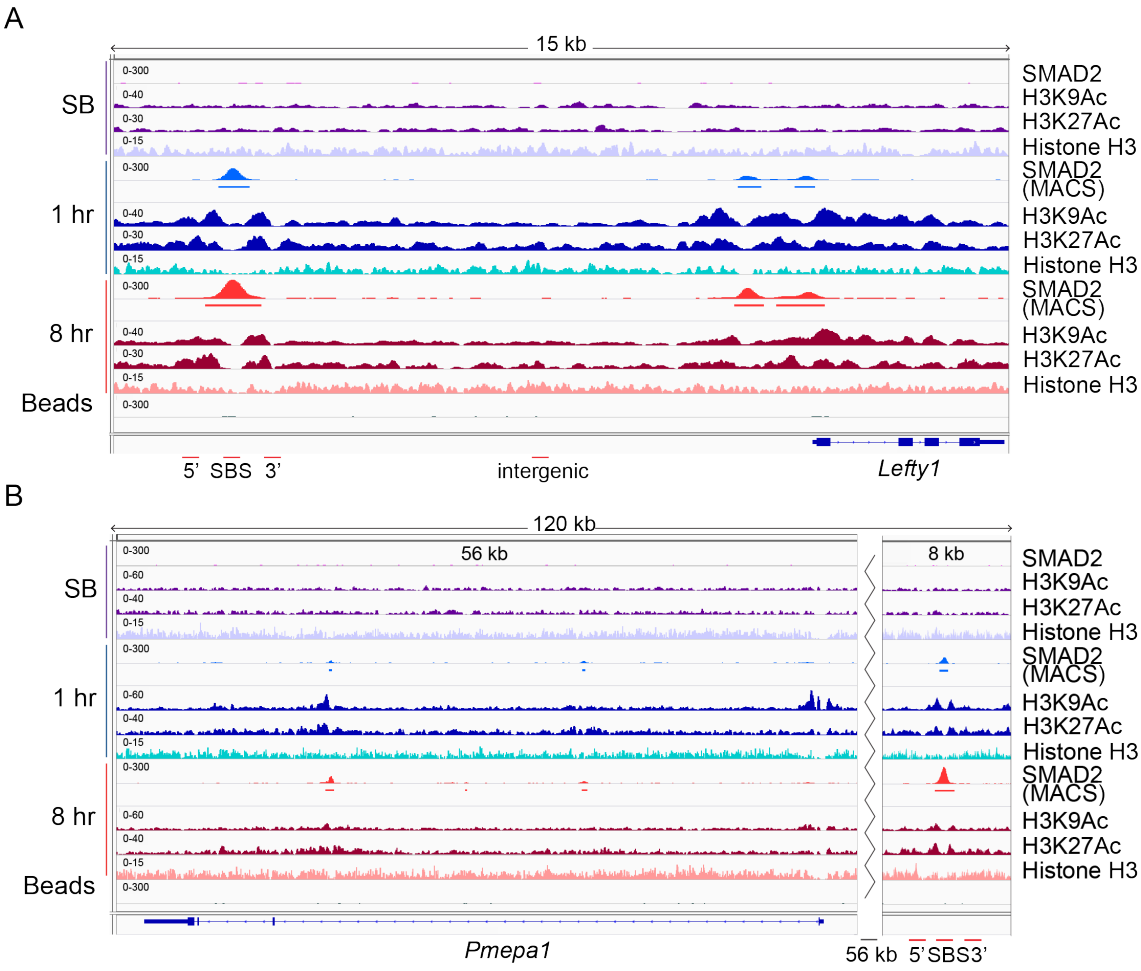


Figure 5.1. NODAL/Activin signalling induces changes in the chromatin landscape around 'baseline off' genes.
IGV browser visualisation of ChIP-seq data for SMAD2, total H3, H3K9ac and H3K27ac obtained in P19 cells treated as indicated. For the SMAD2 ChIP-seq the MACS-called peaks are also shown. The genomic loci displayed refer to the 'baseline off' genes *Lefty1* **(A)** and *Pmepa1* **(B)**. Below each panel are denoted the regions amplified in the ChIP-PCR experiment shown in Figure 5.3. SB, SB-431542; 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

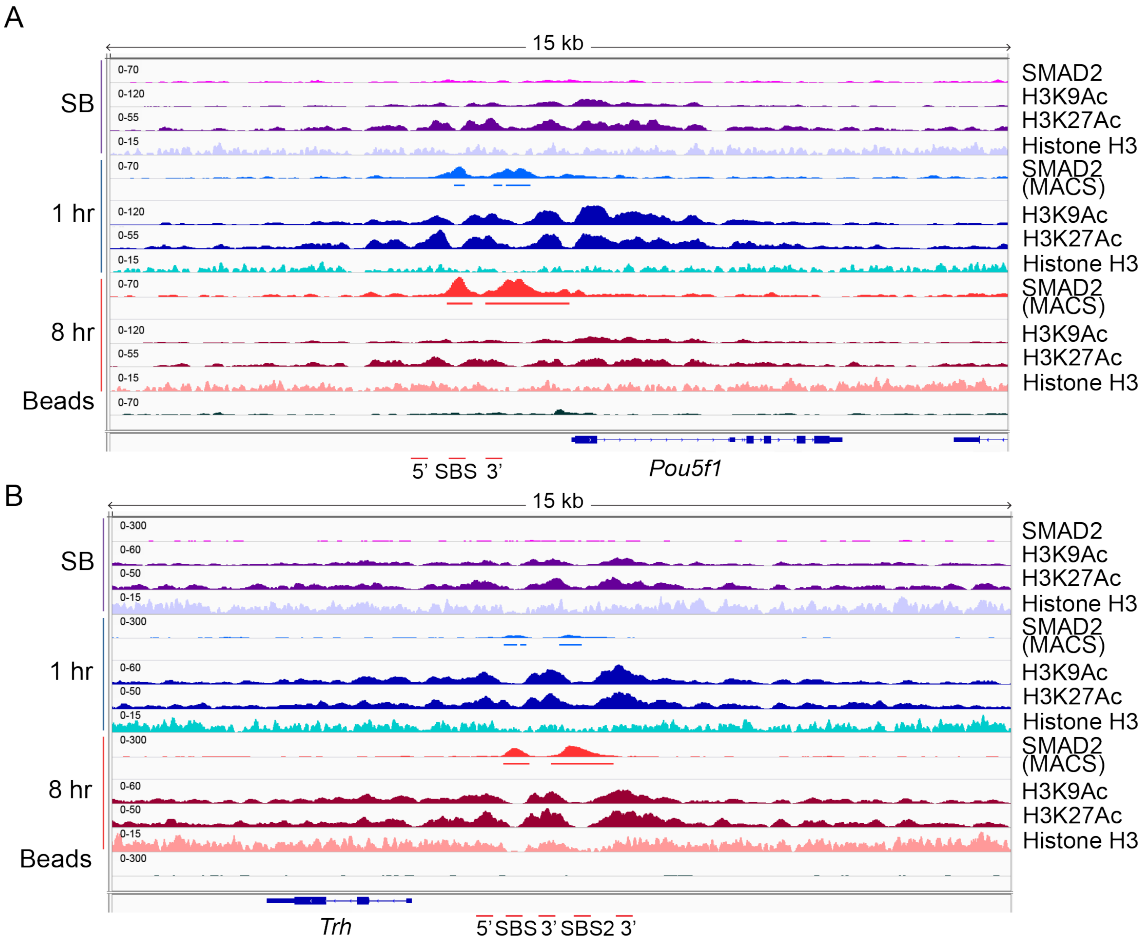


Figure 5.2. NODAL/Activin signalling induces changes in the chromatin landscape around 'baseline on' genes.
IGV browser displays of the same ChIP-seq data described in Figure 5.1 over the genomic regions around the 'baseline on' genes *Pou5f1* **(A)** and *Trh* **(B)**. Below each panel are denoted the regions amplified in the ChIP-PCR experiment shown in Figure 5.3.

5.2.2 SMAD2 binds to closed sites around *Lefty1* and *Pmepa1* and promotes their chromatin remodelling

First, I validated the observations just described by performing ChIP-PCR using primers spanning around 300 bp either side of the SBSs. Indeed, the plots in Figure 5.3 confirm that SMAD2 binding was occurring at sites which had very different chromatin states. In the absence of signalling, *Lefty1* and *Pmepa1* SBSs were in fact in a 'closed' conformation, with high H3 occupancy at the SBSs and low H3K9ac and H3K27ac signals at either side of it. *Trh* and *Pou5f1* loci were instead pre-acetylated and nucleosome depleted, showing the 'pocket-like' shape characteristic of 'open' chromatin. Within 1 hr of Activin treatment, however, *Lefty1* and *Pmepa1* SBSs became more accessible, with H3 level dropping and the flanking nucleosomes acquiring H3 acetylation. Similarly, in response to NODAL/Activin signalling the open conformation of the *Trh* and *Pou5f1* SBSs was further reinforced, as suggested by the increase in H3 acetylation detected at the SBSs 5' and 3' (Figure 5.3, top panel). Importantly, no significant changes were observed over time in the levels of H3K9ac, H3K27ac and total H3 at *Gapdh* TSS. *Gapdh* is not a NODAL/Activin target gene so I used it as the negative control region, and it increased confidence in the results observed at other loci (Figure 5.3, bottom panel). From these ChIP-PCR experiments I therefore concluded that at least at the *Lefty1* and *Pmepa1* gene regulatory regions SMAD2 binding was able to induce chromatin remodelling.

To provide further evidence of an Activin-dependent 'switch' from closed to open chromatin I decided to investigate the presence of other histone modifications commonly associated with active enhancers (Calo and Wysocka, 2013, Rajagopal et al., 2014). I thus performed ChIP-PCR for H3K18ac and H3K23ac over an Activin time course, using the *Lefty1* locus as a representative example. As expected, in the SB-431542 condition the signal for these two marks was uniformly low across the *Lefty1* SBS region. However, after 1 hr of Activin treatment the flanking nucleosomes became significantly acetylated at H3K18 and H3K23, with induction dynamics similar to those described for H3K9ac and H3K27ac (Figure 5.4). This finding suggested that the acetylation of the H3 tail on these different lysine residues might be carried out by the same acetyltransferase, recruited in an Activin-dependent manner to *Lefty1* SBS. EP300 was a good candidate, since it was known to promiscuously acetylate different substrates and it had been previously shown to

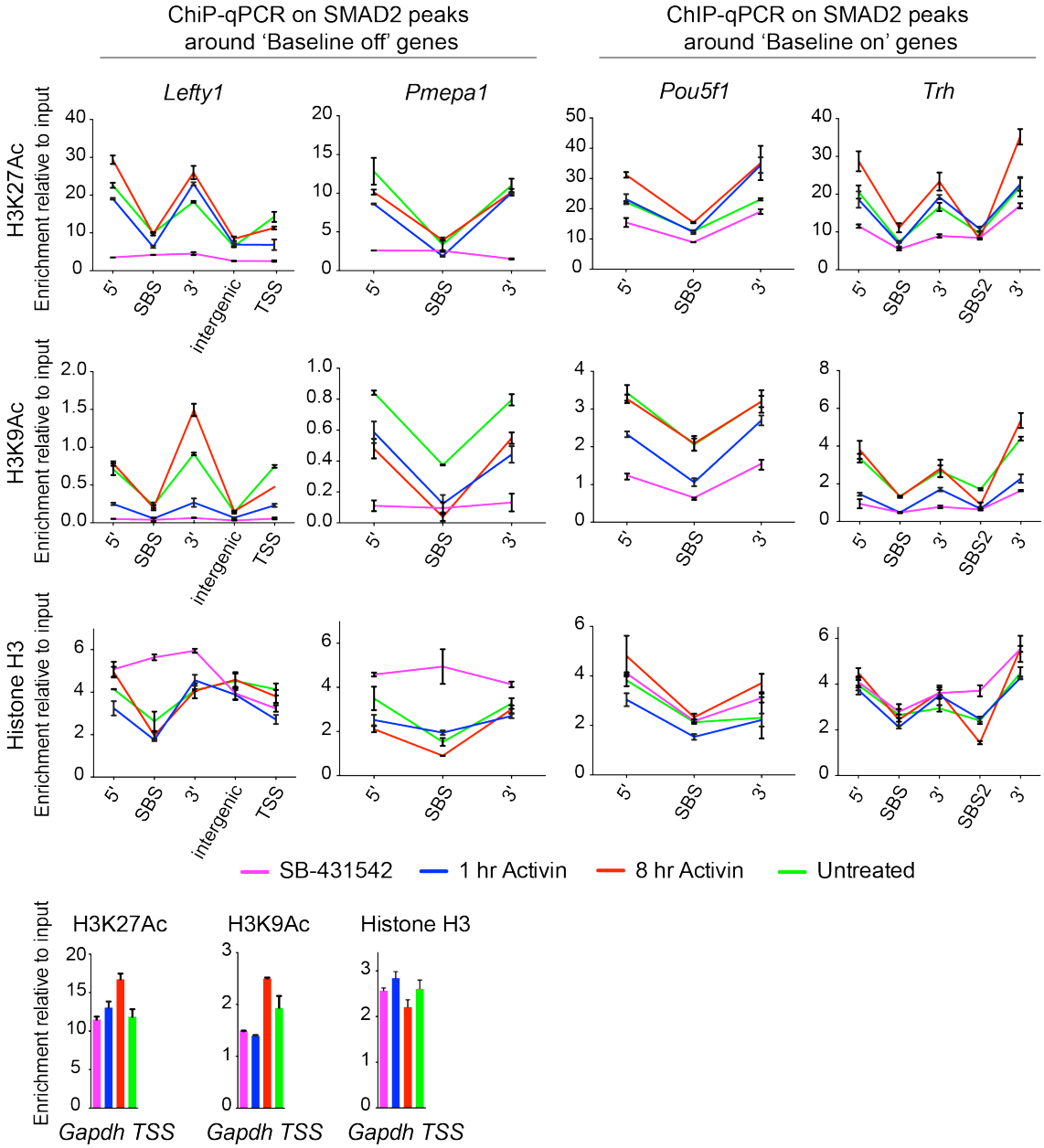


Figure 5.3. Validation of Activin-dependent chromatin remodelling at and around SBSs.
ChIP-PCR for H3K27ac, H3K9ac or Histone H3 was carried out in P19 cells treated as indicated. The amplified regions refer to the SMAD2 peaks around the 'baseline off' genes *Lefty1* and *Pmepa1* or around the 'baseline on' genes *Pou5f1* and *Trh*, denoted by red lines in Figure 5.1 and Figure 5.2. H3Ac and H3 enrichments were also measured 300 pb either side of the SBSs. In the case of *Lefty1*, additional primers were used at the TSS and in the region between the SBS and TSS (intergenic). As a control, the signals obtained for the two histone marks and for H3 over the *Gapdh* TSS are displayed in the bottom panels. A representative experiment is shown (means ± SD).

interact with SMAD2 (Ross et al., 2006). Indeed, ChIP-PCR experiments performed in P19 cells in the same experimental settings showed that EP300 is recruited to the *Lefty1* SBS and other SBSs in response to Activin signalling with the same temporal patterns observed for SMAD2 (Coda et al., 2017).

Since no evidence of NODAL/Activin-mediated nucleosome displacement has been previously reported, I verified this finding by performing FAIRE-PCR, which detects open chromatin with higher sensitivity than ChIP-PCR for histone H3. Indeed, in the SB-431542 condition the plots for *Trh* and *Pou5f1* already had a 'peak-like' shape, suggesting that the SBSs were devoid of nucleosomes compared to the flanking regions. A nearly flat baseline was instead observed for *Lefty1* and *Pmepa1* in absence of signalling, with almost no evidence for open chromatin detected at these SBSs, compared with the other control regions (Figure 5.5A). Upon Activin treatment, however, *Lefty1* and *Pmepa1* plots also acquired the 'peak-like' shape described above, with signal at these SBSs increasing more than five fold compared with the control condition. As expected, for *Trh* and *Pou5f1* the chromatin remained in an open conformation in response to NODAL/Activin signalling (Figure 5.5A-B). In conclusion, this FAIRE-PCR experiment fully confirmed the results obtained with the ChIP-PCR and ChIP-seq for histone H3, and further emphasized the existence of nucleosome displacement downstream of activated SMAD2.

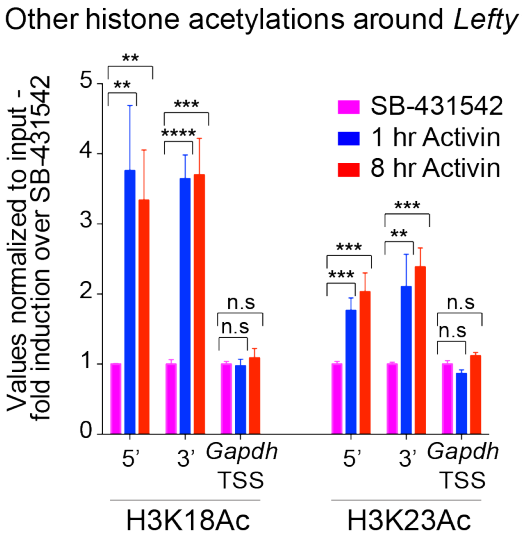


Figure 5.4. Acetylation of H3K18 and H3K23 is induced in response to Activin at the *Lefty1* SBS flanking nucleosomes.
P19 cells were treated as indicated and ChIP-PCR for H3K18ac and H3K23ac was performed over the genomic regions around *Lefty1* SBS and at the *Gapdh* TSS (negative

Figure 5.4 continued.
control) The primers pairs used in this experiment are the same as in Figure 5.3. Plotted are the means and SEM of three independent experiments. n.s not significant; **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01.

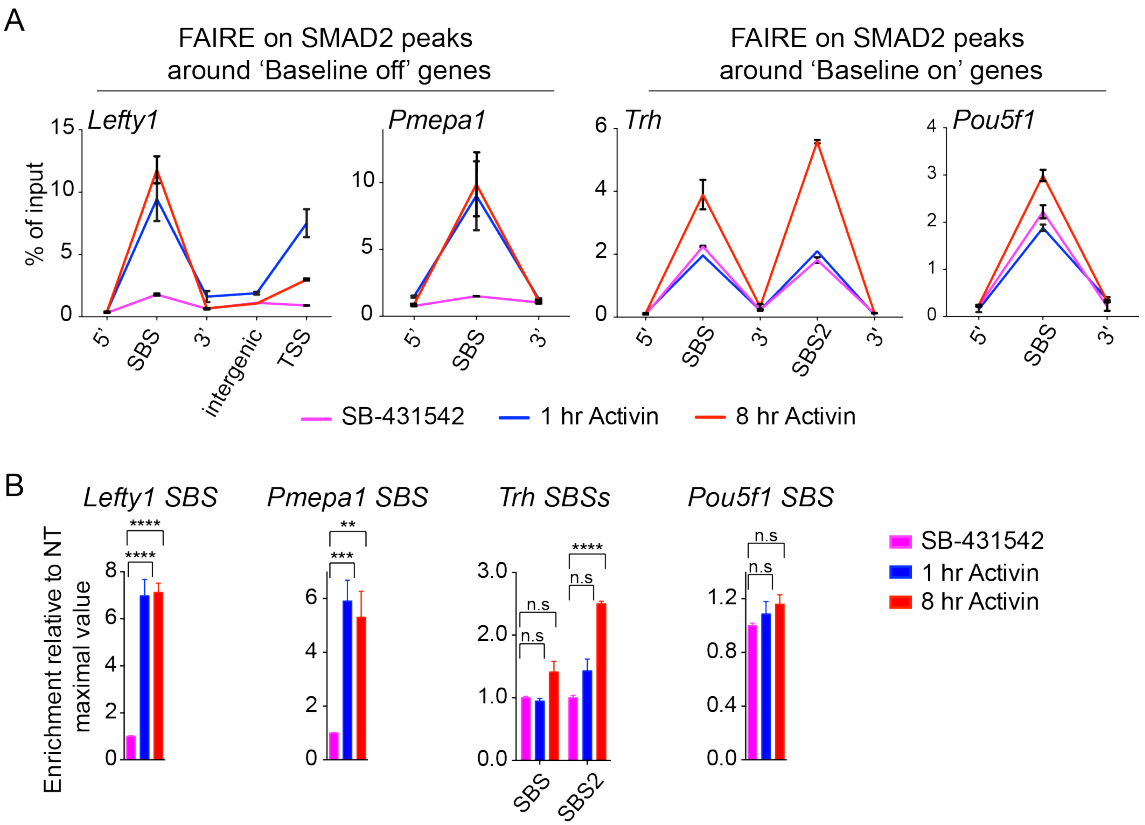


Figure 5.5. NODAL/Activin signalling controls nucleosome displacement at *Lefty1* and *Pmepa1* SBSs.
(A) P19 cells were treated as indicated and FAIRE-PCR was performed. The same regions described in Figure 5.3 were analysed for the enrichment of nucleosome-depleted fragments, using the sets of primers already described. A representative experiment is shown (means ± SD). **(B)** Multiple experiments as in (A) were combined for statistical analysis. Plotted are the means and SEM of three independent experiments for the enrichment at the SBSs. n.s not significant; **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01.

5.2.3 SMAD2 has two different modes of chromatin binding genome-wide

Since the IGV browser displays suggested that NODAL/Activin signalling induced changes in histone modification and/or occupancy at SMAD2 binding sites and associated genes, I set out to test the relevance of this hypothesis for the 'high confidence' dataset of SMAD2 peaks and target genes. The results obtained for the SMAD2 binding sites are presented below and in Section 5.2.4, whereas the analysis concerning the target genes is reported in Section 5.2.5.

First, I investigated how well the two H3ac modifications correlated at the level of SBSs, as exemplified by the plots in Figure 5.6. Here, for H3K9ac and H3K27ac the normalised read depth and the $\log_2$FC relative to SB-431542 are compared for the 1 hr time point only, but analogous results were obtained for the other time points as well (data not shown). These findings provided a confirmation that NODAL/Activin signalling induced similar changes in H3K9ac and H3K27ac around SBSs, however could not provide information about how nucleosome occupancy or H3 acetylation evolved over time with respect to different sets of SMAD2 peaks.

To address this aspect, Philip East (Francis Crick Institute, BABS facility) generated a series of 5 kb window metaprofiles for total histone H3, H3K9ac and H3K27ac. The plots were centred around distinct groups of loci that had a SMAD2 peak in at least one signalling condition. When initially considering the 478 SBSs together, it looked as if in SB-431542 all the loci were already nucleosome-depleted and acetylated, with signalling only leading to a moderate increase in H3 acetylation (Figure 5.7). At first, this finding seemed to contradict the presence of two distinct scenarios, as previously hypothesized after visual inspection of the data. However, I reasoned that the signals from 'closed', unacetylated sites were difficult to detect in profiles averaging hundreds of loci. Indeed, when the loci associated to 'baseline on' or 'baseline off' genes were plotted separately, the differences between these two groups in term of NODAL/Activin dependent histone modification/occupancy readily emerged.

For 'baseline on' genes SMAD2 binding occurred at pre-acetylated nucleosome-depleted sites and resulted in a modest increase in H3K9Ac and H3K27Ac. Conversely, for 'baseline off' genes NODAL/Activin signalling induced a sharp decrease in H3 over SBSs and a robust acetylation of the nucleosomes at either side of the SMAD2 peak (Figure 5.7). Such effects were even more striking

when focusing only on those sites which showed the highest induction of histone acetylation (for the criteria used to identify these sites, see Figure 5.9). At these loci in the SB-431542 condition H3 acetylation was completely absent and histone H3 occupancy was uniformly distributed across the 5 kb window, indicating that the SBSs were in a completely 'closed' state. Here, Activin treatment clearly resulted in chromatin 'opening-up', with nucleosome displacement occurring at the SBSs and acquisition of H3K9ac and H3K27ac at 1.5 kb either side (Figure 5.7).

Overall, these plots confirmed on a genome-wide level the existence of two distinct modes of SMAD2 chromatin binding. In fact, they show that SMAD2 does not only target pre-acetylated nucleosome-depleted sites, but it also binds to closed, inactive chromatin where it induces nucleosome release and acetylation.

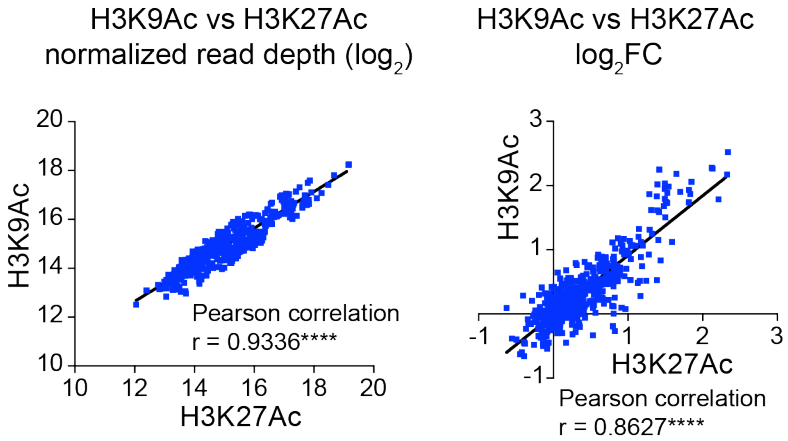


Figure 5.6. Acute Activin treatment induces similar changes in H3K9ac and H3K27ac at SMAD2 binding sites.
Correlation plots comparing the mean normalised read depth of H3K9ac and H3K27ac over a 5 kb window around each of the 478 SMAD2 peaks (left panel). The data refer to the 1 hr Activin time point. For this sample, the log_2FC of H3K9ac and H3K27ac in the 5 kb window relative to SB-431542 were also calculated and the correlation plot of the values obtained for the two histone marks is displayed in the right panel.

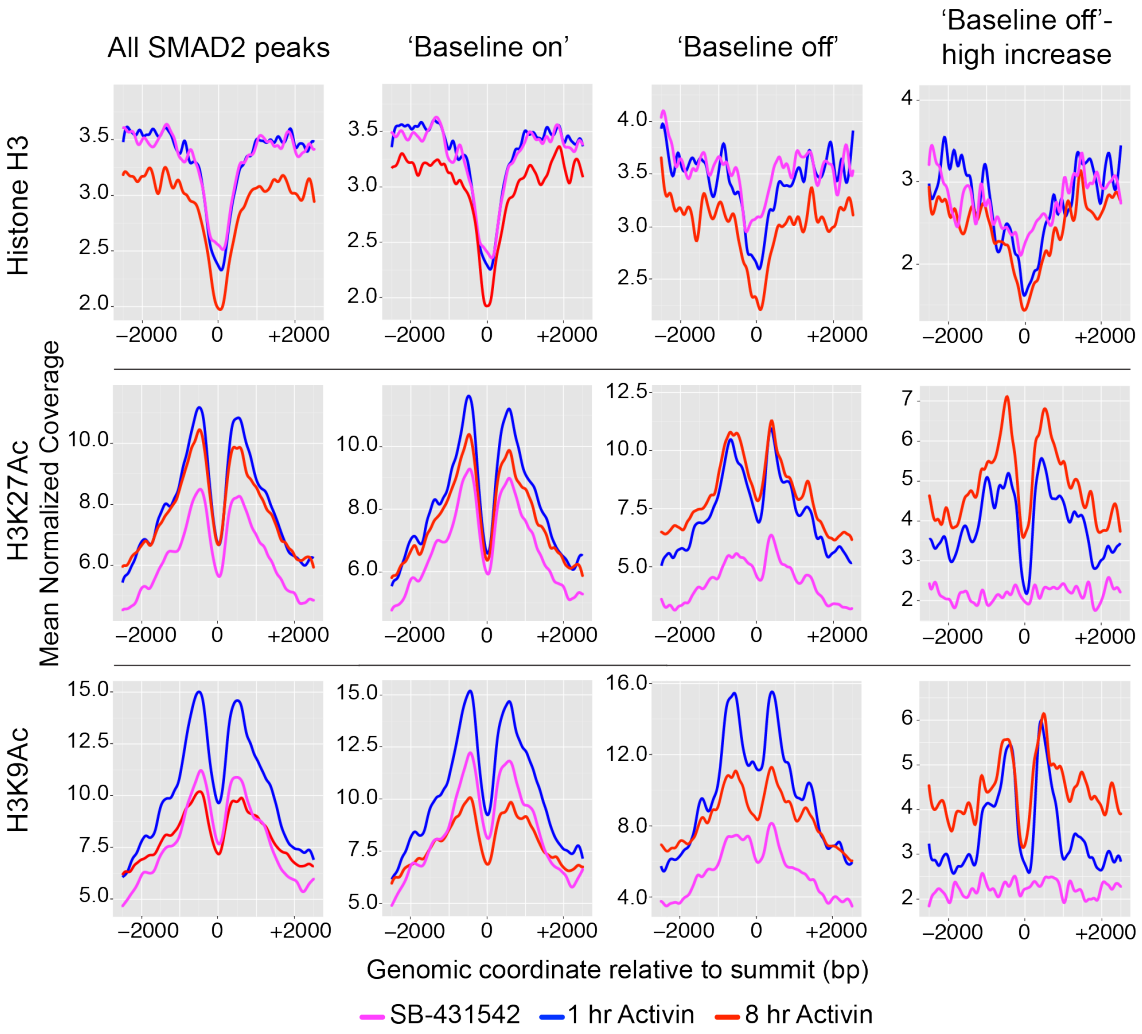


Figure 5.7. Different modes of SMAD2 chromatin binding genome-wide.
The averaged signals of histone H3, H3K27ac or H3K9ac ChIP-seq at each condition was computed and plotted as coloured lines. The metaprofiles span a 5 kb window centred on the average summit across SMAD2 consensus peaks. The plots referring to the same group of SMAD2 peaks are displayed in the same column. To generate the metaprofiles on the far left, all 478 SMAD2 peaks were taken into account. For the other plots, only the SMAD2 peaks associated to 'baseline on' or 'baseline off' genes were considered, respectively. The column of metaprofiles on the far right refers to a subset of SMAD2-bound regions which are found near 'baseline off' genes and acquire H3 acetylation from a low baseline as defined in Figure 5.9.

5.2.4 SMAD2 binding correlates with high histone acetylation

From previous work performed in the lab, it was known that SMAD requires the chromatin context to bind to DNA (Ross et al., 2006). Coupling this observation with the fact that SMAD2 binds to both acetylated and non-acetylated sites as I have shown above, it could be asked which are the chromatin characteristics that make a particular genomic locus a potential SMAD2 target. Addressing this question could ultimately help us to understand how SMAD2 recognises its binding sites in different chromatin contexts. To this end, I investigated if there was any correlation between SMAD2 binding and histone acetylation. This analysis was performed using my data in conjunction with Philip East from BABS and Tessa Gaarenstroom (another student in the lab).

First, for each of the 478 SMAD2 peaks Philip East quantified the mean normalised read depth for H3K9ac or H3K27ac across a 5 kb window over time. In line with the fact that SMAD2 binds to both open and closed chromatin, a spectrum of H3 acetylation levels was observed across the different loci in the SB-431542 state. Not surprisingly, with NODAL/Activin signalling, acetylation generally remained constant or increased following different temporal dynamics depending on the sites considered (Figure 5.8). Based on previous results one could have expected clearer patterns and bigger changes for both histone marks, compared with what was observed in the heatmaps. I reasoned that this discrepancy was likely due to the fact that SMAD2-dependent acetylation was limited to the 1.5 kb either side of the binding site, whilst here the signal was averaged across a larger window. As a consequence, very localised changes could have been cancelled out, increasing the general noise of the data. Nevertheless, two different groups of peaks were identified based on ‘low’ or ‘high’ levels of H3K9ac or H3K27ac acetylation (Figure 5.8). Comparing the signals from the two histone modifications, it could be noted that the number of sites with generally low or generally high H3K9ac was bigger than what was observed for H3K27ac, possibly as a consequence of the hierarchical clustering amplifying the differences existing between the two marks. To obviate to these differences, Tessa Gaarenstroom defined the groups of low overall H3 acetylation and high overall H3 acetylation based on the intersection of the individual datasets, therefore excluding the sites which had just low/high H3K9ac but not H3K27ac, and vice-versa (Figure 5.8).

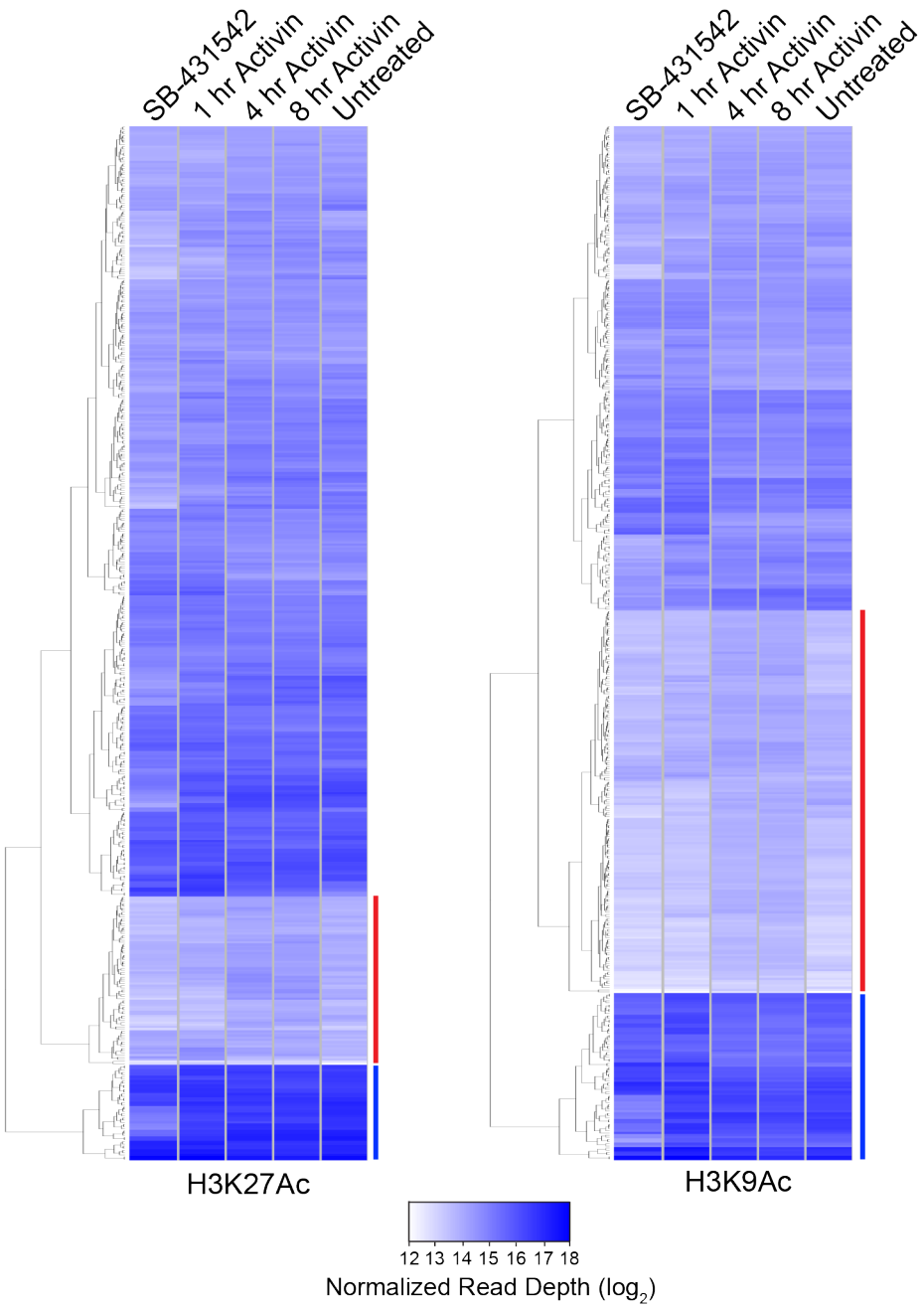


Figure 5.8. Histone acetylation around SMAD2 binding sites is highly variable in the SB-431542 condition and changes in response to NODAL/Activin signalling. For each SMAD2 binding site, normalised read depth (Log_2) over a 5 kb window centred around the SMAD2 consensus peak summit is shown for H3K27ac (left panel) and H3K9ac (right panel). For each histone modification, the red line indicates the group of sites which has low overall acetylation; the group of sites with high acetylation is marked by the blue line.

Initially, we asked how SMAD2 occupancy changed over time relative to these two groups of loci. Interestingly, at every time point SMAD2 binding was stronger at those sites which had overall high H3K9 and H3K27 acetylation (Figure 5.9A). Then, focusing on the SB-431542 condition only, we observed that at 1 hr of Activin treatment SMAD2 was more enriched at loci pre-acetylated compared to non-acetylated ones, whilst no significant difference was observed at the 8 hr time point (Figure 5.9B). Considering these two results it appeared as if pre-existing acetylation was a necessary requirement for stable SMAD2 binding, especially following acute induction. However, such a hypothesis was at odds with the observation that many SMAD2 peaks were detected at sites of closed conformation in the absence of signalling, such as the *Lefty1* and *Pmepa1* loci. To clarify this aspect, the sites which were identified as having low H3 acetylation in the SB-431542 condition were further divided into those showing a high increase in acetylation and into those for which acetylation levels did not change with signalling. Indeed, SMAD2 occupancy over time was significantly higher for the first group of peak compared to the second one (Figure 5.9C). Trying to predict the likelihood of SMAD2 binding to different loci only on the basis of their acetylation levels in the SB-431542 state was therefore misleading. Considering all the analyses together, we finally concluded that SMAD2 preferentially occupies sites which are constitutively acetylated or where acetylation is robustly induced in response to NODAL/Activin signalling.

Having showed that, it remained unclear how SMAD2 was able to target non-acetylated loci apparently indistinguishable in term of chromatin characteristics from the genomic context in which they were located. It then seemed plausible to hypothesise the presence of an additional ‘landmark’ specifying potential SMAD2 binding sites, rather than H3 acetylation. Since in other cell systems enhancers devoid of H3K27ac have been shown to be enriched with H3K4 monomethylation (Calo and Wysocka, 2013), I decided to investigate the presence of this mark in P19 cells. I therefore performed H3K4me1 ChIP-PCR on the SBS region upstream of *Lefty1*, which I chose as a representative example of a closed, non-acetylated binding site. As shown in Figure 5.10, in the absence of signalling, H3K4me1 was strongly enriched at the *Lefty1* SBS and flanking nucleosomes compared to the negative control region. Upon Activin treatment, the H3K4me1 signal dropped at the binding site as a consequence of signal-induced nucleosome displacement, but it

remained constant at either side of the SBS. Thus, at least in the case of *Lefty1*, SMAD2 targets a locus which is pre-marked with H3K4me1.

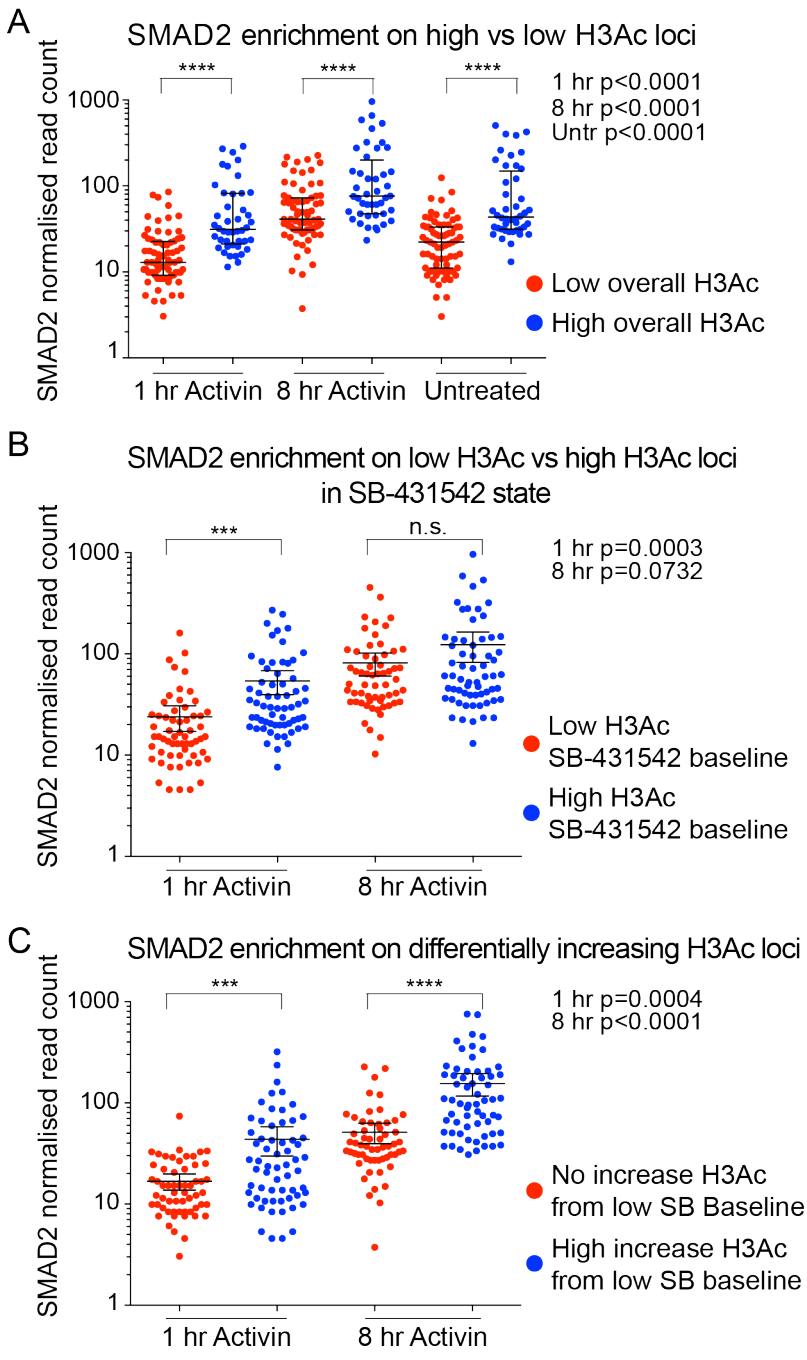


Figure 5.9. High H3 acetylation levels upon NODAL/Activin signalling are predictive of strong SMAD2 chromatin binding.
(A-C) Correlation plots between SMAD2 normalised read counts and groups of binding sites defined on the basis of distinct acetylation states. **(A)** SMAD2 binding sites with 'low overall acetylation' (red) or 'high overall acetylation' (blue) were identified and the SMAD2 read counts over these loci for the 1 hr, 8 hr Activin and Untreated samples are
Figure 5.9 continued on next page.

Figure 5.9 continued.
are displayed. **(B)** SMAD2 binding occupancy is quantified at 1 hr and 8 hr of Activin treatment over those sites with low (red) or high (blue) acetylation in the SB-431542 state. **(C)** The loci in the category of low H3Ac SB-431542 baseline were further separated into two groups based on no increase (red) or high increase (blue) in acetylation following Activin treatment. The relative SMAD2 read counts for the 1 hr and 8 hr Activin samples are shown. For all graphs the black bars indicate the mean and 95% confidence interval. n.s., not significant. The p-values are reported in the plots.

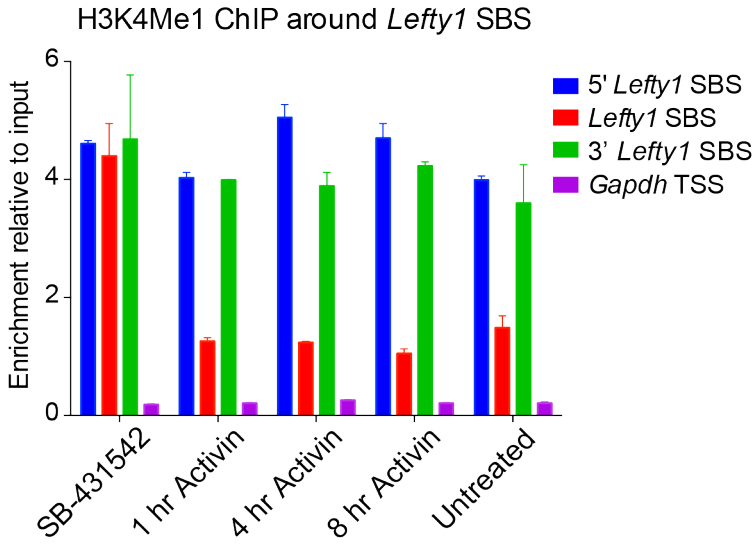
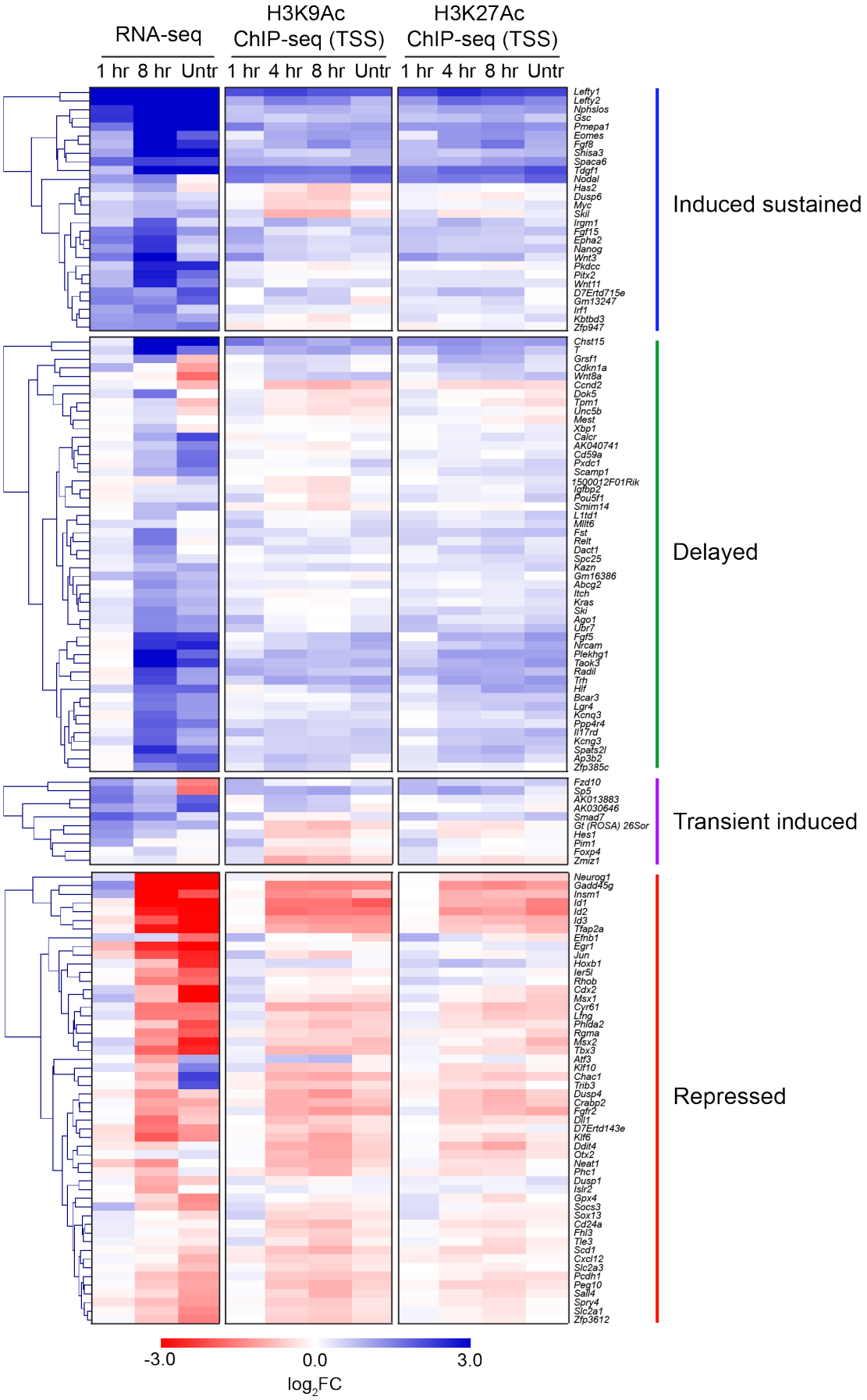


Figure 5.10. In the absence of NODAL/Activin signalling *Lefty1* SBS locus is marked with H3K4me1.
H3K4me1 ChIP-PCR experiment on P19 cells treated as indicated. H3K4me1 signal was measured at the *Lefty1* SBS and flanking nucleosomes using the primers introduced in Figure 5.3. Enrichment over the *Gapdh* TSS is also displayed as a negative control. A representative experiment is shown (means ± SD).

5.2.5 Histone H3 acetylation at SMAD2 target genes correlates with transcription

Since transcription activation is commonly associated with the acquisition of histone acetylation (Zhou et al., 2011), I then wanted to verify if this was observed also for the Activin-regulated genes. Changes in H3K9ac and H3K27ac over a 5 kb window surrounding the TSS of each SMAD2 target gene were quantified and plotted alongside the RNA-seq data. Not surprisingly, H3 acetylation dynamics generally correlated well with the kinetics of gene expression, with a considerable number of the 'induced sustained' and the 'delayed' genes showing an increase in their acetylation levels either at 1 hr or at 4 hr of Activin treatment (Figure 5.11). However, for half of the TSSs in these two groups the signals of the two marks did not seem to change much upon ligand stimulation compared with SB-431542. This is likely due to the fact that they were already robustly expressed in the absence of signalling, and therefore heavily acetylated. I also observed that H3 acetylation was lost on the large majority of 'repressed' genes at 4 hr of Activin stimulation, in line with what was observed for the Pol II ChIP-seq (Figure 5.11). Moreover, as already noted for the Pol II data, generally the 4 hr time point and the 8 hr one looked very similar to each other. For this reason, the 4 hr time point was not taken in consideration in some of the analyses presented in this Chapter. Finally, the heatmap in Figure 5.11 showed that, overall, H3K9ac and H3K27ac changes over time mirrored each other in terms of both temporal patterns and numeric values, suggesting that the deposition/removal of the two marks at the TSSs was orchestrated simultaneously.

Figure 5.11 on next page.
Figure 5.11. Histone H3 acetylation changes correlate with gene expression.
The genes in the high confidence dataset were divided accordingly to the four kinetics categories. For each group of genes, the heatmaps display the log_2FC relative to SB-431542 for RNA-seq (left) and the mean normalized read depth for H3K9ac or H3K27ac (middle and right). The signal for the two histone marks was quantified over a 5 kb window around the TSS of each target gene. The order of genes in each group reflects the hierarchical cluster of the data presented on the left. Untr, Untreated.


RNA-seq
H3K9Ac ChIP-seq (TSS)
H3K27Ac ChIP-seq (TSS)
1 hr 8 hr Untr
1 hr 4 hr 8 hr Untr
1 hr 4 hr 8 hr Untr
Induced sustained
Delayed
Transient induced
Repressed
-3.0
0.0
3.0
log2FC

5.2.6 SMAD2 co-factors on chromatin: characterisation of the distinct roles of FOXH1 and POU5F1

So far, I have highlighted the importance of the chromatin landscape in relationship to the SMAD2 binding events. However, in an attempt to understand how mechanistically SMAD2 recognises different loci on chromatin, the interaction with other TFs must be examined. It is known in fact that SMAD2 does not bind the DNA itself, but it requires the presence of additional TFs (Gaarenstroom and Hill, 2014). Indeed, a MEME-ChIP analysis revealed that the sequences under the summits of the 478 SMAD2 peaks were enriched for motifs bound by known SMAD2 co-factors such as FOXH1, POU5F1 and SOX factors, alongside with unknown motifs (Figure 5.12). It can be noted that the SMAD3/SMAD4 binding element was also found, and this is likely due to DNA binding of SMAD4 in the activated SMAD complex (Inman and Hill, 2002).

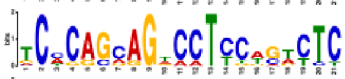
Motif found	E-value	% of peaks (FIMO)	Known Motifs
	2.4e-055	26	FOXH1
	3.4e-042	29	POU5F1/SOX2
	6.8e-038	34	SMAD2/SMAD3/SMAD4
	4.4e-030	33	Unknown
	1.1e-026	29	Unknown
	3.0e-025	17	Unknown
	1.6e-019	33	SMAD2/SMAD3/SMAD4
	4.6e-010	24	SOX Factors

Figure 5.12. DNA motifs for FOXH1, POU5F1 and SOX factors are enriched at SMAD2 binding sites.
The most enriched known and *de novo* motifs at the 478 SMAD2-bound sites were identified using MEME-ChIP. For each of the 478 consensus peak, the sequence of the 500 bp around the summit of each contributing MACS peak was retrieved and used as an input, for a total of 757 sequences. Displayed are E-values from the MEME-ChIP analysis, % of peaks positive for the indicated motif as obtained from FIMO and matching TFs (MEME-ChIP).

The role played by FOXH1 and POU5F1 in Activin-dependent transcriptional responses was first assessed with siRNAs by Tessa Gaarenstroom in the lab. In this experiment she measured the induction in response to Activin of some representative target genes associated with at least one SMAD2 binding site containing the motif for FOXH1 or POU5F1. From the plots in Figure 5.13A it was clear that the effects of knocking-down these TFs were gene dependent. The correct expression of *Lefty1* and *Eomes* required in fact both FOXH1 and POU5F1, whilst the induction of *Tdgf1* was dependent only on POU5F1. Activin-mediated transcription of *Pmepa1* instead did not required neither of the two co-factors, suggesting that the presence of a particular TF motif in the vicinity of a SMAD2 binding site did not necessary imply a functional role of the TF itself in regulating the expression of the target gene (Figure 5.13A). It can also be noted that for those transcripts affected by the knockdowns, the induction in response to Activin was severely impaired, but not completely lost. This could be due to the fact that *Foxh1* and *Pou5f1* mRNAs were still detectable upon silencing, although reduced to the levels generally expected of successful knockdowns using this technique (Figure 5.13B).

Since FOXH1 and POU5F1 both appeared to be relevant for the expression of some NODAL/Activin target genes we then asked if they had the same role in recruiting SMAD2 to its binding sites. First, Philip East averaged the position of FOXH1 and POU5F1 motifs relative to the summit of all SMAD2 peaks and found that the distribution of the motifs was different for the two TFs. For almost all the SBSs containing a FOXH1 motif, the sequences recognised by this TF were found to be just few bp away from the SMAD2 peak summit, with a maximum distance of 50 bp either side. Conversely, for POU5F1 the occurrences of its motif were spread uniformly across the 500 bp window, without a clear pattern (Figure 5.14).

This result suggested that FOXH1, rather than POU5F1, could play a more direct role in recruiting SMAD2 to its target sites. Evidence of the importance of FOXH1 for SMAD2 binding events came also from observing that the presence of the FOXH1 motif positively correlated with SMAD2 enrichment in both the 1 hr and 8 hr Activin samples (Figure 5.15A). Moreover, when considering the sites that were bound by SMAD2 within 1 hr of Activin treatment or only after 8 hr, the percentage of loci having a FOXH1 motif was found to be significantly higher in the first group of SBSs, than in the second (Figure 5.15B). Thus, the presence of sequences bound

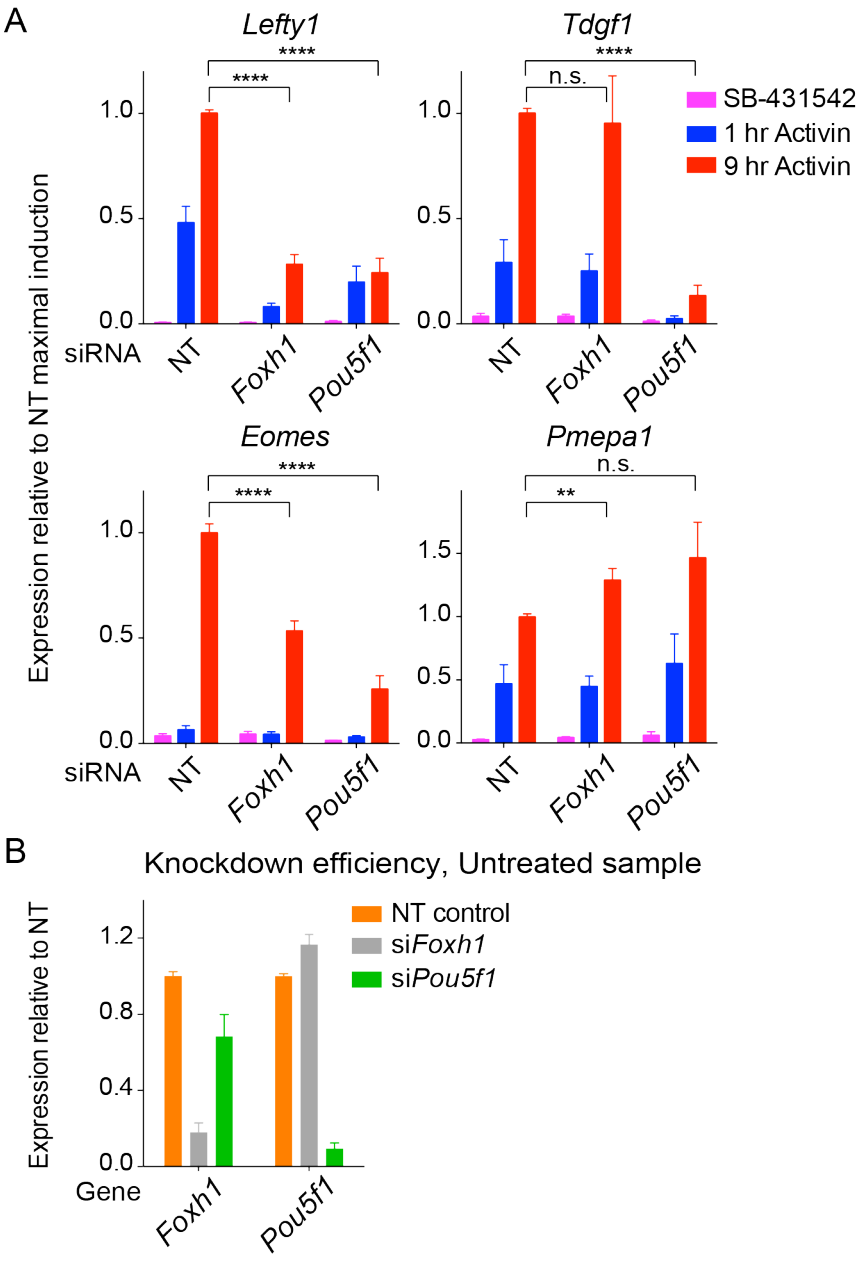


Figure 5.13. FOXH1 and POU5F1 are required for the induction of some SMAD2 target genes.
(A-B) P19 cells were transfected with siRNAs against either *Foxh1* or *Pou5f1*, or with a non-targeting control (NT). **(A)** cells were treated overnight with SB-431542, washed out, then either treated with SB-431542 for 1 hr or with Activin for the indicated times. The mRNA levels for the indicated genes were measured by qPCR, normalised relative to *Gapdh* and plotted relative to the maximal induction observed in the NT sample. Plotted are the means and SEM of three independent experiments performed in duplicate. **** corresponds to a p value of < 0.0001 and ** corresponds to a p value of < 0.01; n.s., not significant. **(B)** After transfection cells were left untreated for 72 hr and siRNAs knockdown efficiency was assayed by qPCR. *Foxh1* and *Pou5f1* levels were measured relative to *Gapdh* and displayed relative to their expression in the NT sample. Displayed are the means and SEM of three independent experiments performed in duplicate.

by FOXH1 was predictive of strong and rapid SMAD2 binding events in response to NODAL/Activin signalling.

Importantly, FOXH1 is part of the Forkhead family of TFs whose members frequently function as pioneer factors to recruit their transcriptional partners to closed sites on chromatin (Spitz and Furlong, 2012, Zaret and Carroll, 2011). For this reason, FOXH1 emerged as a good candidate for helping SMAD2 binding, particularly in the case of non-acetylated, nucleosome dense target sites. Thus, I set-out to first verify the requirement of FOXH1 for the binding of SMAD2 to selected loci, and then to address the sequence of events that occurred on chromatin at these sites.

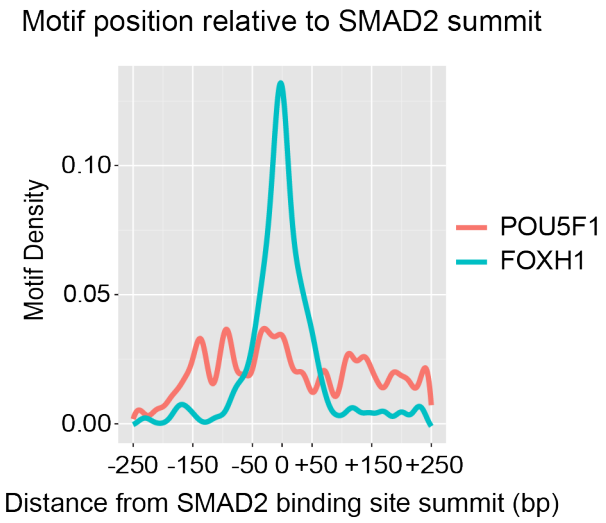


Figure 5.14. FOXH1 motifs, but not POU5F1 motifs, are generally close to SMAD2 peak summit.
Plotted is the distance of either FOXH1 or POU5F1 motifs (independently from DNA strand) to the summit of all the 757 MACS SMAD2 peaks mentioned in Figure 5.12.

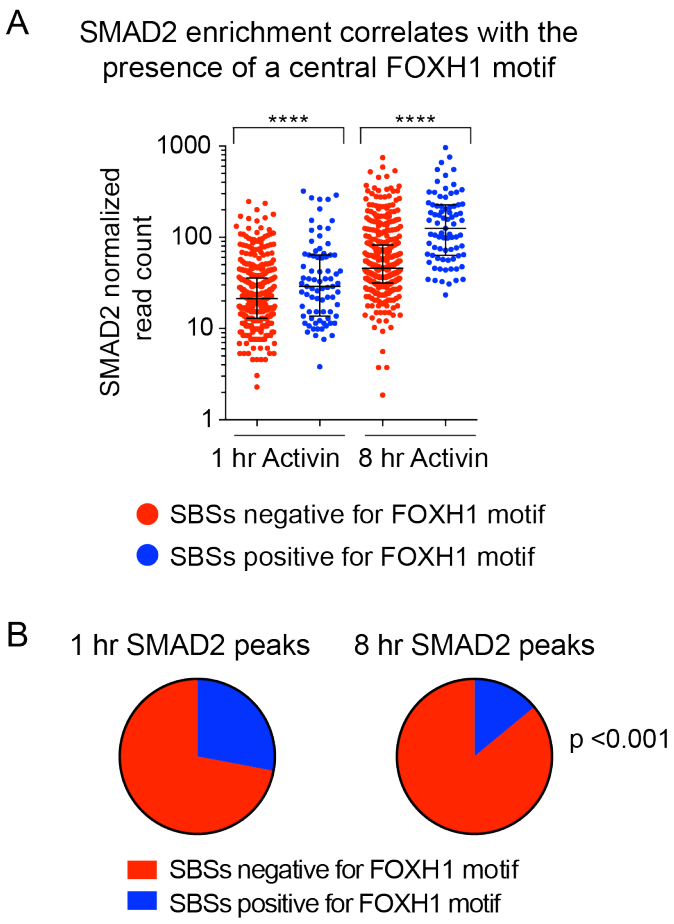


Figure 5.15. SMAD2 enrichment and acute SMAD2 binding correlate with the presence of FOXH1 motif.
(A) The 478 consensus SMAD2 peaks were segregated into FOXH1 negative and FOXH1 positive based on the presence of the FOXH1 motif in 50 bp either side of the consensus peak summit, as obtained by using FIMO software. SMAD2 binding occupancy over FOXH1 negative (red) and FOXH1 positive (blue) peaks is quantified at 1 hr and 8 hr of Activin treatment. For all samples the black bars indicate the mean and 95% confidence interval. **** corresponds to a p value of < 0.0001. **(B)** The 478 consensus SMAD2 peaks were divided into two groups based on the presence of a contributing MACS peak at 1hr or 8 hr of Activin treatment. For each group, the percentage of FOXH1 negative (red) or FOXH1 positive (blue) peaks is indicated. The presence of a FOXH1 motif was defined applying the same criteria described in **(A)**. An un-paired Chi-square test was performed on the data using a 95% confidence interval and the resulting p-value is reported on the graph.

5.2.7 FOXH1 is required for the induction of a subset of SMAD2 target genes

To functionally assess the role of FOXH1 with respect to SMAD2 binding and related chromatin modifications I decided to use the siRNA approach. Since the results described in Figure 5.13A were obtained using a pool of four different sequences against *Foxh1* mRNA, I first repeated the experiment using the individual siRNAs separately. In fact, it could not be excluded that the transcriptional effects observed by knocking-down *Foxh1* reflected on off-target activity of one or more siRNAs, rather than being the biological consequence of a lack of FOXH1.

Importantly, the four siRNAs were all able to efficiently silence *Foxh1* and to individually recapitulate the different effects on the expression of SMAD2 target genes observed when they had been pooled together (Figure 5.16). In all four cases *Lefty1* induction in response to 1 hr of Activin treatment was severely impaired, whilst in the case of *Pmepa1*, no significant differences were observed for any of the siRNAs compared with the control, as expected (Figure 5.16). In order to increase the relevance of the studies I was proposing to perform, the expression of others target genes was also tested. In particular, I focused on two NODAL/Activin targets, *Lefty2* and *Pitx2,* which could potentially be affected by FOXH1 knockdown, since they both had SBSs positive for FOXH1 motif. Furthermore, as a negative control, I included in the analysis a gene, *Smad7*, which has no FOXH1 motifs within its SBSs. As shown in Figure 5.16, *Lefty2* and *Pitx2* induction following acute Activin stimulation was significantly dependent on FOXH1 regardless of the individual siRNA used. As expected, *Smad7* mRNA levels were not affected by knocking down *Foxh1* (Figure 5.16).

Overall, for the genes examined the results obtained with the 4 different siRNAs against *Foxh1* were almost identical to one another, ruling out the possibility of technical artefacts. I therefore decided to perform the experiments described in the next section using the pool of the 4 siRNAs.

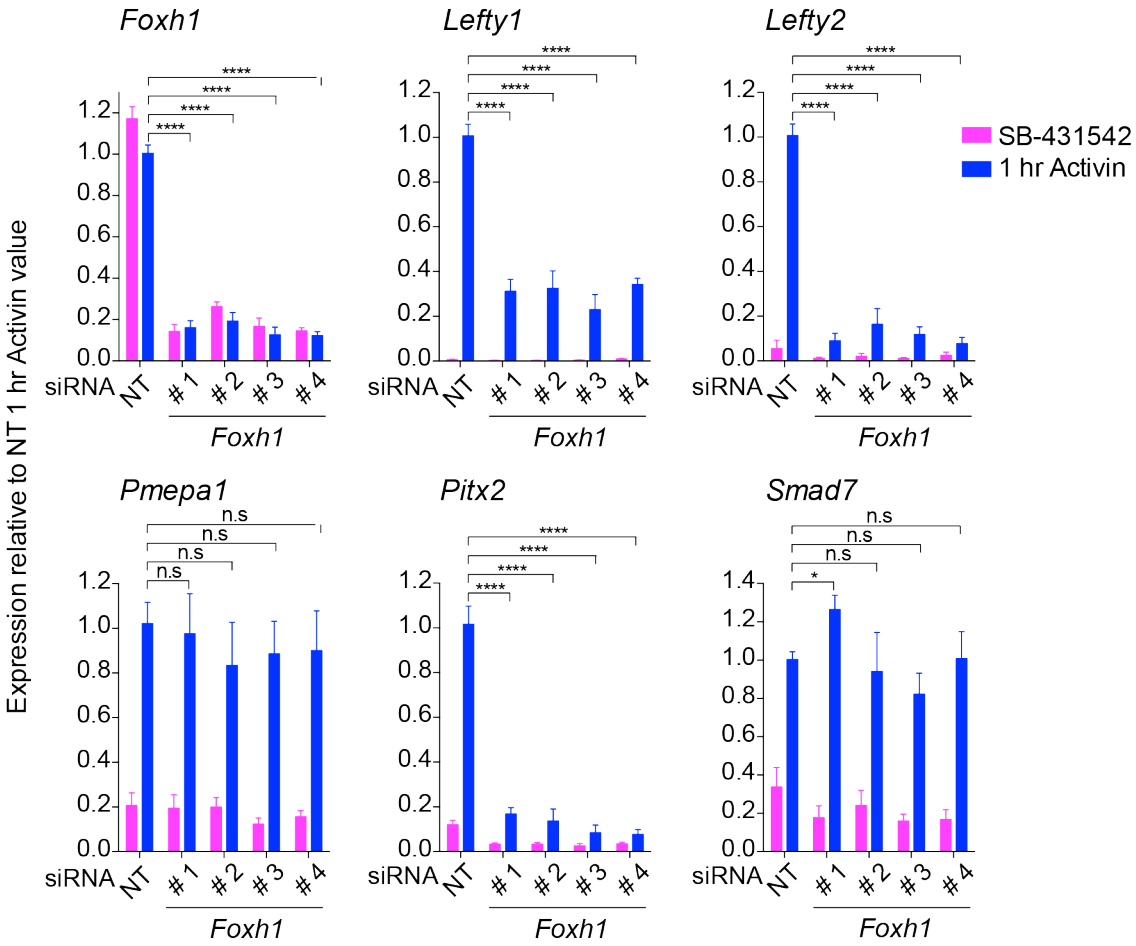


Figure 5.16. Individual siRNAs against *Foxh1* are comparable to one another and recapitulate the results obtained using the siRNA pool.
P19 cells were transfected with four individual siRNAs directed against *Foxh1* (referred to as # 1, # 2, # 3 or # 4) along with a non-targeting control (NT). Following signal inhibition or Activin induction, qPCR was performed for the genes shown. The values obtained were normalised to *Gapdh* and plotted relative to the NT 1 hr Activin sample. Plotted are the means and SEM of three independent experiments each performed as biological replicates. **** corresponds to a p value of <0.0001 and * corresponds to a p value of 0.1. n.s., not significant.

5.2.8 FOXH1 mediates SMAD2 binding and chromatin remodelling at a subset of target loci

I next asked whether FOXH1 was required for SMAD2 binding to target sites. I therefore performed ChIP-PCR experiments for SMAD2 at different loci in the context of an efficient FOXH1 knockdown (Figure 5.17B). As expected, in the control condition SMAD2 was detected at all the SBSs analysed in an Activin inducible manner. However, when *Foxh1* was silenced, the binding of SMAD2 upon 1 hr of Activin treatment was significantly reduced for *Lefty1* SBS, *Lefty2* SBS and *Pitx2* SBS (Figure 5.17A). This result is not surprising since I showed that the induction of these genes in response to Activin is dependent on FOXH1. As expected, SMAD2 enrichment at the SBSs of FOXH1 independent genes such as *Pmepa1*, *Smad7* and *Pou5f1* was not affected by its knockdown (Figure 5.17A). Thus, I concluded that FOXH1 was required for SMAD2 binding to a subset of target sites.

I then asked if the chromatin remodelling events observed at these loci in response to NODAL/Activin signalling were also dependent on FOXH1 presence. This seemed likely, since I had shown that the changes in the chromatin landscape required the presence of bound SMAD2. I therefore performed ChIP-PCR and FAIRE-PCR experiments to verify if H3K27ac and nucleosome displacement at different loci were affected by FOXH1 knockdown (Figure 5.18 and Figure 5.19). Indeed, the acute induction of H3K27 acetylation observed around *Lefty1* and *Lefty2* SBSs in the control condition was severely perturbed in the absence of FOXH1. In contrast, signal inducible H3K27ac was not affected for the nucleosomes flanking *Pmepa1* and *Smad7* SBSs, consistent with the fact that SMAD2 biding to these loci did not require FOXH1 (Figure 5.18A-B). When the region around the *Lefty1* SBS was assayed for enrichment in open chromatin, it was clear that Activin-induced nucleosome displacement at the SBS was also substantially affected by silencing *Foxh1*. Importantly, the nucleosome eviction in response to Activin signalling was not reduced at the *Pmepa1* SBS and constitutively open chromatin was observed at *Pou5f1* SBS across all conditions (Figure 5.19A-B). The last findings were in line with the fact that Activin-mediated induction of *Pmepa1* and *Pou5f1* did not require the presence of FOXH1.

Taken together, these results clearly confirmed that the transcriptional dependency on FOXH1 for some Activin induced genes was due to the crucial role played by this TF in directing SMAD2 to its target sites on chromatin.

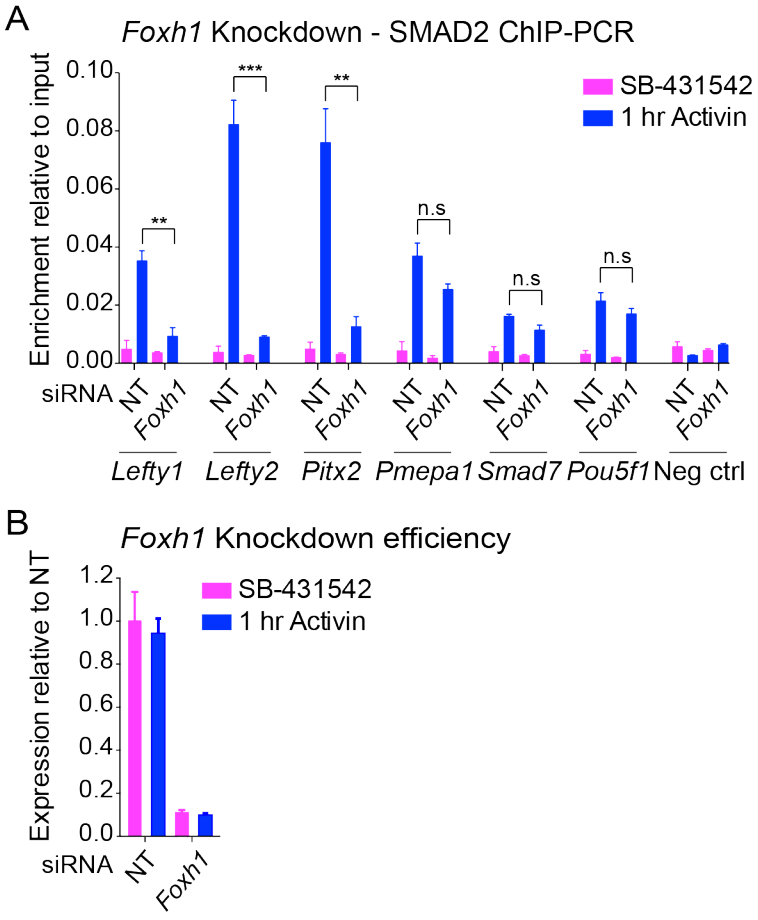


Figure 5.17. FOXH1 is required for SMAD2 binding at a subset of target loci.
(A) P19 cells were transfected with a non-targeting control (NT) or with an siRNA pool against *Foxh1*. Following signal inhibition or Activin induction, SMAD2 ChIP-PCR was performed at the SBSs indicated and over a negative control region (Neg cntrl). Plotted are the means and SEM of three independent experiments performed in duplicate. *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01; n.s., not significant. **(B)** For the cells used in (A), the knockdown efficiency was tested by measuring *Foxh1* levels in a qPCR assay. The values obtained were normalised to *Gapdh* and plotted relative to the NT 1 hr Activin sample. Since the same cells from (A) were also used for the experiments in Figure 5.18 and 5.19, the knockdown efficiency reported in (B) is valid also for these figures.

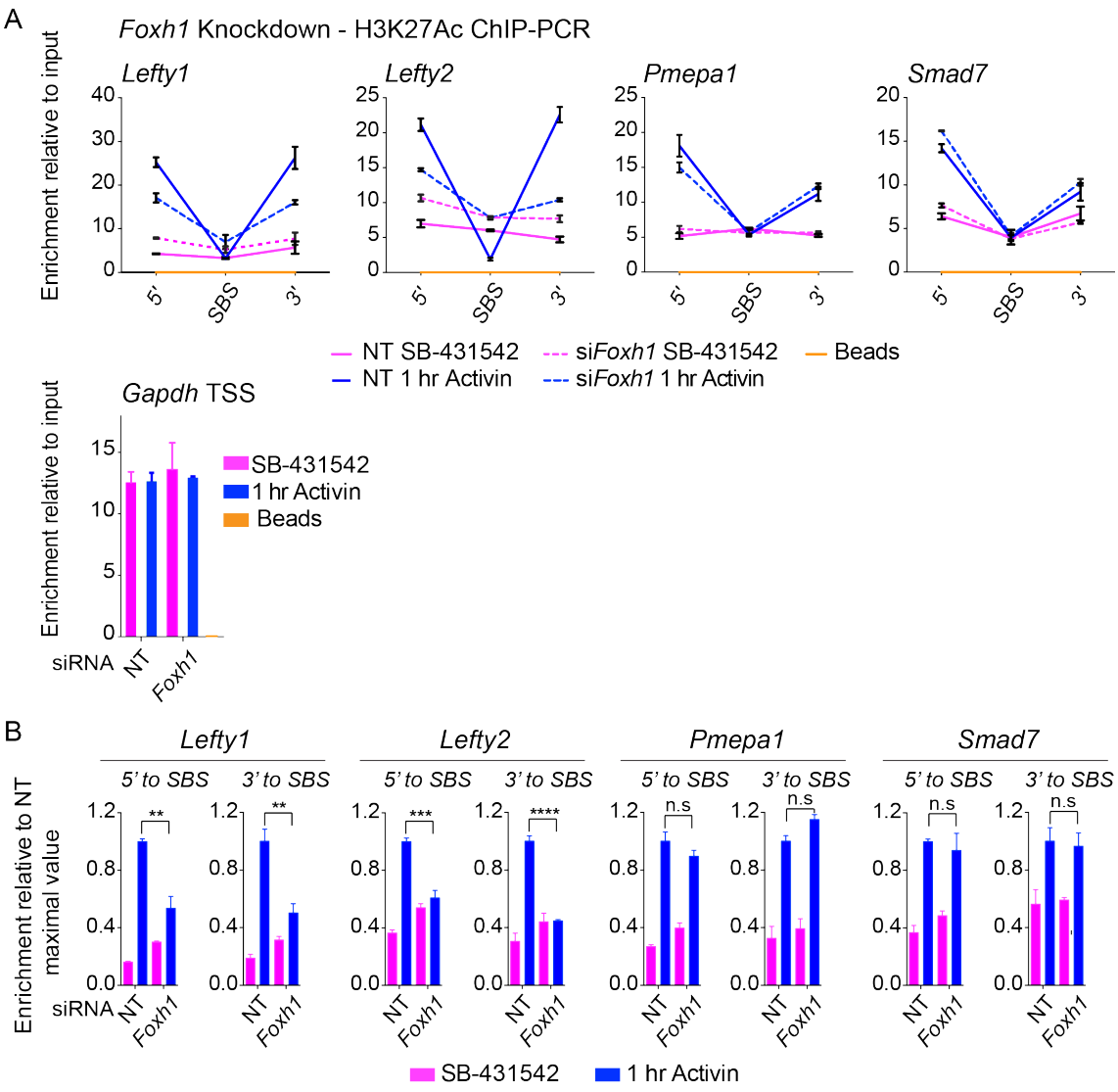


Figure 5.18. At a subset of SMAD2 target loci, the induction of H3K27ac in response to NODAL/Activin signalling depends on FOXH1.
(A) P19 cells were transfected and treated as described in Figure 5.17. H3K27ac ChIP-PCR was performed at the SBSs and flanking sites indicated. As a control, the signals obtained for H3K27ac over *Gapdh TSS* is displayed in the bottom panel. A representative experiment out of three is shown (means ± SD). **(B)** Three independent experiments were carried out as in (A), and the average of H3K27ac enrichment at the nucleosomes either side of the reported SBSs is displayed relative to the averaged NT maximal value. Plotted are the means and SEM. **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01. n.s not significant.

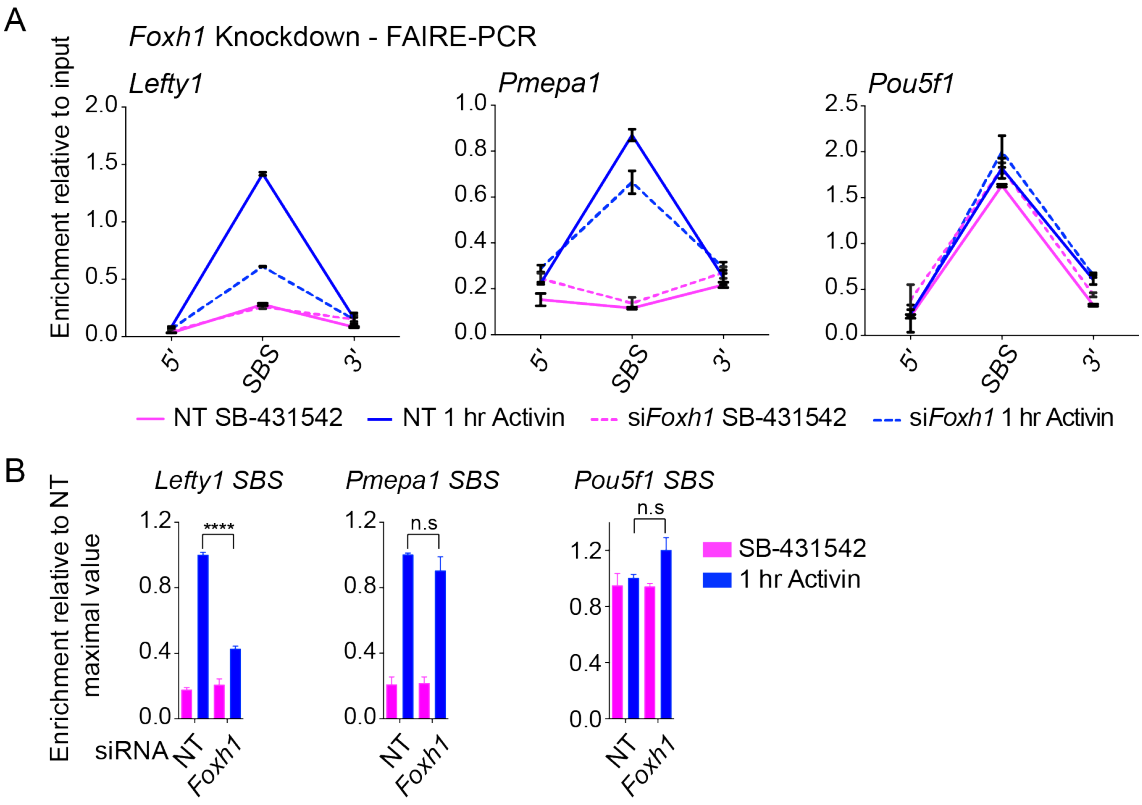


Figure 5.19. FOXH1 is necessary for Activin-induced nucleosome eviction at the *Lefty1* SBS, but not at the *Pmepa1* SBS.
(A) P19 cells were transfected and treated as described in Figure 5.17. FAIRE-PCR was performed and the enrichment of open chromatin for the SBSs and flanking sites indicated was measured. A representative experiment out of three is shown (means ± SD). **(B)** Three independent experiments were carried out as in (A), and the average nucleosome occupancy at the reported SBSs is displayed relative to the averaged NT maximal value. Plotted are the means and SEM. **** corresponds to a p value of < 0.0001; n.s not significant.

5.2.9 FOXH1 does not act as a pioneer factor, but it binds to DNA together with SMAD2

When reasoning about possible mechanisms of interaction between FOXH1 and SMAD2 on chromatin, two different scenarios seemed plausible. As anticipated before, FOXH1 could act as a pioneer-factor stably bound to occluded sites to facilitate the subsequent recruitment of SMAD2. Alternatively, a cooperative model of DNA binding could be hypothesized, with FOXH1–SMAD2 complexes targeting closed chromatin devoid of any pre-bound TFs (Reiter et al., 2017). To discriminate between these two scenarios, it was therefore necessary to asses FOXH1 occupancy at distinct SBSs in different signalling conditions. In fact, if FOXH1 was

indeed recruiting SMAD2 to its target sites according to the pioneer-factor model, chromatin-bound FOXH1 should be detected at these loci even in the absence of NODAL/Activin signalling. Performing ChIP-PCR for FOXH1 over a time course of Activin treatment seemed the straightforward approach to address this question, but it was complicated by the fact that there are no commercial antibody against mouse FOXH1. I therefore decided to generate one (see below).

In the meantime, I did some preliminary experiments in a P19 cell line previously engineered by Thodoris Petrakis in the lab to stably express MYC-tagged FOXH1. In fact, I reasoned that the presence of the tag would allow me to easily carry-out ChIP-PCR experiments using a validated anti MYC antibody. Nevertheless, it was first necessary to verify that the introduction of MYC-FOXH1 in P19 did not alter the cell responses to NODAL/Activin signalling. As shown in Figure 5.20, SMAD2 phosphorylation levels in MYC-FOXH1 P19s for the untreated condition and over a time course of Activin treatment were not different from those observed in WT P19. Also, when the blot was probed with anti MYC antibody, a band at the predicted molecular weight of FOXH1 was detected in MYC-FOXH1 cells, but not in WT cells, thus confirming the presence of the tagged protein (Figure 5.20).

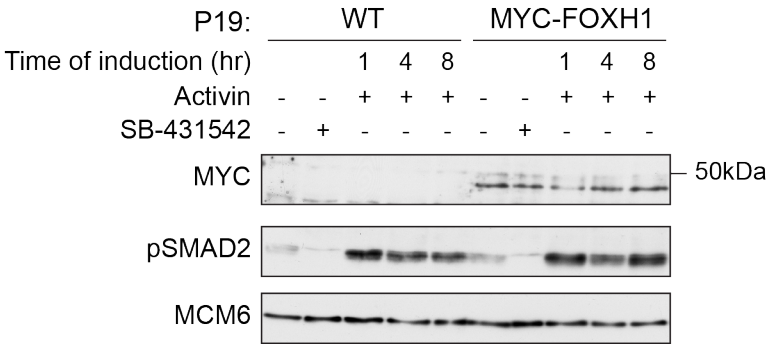


Figure 5.20. The dynamics of SMAD2 phosphorylation in response to NODAL/Activin signalling are conserved in MYC-FOXH1 P19 cells compared with WT P19 cells.
WT P19 cells or P19 cells stably expressing MYC-FOXH1 were treated with 10 μM SB-431542 overnight or untreated; after which the SB-431542 was washed out and the cells were incubated with 10 μM SB-431542 or 20 ng/ml Activin for the times indicated. Immunoblots of cell extracts were probed with antibodies against pSMAD2, MYC and MCM6 as a loading control. The 50 kDa protein marker is shown as a reference indicator for the molecular weight.

Similarly, the different patterns of transcriptional responses downstream of NODAL/Activin signalling were also conserved in MYC-FOXH1 P19. No substantial differences were noticed when the expression of some SMAD2 target genes belonging to distinct kinetic categories was compared to WT cells (Figure 5.21). It is important to highlight that the amount of *Foxh1* mRNA in MYC-FOXH1 P19s was actually just twice as much than in WT P19s, thus suggesting that MYC-FOXH1 was expressed in P19 at near-endogenous levels (Figure 5.21). This observation gave me confidence in using this cell line to assay MYC-FOXH1 binding to chromatin.

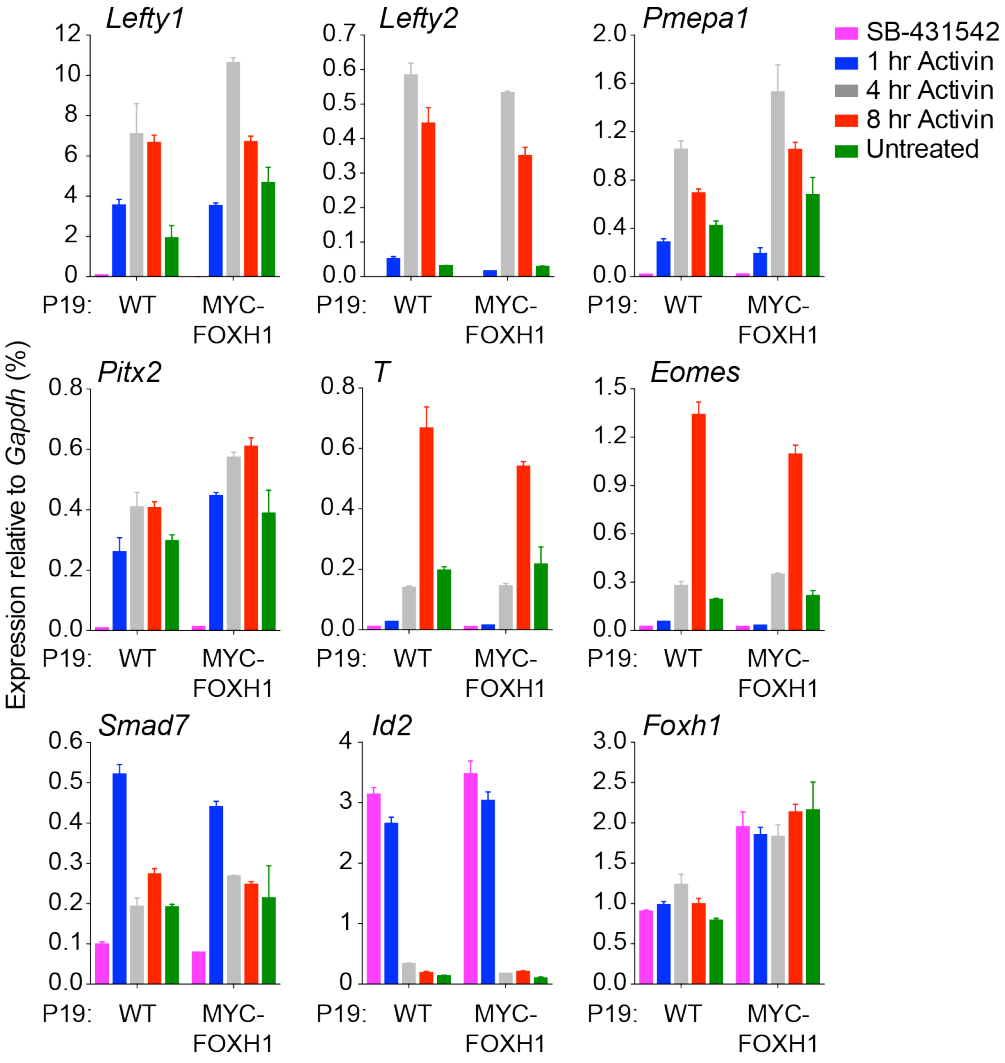


Figure 5.21. The kinetics of the transcriptional responses to NODAL/Activin signalling is not altered in MYC-FOXH1 P19 cells.
For the same samples described in Figure 5.19, the expression of a subset of NODAL/Activin target genes was assayed by qPCR. Note that *Foxh1* levels were also
Figure 5.21 continued on next page.

Figure 5.21 continued.
measured in all samples. A representative experiment (means ± SD) is shown. Untr, untreated

---

I therefore performed ChIP-PCR for MYC over a time course of Activin treatment and analysed the signals at SMAD2 binding sites of FOXH1 dependent and independent genes. Strikingly, very little signal was observed at *Pitx2*, *Lefty1* and *Lefty2* SBSs in the SB-431542 condition, but it was highly enriched at the same loci in the 1 hr and 8 hr Activin samples (Figure 5.22). Thus, for those sites where SMAD2 recruitment was impaired by FOXH1 knockdown, binding of MYC-FOXH1 was induced by Activin. This result suggested that at least at the loci examined FOXH1 was not acting as a pioneer factor. Almost no MYC-FOXH1 was detected at any time at *Pmepa1* and *Pou5f1* SBSs, as expected by the fact that these genes did not require FOXH1 for their expression in response to Activin. The extremely low signal measured over the intergenic region used as negative control confirmed the specificity of the antibody used in the experiment (Figure 5.22).

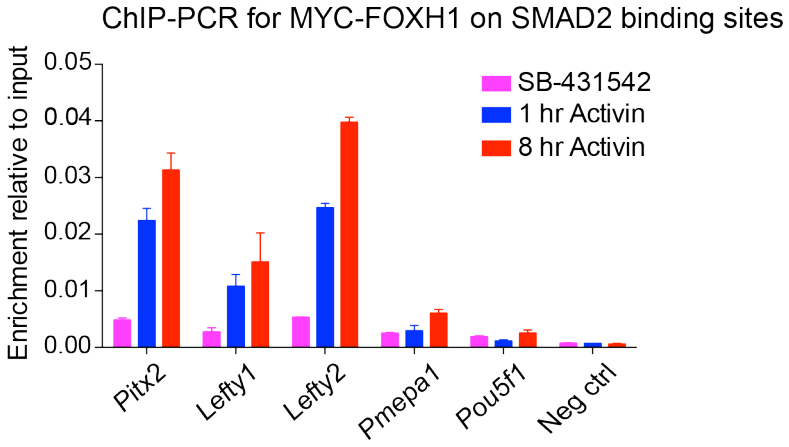


Figure 5.22. MYC-FOXH1 binding at a subset of SMAD2 target sites is Activin signalling dependent.
P19 cells stably expressing MYC-FOXH1 were treated overnight with SB-431542, washed out, then either treated with SB-431542 for 1 hr or with Activin for the indicated times. ChIP-PCR using an anti-MYC antibody was performed and the signals for the SBSs indicated and the negative control region (Neg cntrl) are displayed. A representative experiment (means ± SD) is shown.

In parallel, I set out to raise an antibody against mouse FOXH1 in collaboration with the peptide synthesis facility and the animal facility of the Francis Crick Institute (For details, see Section 2.3.3). After column purifying the antibody from rabbit blood I first tested it for Western Blot on lysates from MYC-FOXH1 P19 or WT P19 transfected or not with siRNA against FOXH1. As shown in Figure 5.23, the FOXH1 antibody failed to recognise endogenous FOXH1 in WT cells on a Western Blot. However, in the MYC-FOXH1 sample a band was detected with the antibody at the same molecular weight at which the signal from MYC antibody was also observed (Figure 5.23). I therefore concluded that the raised antibody was able to recognise FOXH1 in the context of over-expression. Considering this result, the absence of bands for the WT samples could likely be explained as a lack of sensitivity of the FOXH1 antibody in Western Blot rather than being due to its failure to bind FOXH1 per se.

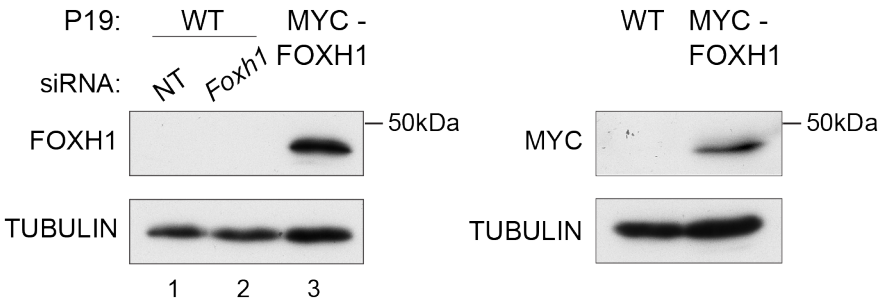


Figure 5.23. The in-house FOXH1 antibody recognises overexpressed FOXH1 in Western blot analysis.
Whole cell extracts of WT P19 transfected with a non-targeting control (NT) or with an siRNA pool against *Foxh1* where loaded on a gel alongside with lysates from P19 stably expressing MYC-tagged FOXH1. Protein levels for FOXH1 and TUBULIN as a loading control were assayed by Western blot (left panel). The samples in lane 1 and 3 were also blotted separately and the membrane was incubated with the anti MYC antibody (right panel). The 50 kDa protein marker is reported as a reference indicator for molecular weight.

Confident in the specificity of the antibody raised against mouse FOXH1, I then performed ChIP-PCR in WT P19 cells to measure FOXH1 binding at selected SMAD2 target sites over an Activin time course. Importantly, I was able to reproduce the results obtained with the MYC antibody in cells expressing MYC tagged FOXH1 (compare Figure 5.22 with Figure 5.24). Endogenous FOXH1 was in fact bound at SBSs of the FOXH1 dependent genes *Pitx2*, *Lefty1* and *Lefty2* only in response to acute or prolonged Activin signalling. As expected, very little enrichment was

reported over the SBSs of the FOXH1-independent genes *Pmepa1* and *Pou5f1*, and also over the negative control region (Figure 5.24).

Overall, this experiment provided with the clear evidence that, at least on a subset of SMAD2 target sites, FOXH1 does not act as a pioneer factor, but it binds to closed chromatin together with SMAD2. The result also suggested that FOXH1 and SMAD2 possibly use a cooperative-like mode of DNA binding, with NODAL/Activin signalling being necessary for FOXH1 to target SMAD binding sites, similarly to what has been previously observed for the *Xenopus* FOXH1 isoform xFAST3 using *in vitro* DNA binding assays (Howell et al., 2002).

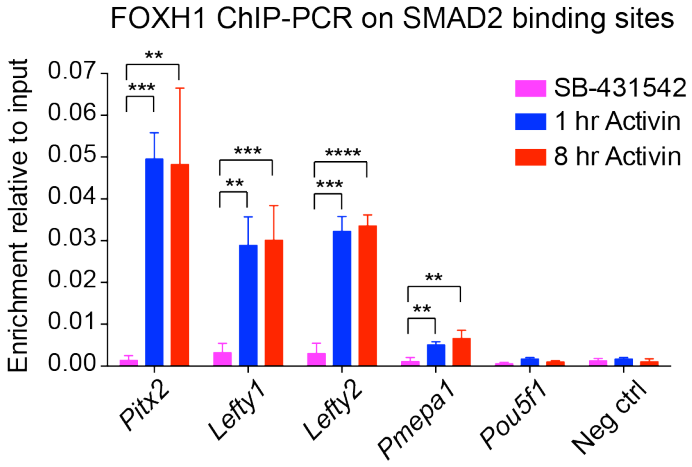


Figure 5.24. FOXH1 binds to a subset of SMAD2 target genes in response to NODAL/Activin signalling.
FOXH1 ChIP-PCR on P19 cells treated as indicated. FOXH1 enrichment was measured at the SBSs indicated and over a negative control region (Neg ctrl). Plotted are the means and SEM of two independent experiments performed in duplicate. **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01; n.s., not significant.

5.2.10 The ATP-dependent helicase SMARCA4 is required for the expression of some NODAL/Activin target genes

Thus far, I have characterised the sequence of events occurring on chromatin in response to NODAL/Activin signalling. I showed that FOXH1–SMAD2 complexes can bind to closed chromatin, inducing nucleosome displacement and H3 acetylation. However, although I provided evidence of EP300 being responsible for the NODAL/Activin dependent histone modifications, the mechanism of nucleosome eviction in response to the signalling has not been addressed. The ATP-dependent

helicase SMARCA4, also known as BRG1, represented a good candidate to be involved in this process. SMARCA4 is part of the large SWI-SNF nucleosome remodelling complex (Lange et al., 2011) and it has previously been shown to mediate the induction of many TGF-β target genes (Xi et al., 2008, Ross et al., 2006). Nevertheless, a mechanistic role for SMARCA4 in TGF-β regulated transcription was not established.

To explore this, I first set out to verify if SMARCA4 knockdown affected NODAL/Activin target gene expression in the context of P19 cells. I used a combination of the same siRNAs already employed by Ross et al., 2006, and measured the acute induction of either the 'baseline off' genes *Lefty1*, *Lefty2* and *Pmepa1* or the 'baseline on' genes *Pitx2* and *Smad7*. As shown in Figure 5.25, the effect of silencing SMARCA4 was highly gene specific. For *Lefty1*, *Lefty2* and to a lesser extent *Pmepa1*, the transcriptional response to 1 hr of Activin treatment was severely impaired. Conversely, the induction of *Pitx2* and *Smad7* was not significantly affected (Figure 5.25).

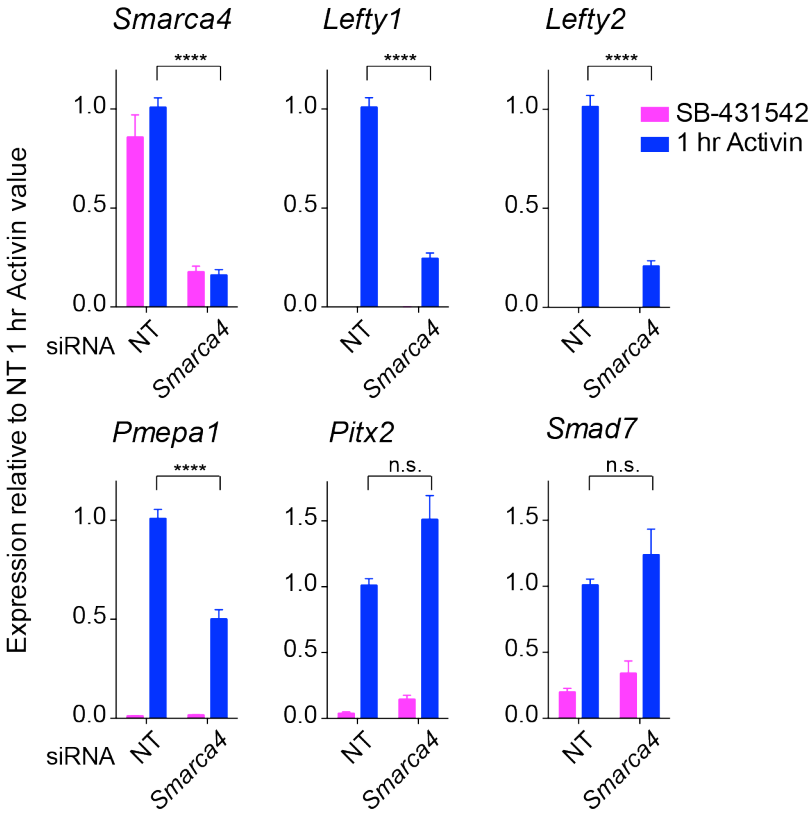


Figure 5.25. See next page for legend.

Figure 5.25. The induction of some SMAD2 target genes in response to NODAL/Activin signalling requires SMARCA4.
P19 cells were transfected with a non-targeting control (NT) or with an siRNA pool against *Foxh1*. Following signal inhibition or Activin induction, qPCR was performed for the genes shown. The values obtained were normalised to *Gapdh* and plotted relative to the NT 1 hr Activin sample. Note that the same samples described here were subsequently used for the experiments described in Figure 5.27, Figure 5.28 and Figure 5.29. Plotted are the means and SEM of four independent experiments each performed as biological replicates. **** corresponds to a p value of <0.0001. n.s., not significant.

Following the same rationale described for the FOXH1 knockdown experiments, I then tested separately the two siRNAs against SMARCA4, which I used as a pool in the experiment just described. It can be noted that sequence # 1 was less efficient in knocking down SMARCA4 than sequence # 3, thus less dramatic effects were observed with sequence # 1 compared to sequence # 3 (Figure 5.26). Nevertheless, the consequences on target gene induction obtained with the individual siRNAs were overall comparable with each other and well recapitulated the results observed when the siRNAs were transfected as a pool (Figure 5.26). In subsequent experiments I used a combination of the two single siRNAs against SMARCA4 as a pool.

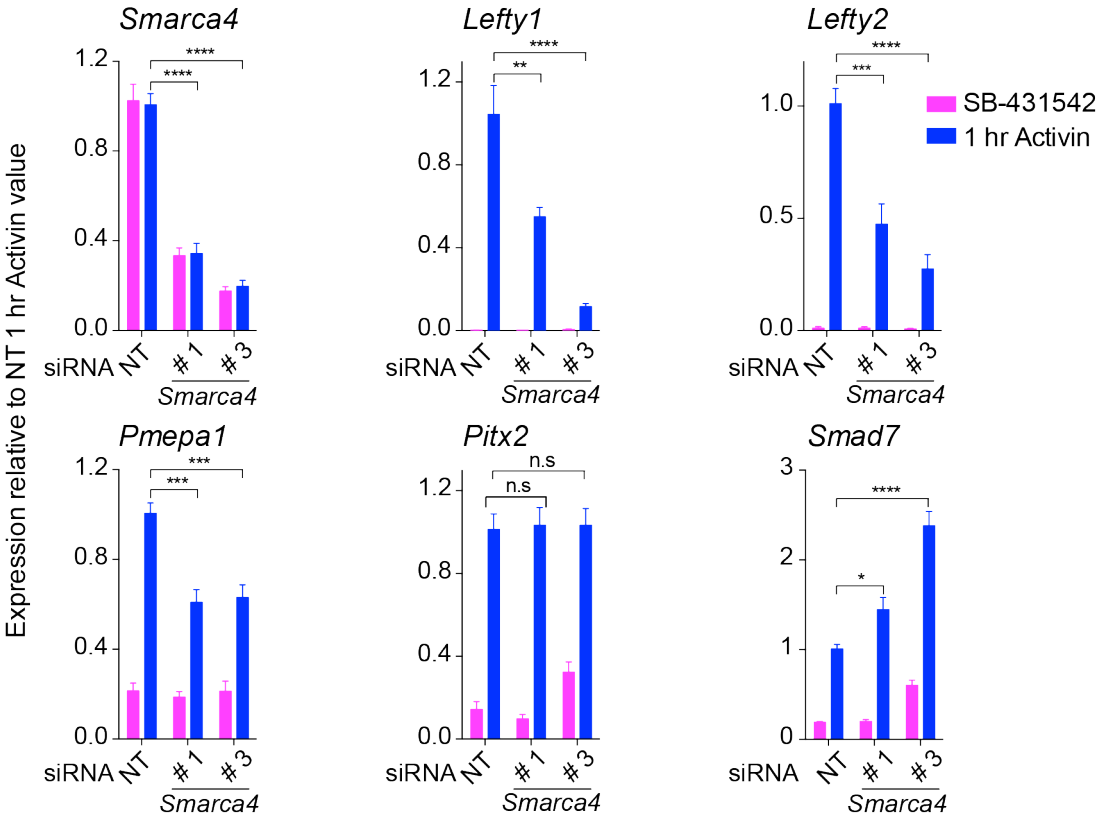


Figure 5.26. on previous page.

Figure 5.26. Individual siRNAs against S*marca4* recapitulate the results obtained using the siRNA pool.

P19 cells were transfected with two individual siRNAs directed against *Smarca4* (referred to as # 1 or # 3) along with a non-targeting control (NT). Following signal inhibition or Activin induction, qPCR was performed for the genes shown. The values obtained were normalized to *Gapdh* and plotted relative to the NT 1 hr Activin sample. Plotted are the means and SEM of three independent experiments each performed as biological replicates. **** corresponds to a p value of <0.0001; *** corresponds to a p value of <0.001; ** corresponds to a p value of <0.01 and * corresponds to a p value of 0.1. n.s., not significant.

5.2.11 For a subset of closed SMAD2 target sites, SMAD2-induced nucleosome displacement requires SMARCA4 activity

I next wanted to address mechanistically the transcriptional effects of silencing SMARCA4, asking if the SMAD2 binding and subsequent chromatin remodelling were dependent on SMARCA4 at some target sites. I therefore knocked down SMARCA4 and performed a series of ChIP-PCR and FAIRE-PCR experiments, focusing on the SBSs associated with the genes analysed above. Indeed, SMAD2 enrichment at *Lefty1*, *Lefty2* and *Pmepa1* SBSs was impaired in the absence of SMARCA4, thus explaining the lack of induction observed for these 'baseline off' genes in Figure 5.25 (Figure 5.27). In contrast, SMARCA4 was not required for SMAD2 binding at the SBSs of the 'baseline on' genes *Pitx2* and *Smad7*, as expected from the transcriptional results described above (Figure 5.27).

Importantly, at those loci where SMAD2 binding required SMARCA4, H3 acetylation and nucleosome occupancy were also perturbed. ChIP-PCR assays showed that the induction of H3K27ac in response to 1 hr Activin was severely reduced at the sites flanking *Lefty1*, *Lefty2* and *Pmepa1* SBSs, but not at those either side of *Pitx2* and *Smad7* SBSs (Figure 5.28A-B). In addition, FAIRE-PCR experiments revealed that SMARCA4 was also necessary for Activin-mediated nucleosome displacement at the *Lefty1*, *Lefty2* and *Pmepa1* SBSs. In contrast, the accessibility of *Pitx2* and *Smad7* SBSs was clearly less affected by silencing SMARCA4 (Figure 5.29A-B).

Therefore, I concluded that SMARCA4 is necessary for SMAD2 binding to a subset of closed sites, likely because it directly elicits the chromatin remodelling events observed at these loci in response to signalling.

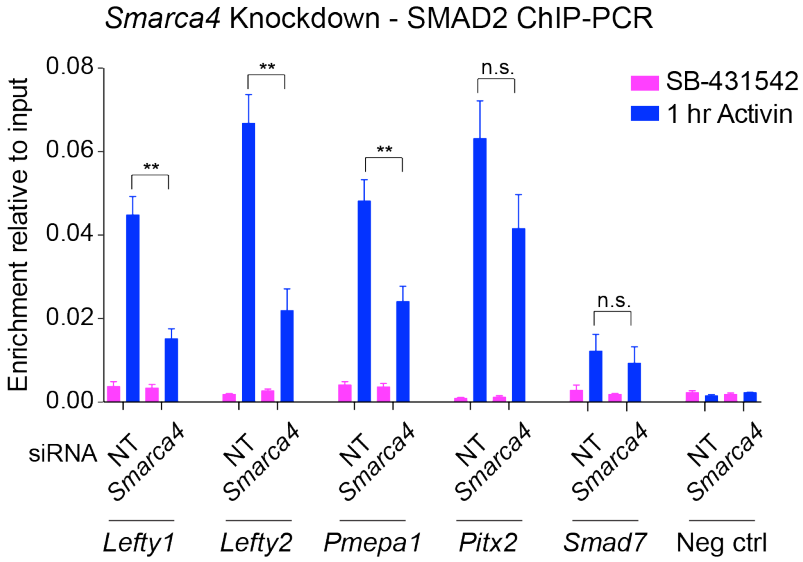


Figure 5.27. SMAD2 binding at a subset of target loci is reduced upon SMARCA4 knockdown.
P19 cells were transfected with a non-targeting control (NT) or with an siRNA pool against *Smarca4*. Following signal inhibition or Activin induction, SMAD2 ChIP-PCR was performed at the SBSs indicated and over a negative control region (Neg cntrl). Plotted are the means and SEM of four independent experiments performed in duplicate. ** corresponds to a p value of < 0.01; n.s., not significant.

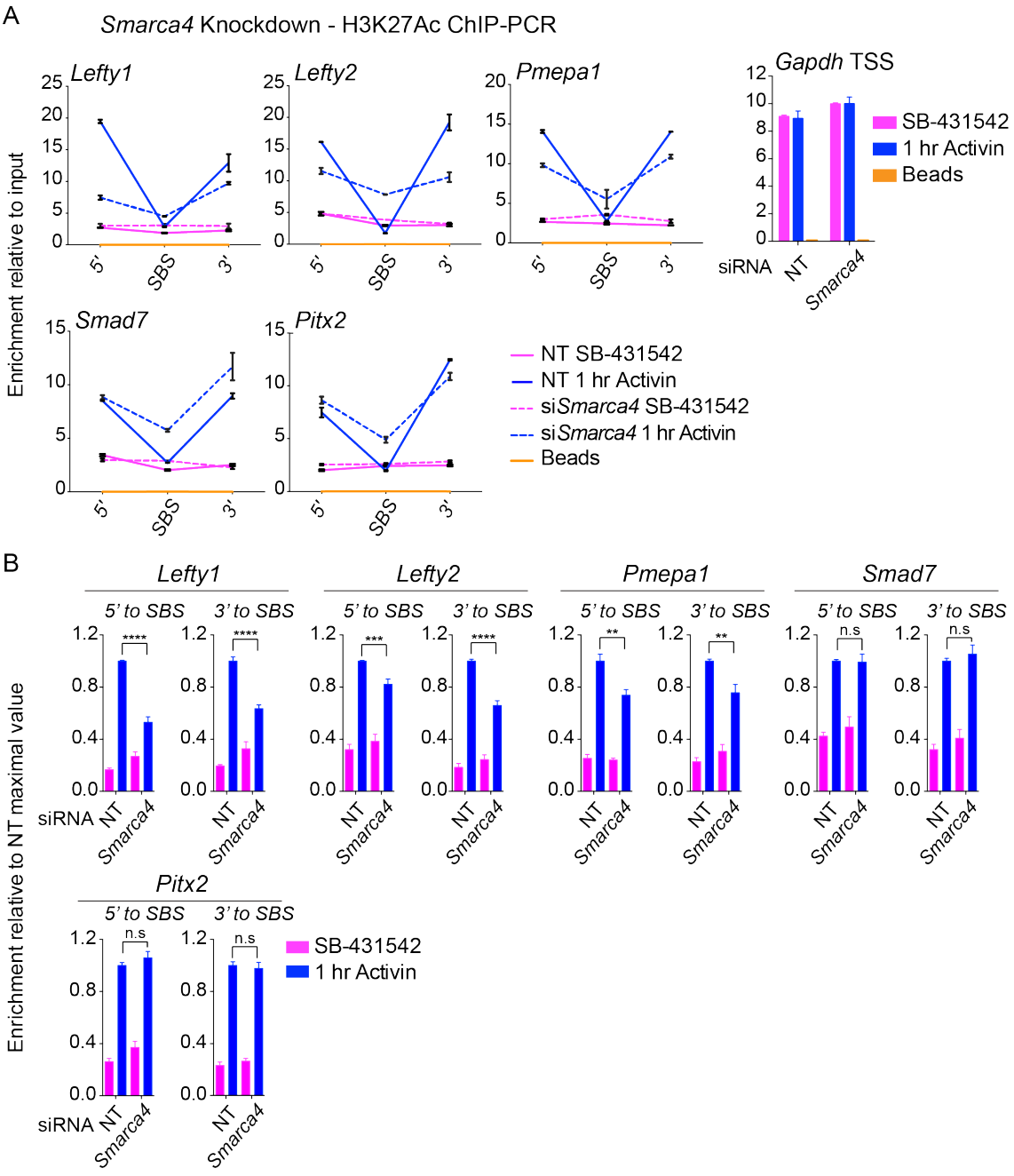


Figure 5.28. For a subset of SMAD2 target loci, the induction of H3K27ac in response to NODAL/Activin signalling requires SMARCA4.
(A) P19 cells were transfected and treated as described in Figure 5.26. H3K27ac ChIP-PCR was performed at the SBSs and flanking sites indicated. As a control, the signals obtained for H3K27ac over *Gapdh TSS* is displayed in the right panel. A representative experiment out of three is shown (means ± SD). **(B)** Three independent experiments were performed as in (A), and the average of H3K27ac enrichment at the nucleosomes either side of the reported SBSs is displayed relative to the averaged NT maximal value. Plotted are the means and SEM. **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.01. n.s not significant.

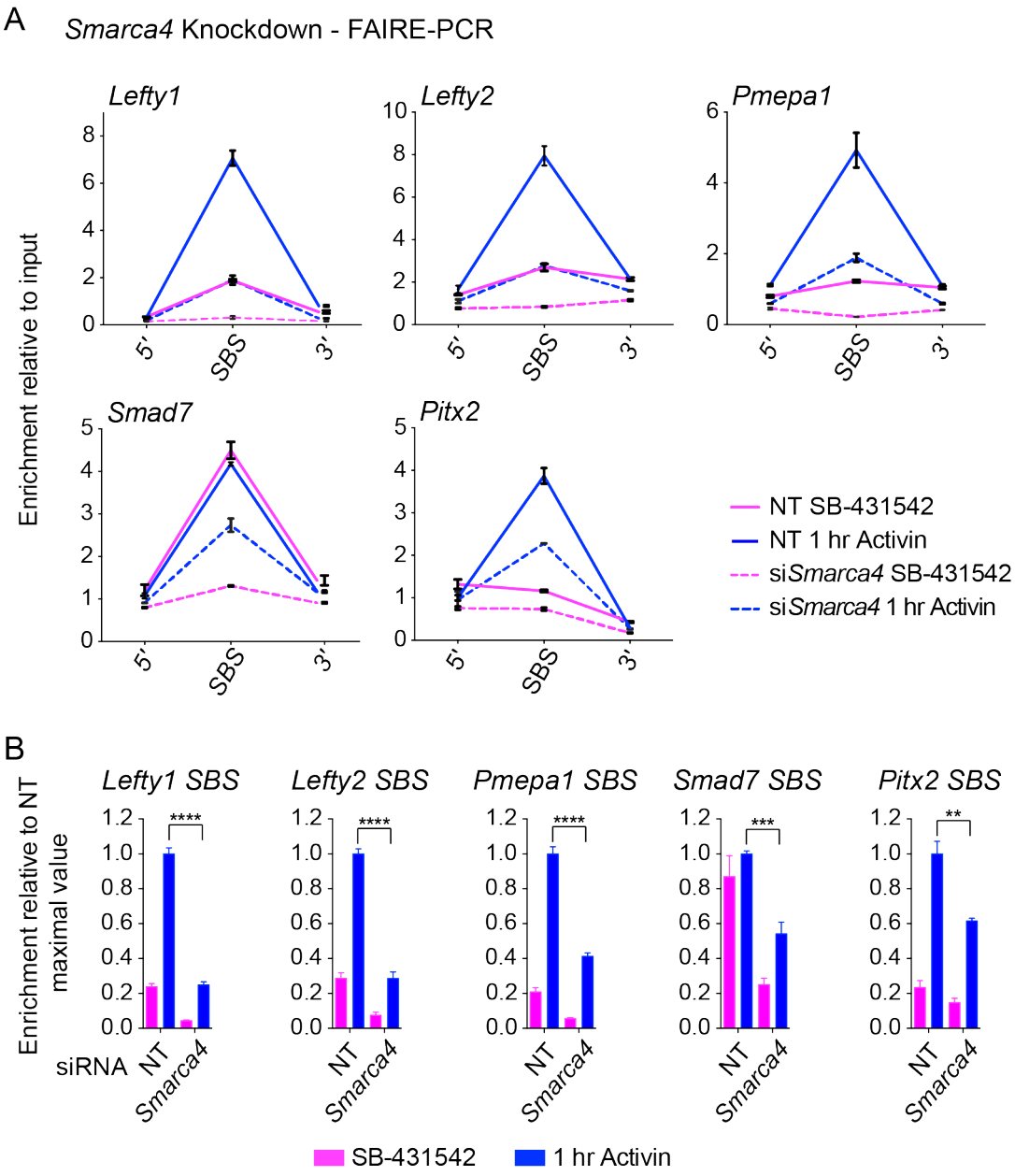


Figure 5.29. SMAD2-mediated nucleosome displacement requires SMARCA4 activity.
P19 cells were transfected and treated as described in Figure 5.27. FAIRE-PCR was performed and the enrichment in open chromatin for the SBSs and flanking sites indicated was measured. A representative experiment out of three is shown (means ± SD). **(B)** Three independent experiments were performed as in (A), and the average of nucleosome occupancy at the reported SBSs is displayed relative to the averaged NT maximal value. Plotted are the means and SEM. **** corresponds to a p value of < 0.0001; *** corresponds to a p value of < 0.001 and ** corresponds to a p value of < 0.0001.

5.3 Discussion

5.3.1 Summary of main findings

- The ChIP-seq experiments for H3K9ac, H3K27ac and total H3 reveal that NODAL/Activin signalling induces changes in the chromatin landscape at the level of target genes and associated SMAD2 binding sites.
- H3 acetylation changes at the TSSs of NODAL/Activin target genes correlate with the gene expression dynamics over time.
- SMAD2 has two different modes of binding: in the case of 'baseline on' genes it binds to open, pre-acetylated chromatin (eg. *Pou5f1*), whereas in the case of 'baseline off' genes it binds to closed, non-acetylated sites (eg. *Lefty1*). In the second scenario, SMAD2 binding induces H3 acetylation and nucleosome displacement. These chromatin loci may be constitutively marked by H3K4me1.
- The expression of a subset of 'baseline off' genes requires FOXH1. The binding of SMAD2 to the SBSs regulating these genes, and the downstream chromatin remodelling events are also dependent on FOXH1.
- FOXH1 does not act as a pioneer factor to recruit SMAD2 to closed chromatin, but it binds to these sites together with SMAD2 in an NODAL/Activin inducible manner.
- The SWI/SNF ATPase SMARCA4 is required for H3K27ac and nucleosome displacement at a subset of closed SMAD2 target sites. As a result, SMAD2 binding at these loci and the induction of regulated genes are also affected by SMARCA4 knockdown.

5.3.2 Distinct modes of SMAD2 chromatin binding: from pre-acetylated, open target sites to latent enhancers

By characterising histone acetylation and nucleosome occupancy in different NODAL/Activin signalling conditions I showed in this chapter that biologically relevant SMAD2 binding events occur at both active and inactive chromatin. At the latter sites, SMAD2 clearly induces direct H3 acetylation and nucleosome eviction. As already pointed out in this thesis, the issue of whether the deposition of epigenetic marks is cause or consequence of TF binding is still a matter of debate (Spitz and Furlong, 2012). Here I provide evidence that at least for NODAL/Activin signalling it is possible

to establish a causal link between TF binding and the chromatin modifications occurring at its target sites.

This finding was rather unexpected, especially because SMAD complexes have previously been thought to passively bind remodelled active chromatin upon their recruitment by master TFs (Mullen et al., 2011). Indeed, I observed examples of such a scenario in the P19 cells. SMAD2 targeting of pre-acetylated, open sites in the proximity of genes already expressed, as in the case of 'baseline on' *Pou5f1* and *Trh, de facto* resembles the modulatory SMAD2/3 binding events described in ESCs by Mullen et al., 2011. Nevertheless, I also demonstrated that SMAD2 can directly bind to closed chromatin and induce the transcription of 'baseline off' genes like *Lefty1* and *Pmepa1*. Thus, two distinct modes of SMAD2 binding exist. Considering the importance of NODAL/Activin signalling for embryonic development, this finding also solves the contradiction between the modulatory role proposed for SMADs pathway by Mullen et al., 2011 and the biological complexity of the downstream responses observed *in vivo* (Schier, 2009). It is important to note here that active SMAD2 binding to closed sites has not been revealed in this previous work because the authors investigated the chromatin state in chronic signalling conditions, but not in the absence of NODAL/Activin (Mullen et al., 2011). A more recent study did analyse changes in histone modifications in the SB-431542 condition versus active signalling (Bertero et al., 2015). These authors proved that SMAD2/3 controls the deposition of H3K4me3 at promoters and target genes by recruiting the DPY30-COMPASS histone modifiers. However, by focusing only on H3K4me3 changes, this study could not provide any insight about SMAD2/3 activity at enhancers. At these sites H3K4me3 is generally poorly enriched (Calo and Wysocka, 2013, Rada-Iglesias et al., 2011), and other markers such as H3K27ac should have been included in the analysis to be able to detect SMAD2/3 binding to active or inactive enhancers.

After the concept of distinct classes of enhancers was introduced, signatures of active chromatin have been used to predict sites of functional TF binding events in distinct cell types or developmental stages (Rada Iglesias et al., 2012). Importantly, our results suggest that employing such approach in a context where cells are exposed to acute extracellular stimuli could be misleading, unless the epigenetic state is characterised over different signalling conditions. Indeed, I showed that P19 chromatin landscape rapidly changes in response to NODAL/Activin signalling and that high H3 acetylation in SB-431542 is only partially predictive of strong SMAD2

occupancy. If I had not considered how H3K9ac and H3K27ac levels changed over time, relevant SMAD2 binding events to silent chromatin would have not been identified. *Lefty1* and *Pmepa1* SBSs are for instance representative examples of functional SMAD2 target loci which are not active in P19 cells in the absence of NODAL/Activin signalling. However, other histone modifications beside acetylation could mark such sites.

A good candidate for such a modification is H3K4me1, since it has been associated with both active and poised enhancers (Zentner et al., 2011). Indeed, I was able to demonstrate its enrichment at the *Lefty1* SBS in the SB-431542 condition, and over the Activin time course. Considering that no TF is bound at this inactive site in the absence of signalling (see Section 5.3.3), it could still be asked how the H3K4me1 mark is established in the first place. Interestingly, work performed in differentiated macrophages suggests that H3K4me1 can be deposited at silent, un-marked chromatin loci by signalling activated TFs (Ostuni et al., 2013). These so called latent enhancers acquire H3K27ac and remain active as long as the extracellular stimulus is present. When the signalling is removed, acetylation is then rapidly lost at these sites and only H3K4me1 is maintained (Ostuni et al., 2013). It is possible to speculate that in P19 cells SBSs like *Lefty1* or *Pmepa1* also behave as such latent enhancers. According to this hypothesis, in the untreated condition (autocrine NODAL signalling) they would be marked by H3K4me1 and H3K27ac. Blocking NODAL/Activin signalling with SB-431542 would then result in loss of H3K27ac, but H3K4me1 would be maintained. Following SB-431542 washout and Activin stimulation, these pre-marked H3K4me1 sites would rapidly re-acquire H3K27ac. Nevertheless, before confirming the existence of latent enhancers in P19, a ChIP-seq experiment for H3K4me1 over an Activin time course should be performed. Then, it should be addressed how H3K4me1 is established in the untreated condition in the first place, and if this happens in a signal dependent manner.

5.3.3 The role of FOXH1 in mediating SMAD2 binding to chromatin

Once I established the ability of SMAD2 to bind to closed chromatin, in the second part of this chapter I addressed the mechanisms behind it. In particular, I focused on the role of FOXH1, which is a member of the forkhead transcription factor family and

a very well characterised SMAD2 cofactor (Chen et al., 1996). First, I showed that of FOXH1 was necessary for SMAD2 binding at a subset of 'latent enhancers' like *Lefty1*, *Lefty2* and *Pitx2*, and hence it was required for the induction of these target genes. Then, since members of the forkhead transcription factor family have been shown to frequently act as pioneer factors, I tested if FOXH1 worked as such to facilitate SMAD2 binding to its SBSs (Spitz and Furlong, 2012). According to this hypothesis, FOXH1 would stably target closed chromatin and coordinate subsequent SMAD2 recruitment. Surprisingly, I found that FOXH1 binds to these sites only in response to NODAL/Activin signalling together with SMAD2.

This result seems at odds with what has been recently described to happen *in vivo* during the early stages of *Xenopous Tropicalis* development by the Ken Cho's lab (Charney et al., 2017). Here, prior to NODAL signalling activation, FOXH1 is bound to regulatory regions of genes such as *Pitx2*, *Cer1, Nodal2* and *Mix1*, and it recruits the transcriptional co-repressor TLE/GROUCHO. Upon the mid-blastula transition, SMAD2 binds at these sites alongside with FOXA1, displaces TLE/GROUCHO and induces gene expression. Moreover, when late embryos are treated with SB-431542, SMAD2 binding at the mesendoderm enhancers is lost, but FOXH1 binding is retained. To overcome the contradiction between what I described and the work from the Ken Cho's lab, the differences existing amongst the two model systems should be considered. First, in *Xenopous* embryos active gene repression is crucial to prevent premature induction of developmental genes. Such a mechanism does not seem to be employed by P19 cells, where genes are maintained inactive without the need for repressors to be bound at regulating enhancers (eg *Lefty1*). Secondly, the GROUCHO interacting domain is poorly conserved in mammal FOXH1 (data not shown), suggesting that FOXH1 itself could function differently in the two species. Nevertheless, further studies need to be performed to verify this hypothesis.

Importantly, the classical pioneer factor model proposed by Charney and colleagues has been recently challenged also by other *in vitro* studies, which culminated in proposing a new mechanism termed 'dynamic assisted loading' (see Section 1.1.3). This model envisages a highly dynamic system where TFs binding is achieved cooperatively through very rapid factors/chromatin interactions (Swinstead et al., 2016b). Crucially, in such a scenario, opening of closed sites is not due to the TFs 'pioneering' activities, but is the result of ATP-dependent remodelling complexes

and TFs dynamically cycling on and off chromatin. Considering my FOXH1 ChIP-PCR results in conjunction with the *Smarca4* siRNA experiments I will discuss below, I speculate that SMAD2 and FOXH1 might bind chromatin through a dynamic assisted loading mechanism in P19 cells. It would be interesting to verify if other modes of recruiting SMAD2 exists, in particular at those sites like the *Pmepa1* SBS where SMAD2 binding to close chromatin relies on an as yet unidentified TF distinct from FOXH1.

5.3.4 Mechanisms of NODAL/Activin-induced chromatin remodelling

Having shown that FOXH1-SMAD2 complexes bind to closed chromatin and induce nucleosome eviction and H3 acetylation, it was important to address the molecular mechanisms behind these chromatin remodelling events. The requirement of the SWI/SNF ATPase SMARCA4 for the induction of a subset of TGF-β genes was already known (Ross et al., 2006, Xi et al., 2008). However, it was still unclear which stage in the cascade of events from SMAD binding to transcription activation was dependent on SMARCA4. In this chapter I clearly demonstrated that SMARCA4 is necessary to mediate SMAD2 recruitment to a subset of 'latent enhancers' associated with 'baseline off' genes such as *Lefty1* and *Lefty2*. As a result, nucleosome displacement and histone acetylation at these sites are impaired when SMARCA4 is knocked down, and the gene expression in response to acute Activin stimulation is also lost. Importantly, the scenario described perfectly matches with the dynamic assisted loading mechanism just introduced (Swinstead et al., 2016b). Here, TFs and chromatin remodelling complexes constantly collaborate to displace nucleosomes at target sites, in a way that both the binding of the TFs and the establishment of open chromatin depend on these interactions. Similarly, in the absence of SMARCA4, FOXH1–SMAD2 complexes are unable to displace nucleosomes at their target sites. Failing in the creation and maintenance of open chromatin at these loci would in turn dramatically decrease the frequency of FOXH1–SMAD2 binding events. Indeed, SMAD2 enrichment at a subset of target sites is severely reduced in the absence of SMARCA4.

Interestingly, the effect of SMARCA4 knockdown was locus-specific. For example, the SBS of the 'baseline on' gene *Pitx2* was much less affected by SMARCA4 loss compared with other SBSs like *Lefty1*, despite displaying the same

FOXH1 dependency and similar chromatin features. It is possible to speculate that SMADs employ different remodellers at different targets sites. For example, interactions of SMAD2/3 with BPTF, a nucleosome remodelling factor subunit which is not member of the SWI/SNF complex have also been reported (Landry et al., 2008). In the future, it would be interesting to test BPTF's requirement for SMAD2 binding and nucleosome displacement at those sites which are less dependent on SMARCA4, like the *Pitx2* SBS.

Considering what has been discussed so far, the reduction in H3K27ac observed at some SBSs upon *Smarca4* silencing is fully explained by the loss of SMAD2 binding at the same sites. Data from our lab demonstrate in fact that the histone acetyltransferase EP300 responsible for catalysing H3K27ac is directly recruited to chromatin by SMAD2 in response to NODAL/Activin signalling (Calo and Wysocka, 2013, Coda et al., 2017, Ross et al., 2006). In this chapter, alongside H3K27ac, I also investigated Activin-induced changes in H3K9ac and concluded that they follow similar dynamics to those observed for H3K27ac. Importantly, H3K9ac is thought to be mediated by HATs distinct from EP300, namely GCN5, PCAF or Tip60 (Karmodiya et al., 2012). Indeed, SMADs have previously been shown to bind PCAF/GCN5 and it is reasonable to think that that these HATs are recruited alongside EP300 to SMAD2 target sites (Ross and Hill, 2008). However, so far I have not been able to detect these proteins on chromatin, possibly due to technical limitations of the antibodies used.

Chapter 6. ATAC-seq identifies changes in chromatin accessibility and transcription factor occupancy in response to NODAL/Activin signalling.

6.1 Introduction

Since its recent introduction, the assay for transposase-accessible chromatin using sequencing (ATAC-seq) has been used in a growing number of studies to identify epigenome signatures and transcription regulatory networks in different biological contexts (Denny et al., 2016, Gray et al., 2017, Rendeiro et al., 2016). Compared to traditional genome-wide profiling technologies, ATAC-seq provides a great advantage as it presents the possibility of characterizing chromatin accessibility, nucleosome positioning and TF binding dynamics at the same time. This method is based on the action of a hyperactive Tn5 transposase, which has been modified *in vitro* to simultaneously cleave and tag genomic double-stranded DNA with adapters suitable for next-generation sequencing (Adey et al., 2010, Buenrostro et al., 2013). Due to steric effects, Tn5 catalyses transposition reactions with higher probability at open chromatin locations compared to sites in a closed conformation. As a consequence, ATAC-seq reads concentrate at regions of accessible chromatin, allowing the identification of the enhancers and TSSs active in a population of cells by standard peak-calling methods (Figure 6.1). Across the nucleosome-free sequences, however, the frequency of Tn5 insertions is not constant, since the presence of DNA-bound TFs also interferes with tagmentation. These stretches of continuous sequence protected from cleavage correspond to classical 'protein footprints' and are identified by quantifying Tn5 insertions at single base resolution (Figure 6.1). Thus, ATAC-seq data complemented with a transcriptomics approach can be used to infer TF occupancy on a genome-wide scale and to reconstruct cell-specific transcription factor networks.

The experiments described in the previous chapters highlighted the unforeseen role of SMAD2 in remodelling chromatin and orchestrating a complex programme of gene expression downstream of NODAL/Activin signalling. Nevertheless, chromatin accessibility has been rigorously investigated only for a subset of target genomic loci, and changes at non SMAD-bound enhancers have not

been explored at all. Also, which cofactors other than FOXH1 are required for SMAD2 binding to closed sites has not been addressed yet. Finally, the key molecular players responsible for shaping the transcriptional responses in prolonged signalling conditions still need to be identified. ATAC-seq seemed an ideal technique to address these questions, since it allows the simultaneously capture of changes in chromatin states and also footprint profiles.

Here, I present the preliminary results obtained by performing ATAC-seq on P19 cells over the same time-course of NODAL/Activin treatment previously introduced. In the first part of the chapter, I provide proof of evidence for the robustness and quality of the datasets obtained. Then, I attempt to quantify chromatin accessibility at SMAD2 binding sites for each NODAL/Activin signalling condition relative to the SB-431542 state. Finally, the footprint analyses carried out at the SMAD2 binding sites will be discussed, alongside other experimental approaches that will be employed in the future to identify novel SMAD2 transcriptional cofactors.

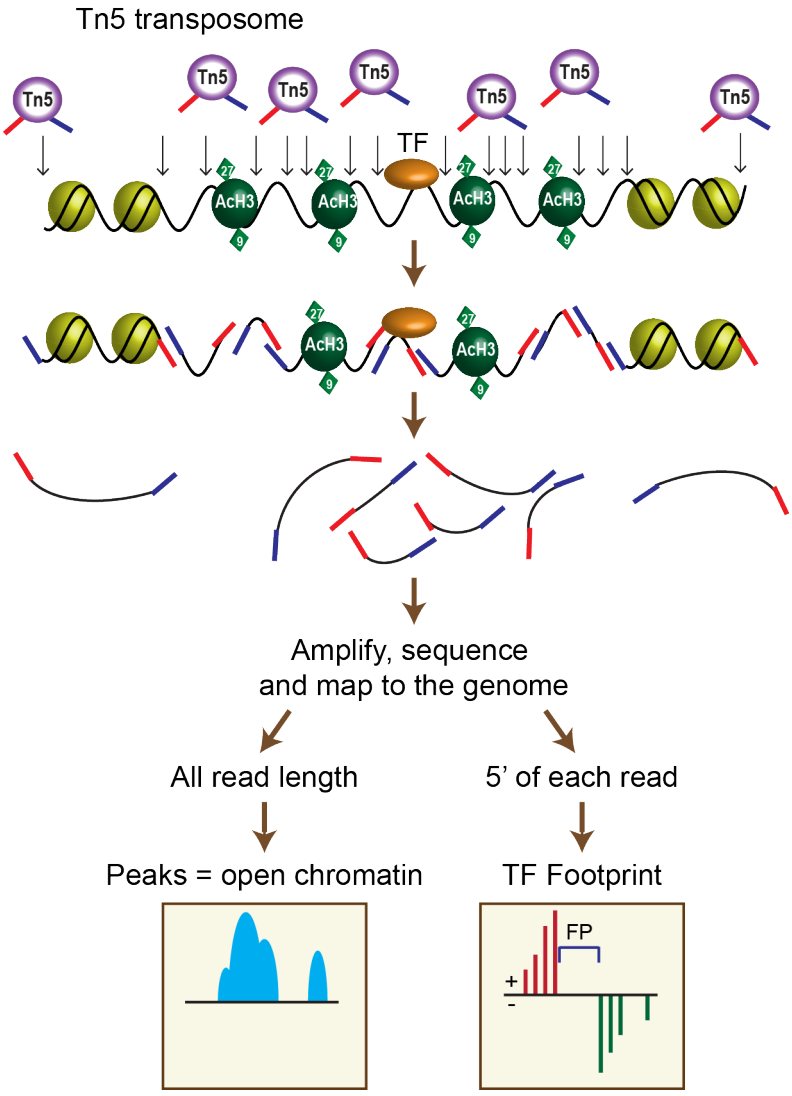


Figure 6.1. ATAC-seq assays open chromatin and provides TF footprints. Schematic of ATAC-seq library preparation and workflow of downstream data-analysis. Genome tagmentation is mediated by Tn5 transposase (purple) loaded with sequencing adapters (blue and red). DNA sequences bound by nucleosomes (green) or by transcription factors (TF, orange) are protected from Tn5 activity. The reads obtained are then mapped to the reference genome over their entire length to identify regions of chromatin accessibility using peak-calling tools. To identify TF footprints, only the 5' end of each read is aligned to the positive (red) or negative (green) DNA strand. Green diamonds, H3K27Ac and H3K9Ac.

6.2 Results

6.2.1 ATAC-seq experiments performed in P19 cells are very reproducible and generate high quality sequencing libraries

From the experiments described so far it was clear that the complex programme of gene regulation downstream of NODAL/Activin signalling relies on chromatin remodelling events and on a poorly characterised network of TFs. To better investigate the relationship between these two aspects it was necessary to capture changes in chromatin accessibility and TF occupancy simultaneously. ATAC-seq provided me with this opportunity. I therefore carried out ATAC-seq (biological replicates) in the same conditions used for the SMAD2 ChIP-seq. After sequencing, the reads were processed for alignment and visualisation by Harshil Patel in the Francis Crick Institute BABS facility. The downstream analyses were also performed in close collaboration.

As an initial step, we verified the technical reproducibility and the quality of the libraries generated. The first was assayed by comparing across replicates the read counts under the ATAC-seq peaks identified in each sample. As shown in Figure 6.2, for each signalling condition the two replicates very well correlated with each other. The quality of the data was verified by plotting the size distribution of the sequenced fragments for each sample. Importantly, all the profiles were similar to each other, both at the level of the individual experiment and across replicates (Figure 6.2). Also, they all exhibited the pattern expected from ATAC-seq data, with a clear periodicity of approximately 200 bp reflecting DNA protection by integer multiples of nucleosomes (Buenrostro et al., 2013). At the same time, all samples were enriched for short inserts, with fragments no longer than 150 bp accounting for approximately half of the reads in each case (Figure 6.3, dotten line). Such a library complexity gave us confidence in the potential of the data to identify both regions of closed chromatin, which relies on sampling long inserts, and TF binding events, which instead requires precise location to an almost single base resolution of Tn5 cutting sites. To further increase the coverage, considering the high reproducibility of biological replicates between experiments we decided to pool the replicates together. Unless otherwise stated, the findings presented in the next sections therefore resulted from the analyses performed on the merged dataset.

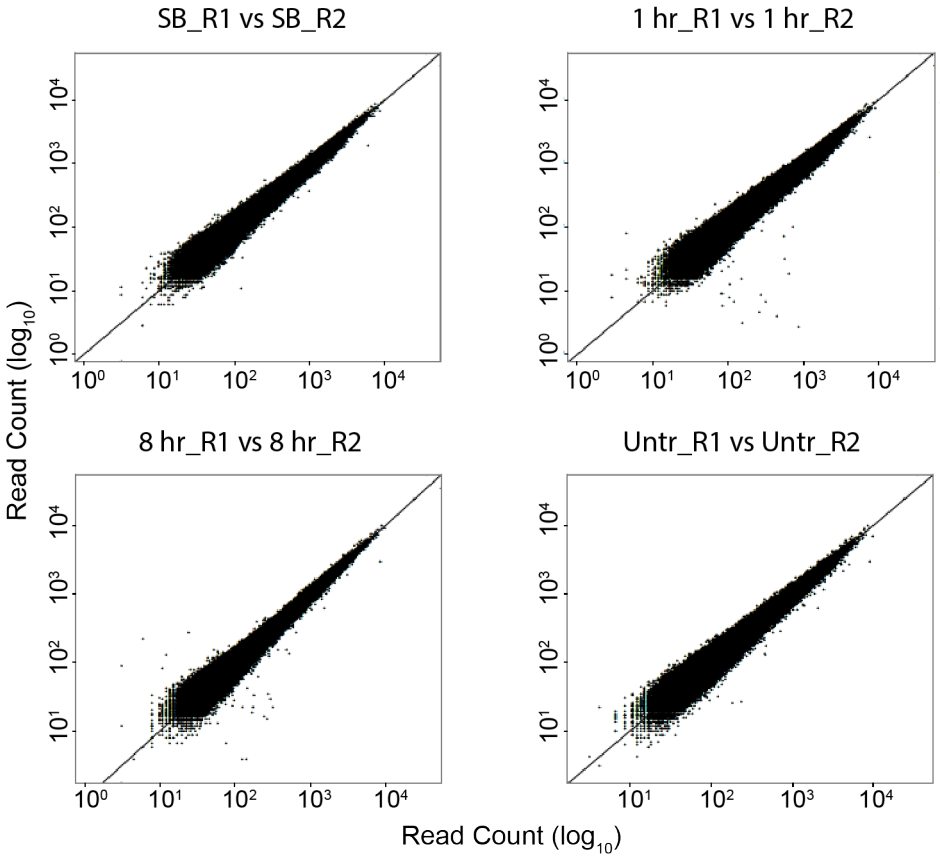


Figure 6.2. The ATAC-seq data from the two biological replicates are highly reproducible.
Scatter plots showing the correlation between replicates for the four time points. Displayed are the normalised read counts for the ATAC-seq peaks identified by MACS.2 in each sample. 1 hr, 1hr Activin; 8 hr, 8 hr Activin; SB, SB-431542; Untr, Untreated; R1, Replicate 1; R2, Replicate 2.

Figure 6.3 on next page.
Figure 6.3. All the sequencing libraries show the typical periodicity in DNA protection expected from a successful ATAC-seq experiment.
(A and **B)** The insert size distribution of the ATAC-seq fragments is displayed as a continuous red line for each sample of either the first **(A)** or the second **(B)** biological replicate. The frequency of insert size is also reported on all plots as a dotted red line. 1 hr, 1hr Activin; 8 hr, 8 hr Activin; SB, SB-431542; Untr, Untreated; R1, Replicate 1; R2, Replicate 2.

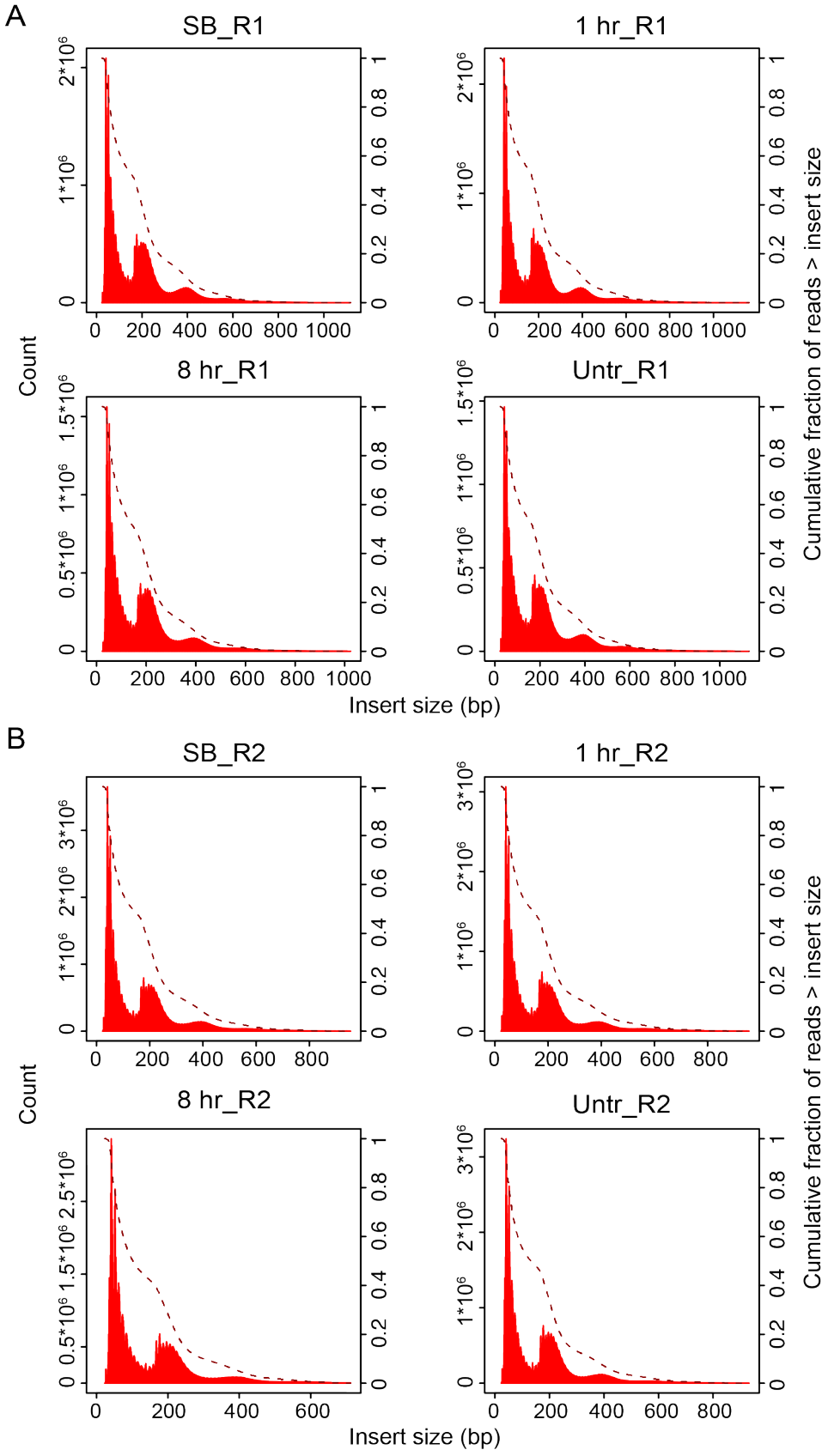


Figure 6.3. See previous page for legend.

6.2.2 NODAL/Activin signalling induces changes in ATAC signal over representative target genomic regions

Since we had discovered that SMAD2 binding occurs with different characteristics and timings for different target genes, I initially inspected ATAC-seq coverage over time at a couple of representative loci by visualising the data with IGV genome browser. As exemplified by the *Lefty2/Lefty1* genomic region, the ATAC signal was enriched at SMAD2 binding sites and also at the promoters/TSSs of the expressed genes. Importantly, the intensity of the peaks at *Lefty1* and *Lefty2* SBSs substantially increased with NODAL/Activin compared to the SB-431542 condition, but it was unchanged at other sites not bound by SMAD2, like at the *Pycr2* promoter (Figure 6.4A). This observation was consistent with the FAIRE results described in the previous chapter, thus confirming that these loci were in a closed conformation in the absence of signalling and 'opened-up' following SMAD2 activation. However, the ATAC signal at *Lefty1* and *Lefty2* SBSs was not zero in the SB-431542 condition, suggesting a constitutive higher chromatin accessibility at these sites compared to the flanking genomic regions.

The ability of ATAC-seq to identify regions of active chromatin such as enhancers and transcriptionally engaged promoters/TSSs became even more apparent when I visualised the ATAC tracks alongside ChIP-seq for H3K27ac and Pol II over the same genomic locus. After 1 hr of Activin treatment, the ATAC signal was in fact enriched for no more than 300 bp in between peaks of histone acetylation at SBSs, in line with the idea that upon SMAD2 binding only one/two central nucleosomes are displaced, whereas the flanking ones are acetylated (Figure 6.4B). Simultaneously, the chromatin at the *Lefty1* TSS was highly accessible, and two ATAC peaks were detected in sharply co-localisation with the SMAD2 peaks and adjacent to the peak for Pol II Ser5P. ATAC signal enrichment was lower over the gene body, as expected from the fact that active TSSs in general are thought to be more nucleosomes devoid than the rest of the gene and as it has been previously observed by others (Wu et al., 2016). Furthermore, it was interesting to note that despite *Lefty1* was not transcribed in absence of signalling, the more distal of the two ATAC peaks just described was already present, suggesting that *Lefty1* proximal promoter/TSS was already partially accessible to Tn5 in the SB-431542 state (Figure

6.4B). Indeed, this was the case for all the other 'baseline off' genes I looked at, possibly reflecting general characteristics of chromatin organisation in P19.

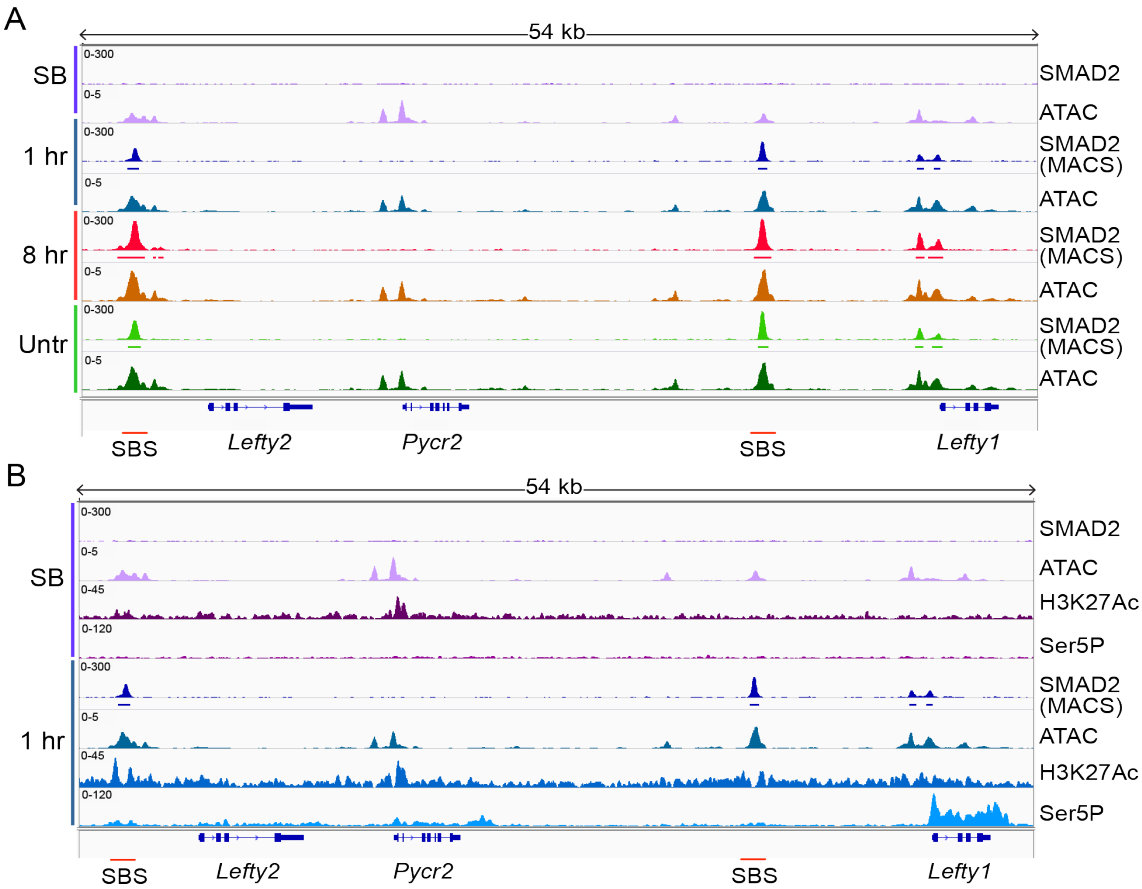


Figure 6.4. The ATAC signal co-localise with SBSs and TSSs at the *Lefty2/Lefty1* genomic locus and its intensity increases with NODAL/Activin signalling.
(A) ATAC-seq was performed on P19 cells treated as indicated. Shown is a screenshot from the IGV genome browser displaying ATAC-seq coverage over the *Lefty2/Lefty1* genomic region alongside tracks from the SMAD2 ChIP-seq and relative peak intervals (MACS). **(B)** As in (A), except that only the SB-431542 and the 1 hr Activin conditions are shown. At these two time-points, the normalised reads from the H3K27ac and Pol II Ser5P ChIP-seq experiments are also visualised. The SMAD2 binding sites are indicated below each panel. SB, SB-431542; 1hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

A similar scenario to the one described for *Lefty1*/*Lefty2* SBSs was also observed at the major *Pmepa1* SBS. Here, the ATAC signal was minimal in absence of signalling and dramatically increased upon SMAD2 binding (Figure 6.5A). In contrast, at the *Trh* SBSs the chromatin was already open in absence of signalling and the ATAC peak intensity remained constant over time at these sites (Figure 6.5B). Altogether these two examples fully confirmed the existence of distinct modes

of SMAD2 binding hypothesised on the basis of the ChIP-seq results, as it is evident by comparing Figure 6.5 with Figure 5.1 and Figure 5.2. Changes in chromatin accessibility in relationship to SMAD2 binding was then analysed at genes that were induced with different temporal dynamics than *Lefty1* and *Pmepa1.*

In the case of the 'transiently induced' gene *Smad7,* SMAD2 binding occurred at an already accessible site, which became more open at later time points alongside increased SMAD2occupancy. Interestingly, the enhancer-like site 5' to the *Smad7* SBS exhibited opposite dynamics, suggesting that it could be functionally relevant for mediating *Smad7* secondary transcriptional repression (Figure 6.6A). The ATAC signal also reflected well the SMAD2 occupancy in the case of repressed genes, as exemplified at the *Tbx3* locus (Figure 6.6B). Finally, I observed that distinct modes of SMAD2 binding could be identified not just in response to acute stimulation, but also at later time points. In the case of *Eomes*, which was a gene induced with delayed kinetics, SMAD2 binding occurred at both open (SBS 1, 4, 5) and closed (SBS 2, 3) chromatin regions 8 hr after Activin treatment, as observed by comparing the changes in ATAC signal at the proximal and distal SBSs over time (Figure 6.6C).

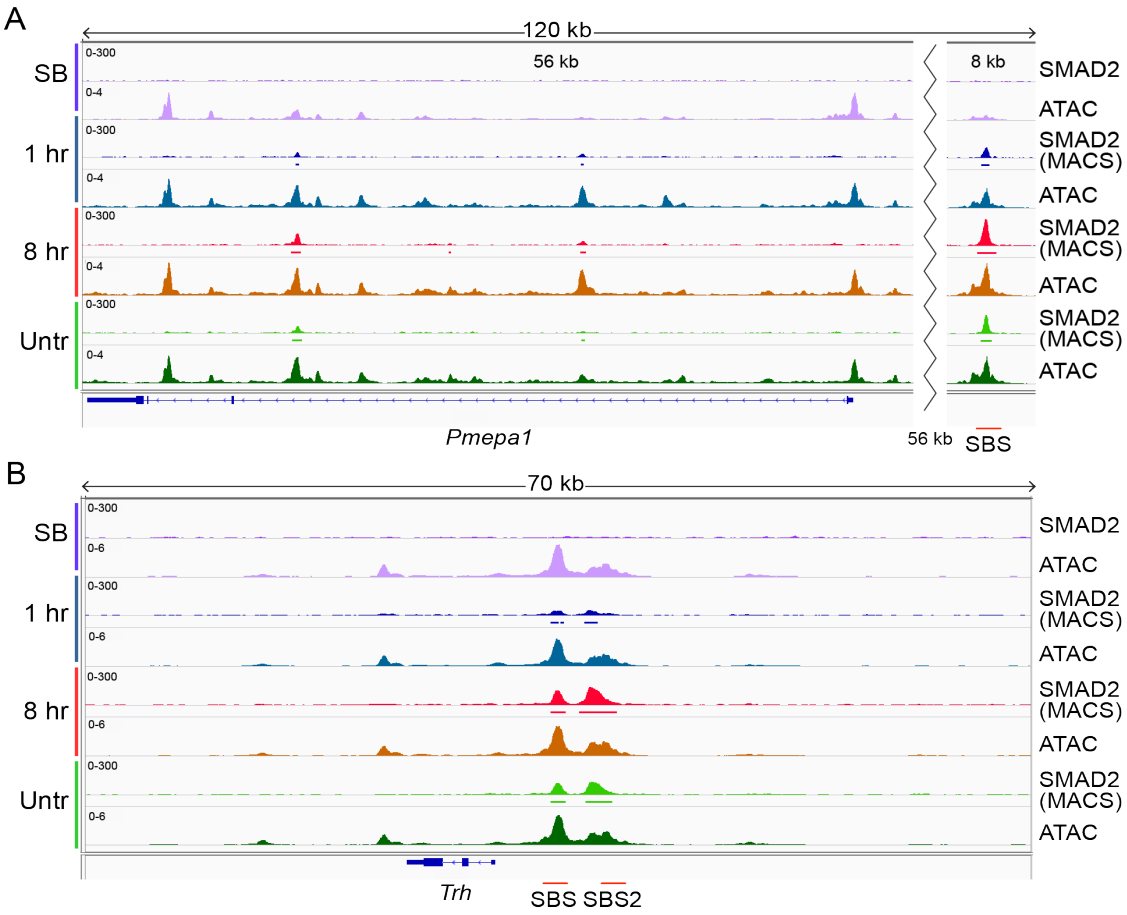


Figure 6.5 on previous page.
Figure 6.5. ATAC-seq confirms SMAD2 distinct modes of chromatin binding. IGV browser visualisation of SMAD2 ChIP-seq and ATAC-seq experiments performed in P19 cells treated as indicated. For the SMAD2 ChIP-seq the MACS-called peaks are also shown. The genomic loci displayed refer to the 'baseline off' gene *Pmepa1* **(A)** and to the 'baseline on' gene *Trh* **(B)**. The SMAD2 binding sites are indicated below each panel. SB, SB-431542; 1hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

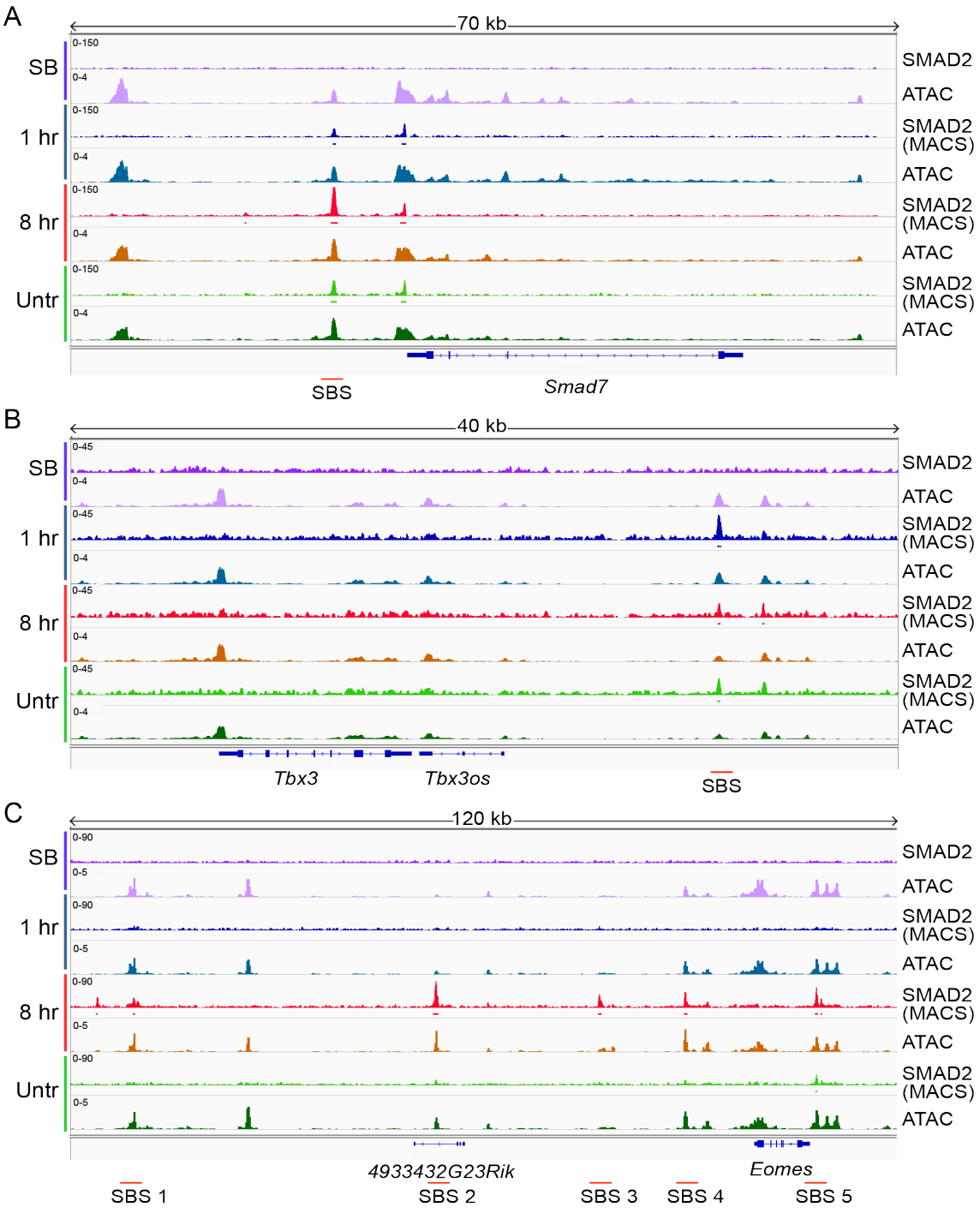


Figure 6.6. See next page for legend.

Figure 6.6 on previous page.
Figure 6.6. ATAC-seq changes over time do not necessarily reflect the different dynamics of gene expression: an overview of representative loci.
IGV browser visualisation of SMAD2 ChIP-seq and ATAC-seq experiments performed in P19 cells treated as indicated. For the SMAD2 ChIP-seq the MACS-called peaks are also shown. The genomic loci displayed refer to the 'transiently induced' gene *Smad7* **(A),** to the 'repressed' gene *Tbx3* **(B)** and to the gene *Eomes*, which is induced with delayed kinetics **(C)**. The SMAD2 binding sites are indicated below each panel. SB, SB-431542; 1hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated.

6.2.3 MACS.2 identifies regions of enriched chromatin accessibility at the SMAD2 binding sites and genome-wide

Following these promising observations, we decided to rigorously identify regions of enriched ATAC signal at the genome-wide level for each of the 4 conditions using the MACS.2 software. As shown in Figure 6.7, the total number of intervals identified and their annotation according to the different genomic features were almost identical across samples, providing further evidence of the quality and technical reproducibility of the data.

Figure 6.7. MACS.2 identifies peaks in the ATAC signal genome wide.
MACS.2 software was used to call ATAC peaks and assign them to either intergenic, exon, intron or promoter/TSS DNA regions. For each sample is reported the total number of peaks and their distribution according to the 4 genomic features used for the annotation.

Since from the visual inspection of representative regions discussed above it appeared that ATAC peaks co-localise with SMAD2 peaks, we asked if this was the case for the all high confidence set of SMAD2 binding sites. For each time point, the percentage of SMAD2 target loci encompassed by an ATAC peak was calculated. As expected, with both acute and prolonged NODAL/Activin signalling almost all (> 90%) the consensus SBSs overlapped for at least 1 bp with an ATAC interval, and

surprisingly this proportion was maintained in the SB-431542 state (Figure 6.8). For example, the very low ATAC signals observed at *Lefty1* and *Pmepa1* SBSs in absence of signalling were called by MACS.2 as enriched compared to the background level of coverage in that sample. This observation thus suggested that despite many SBSs such as *Lefty1* and *Pmepa1* were occupied by nucleosomes in absence of signalling as I had previously shown, additional features could mark these sites making them relatively more accessible to Tn5 compared to the surrounding chromatin.

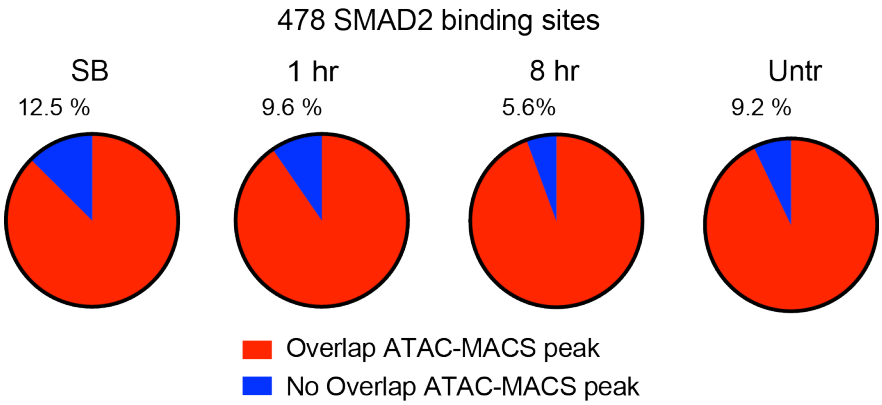


Figure 6.8. In all the signalling conditions, the large majority of SMAD2 binding sites intersect with a chromatin accessible region.
For each of the 478 SMAD2 consensus intervals, the distance to the closest ATAC-seq peak was computed in all signalling conditions. For each sample, the fraction of SMAD2 binding sites overlapping or not with an ATAC-seq interval for at least 1 bp is displayed in red or blue.

Intrigued by this unexpected result, I then focused on those SBSs which did not intersect with an ATAC interval in the SB-431542 state and questioned if these sites acquired an ATAC peak upon NODAL/Activin stimulation. Indeed, it was the case for 15 out of 60 SBSs, while 22 of them were never found to overlap an ATAC peak in any signalling condition (Figure 6.9). The average SMAD2 enrichment at these 22 sites was significantly lower compared with the rest of the dataset and most of the binding events occurred only at one time point, suggesting that they were likely not functionally relevant (data not shown). Moreover, none of the SBSs I used as models to describe my results were included in this group (For the full list, see Appendix 8.3).

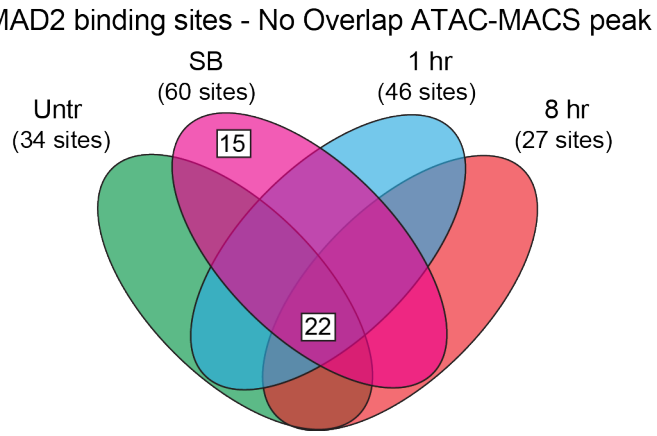


Figure 6.9. Out of the 478 SMAD2 binding sites, 22 SBSs are never found in a chromatin accessible region.
For each signalling condition, the list of SBSs which did not overlap with an ATAC peak was retrieved from Figure 6.7. The individual groups of SBSs were then compared using Venny 2.0 and displayed is the graphical output of the analysis, with highlighted the number of sites in common between all four samples (22), or exclusively present in the SB-431542 one (15). The total number of elements in each list is also reported.

At the same time, since for each genomic locus the number and the length of the ATAC intervals identified was not necessarily completely constant across different samples, we decided to merge them together in order to create a dataset of consensus ATAC peaks. As expected from the results above, the consensus ATAC intervals intersected 456 out of 478 high confidence SMAD2 consensus peaks (Figure 6.10A). Importantly, the large majority of these 456 intervals (listed in Appendix 8.4) was entirely encompassed by a consensus ATAC peak, and no one had less then 40% of its length in common (Figure 6.10B). Overall, this observation confirmed that in general SMAD2 binding sites highly co-localise with regions of accessible chromatin in at least one of the 4 signalling conditions.

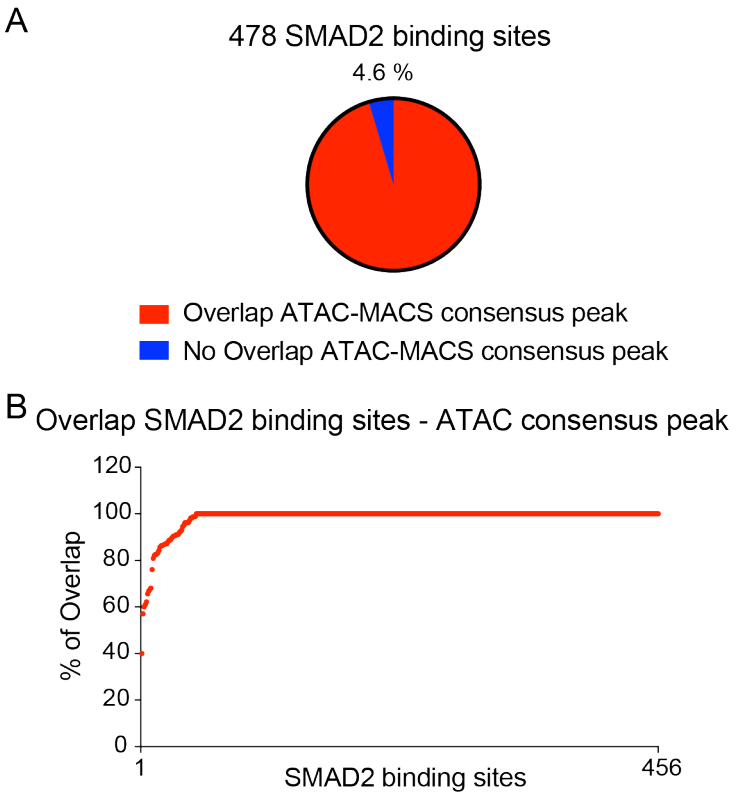


Figure 6.10. The large majority of SBSs perfectly co-localises with a region of open chromatin in at least one signalling condition.
The ATAC intervals identified in each sample by MACS.2 were merged to form a consensus list of ATAC peaks. **(A)** For each of the 478 SMAD2 consensus intervals, the distance to the closest ATAC consensus peak was computed and the fraction of SMAD2 binding sites overlapping or not with an ATAC-seq interval for at least 1 bp is shown in red or blue. **(B)** For each of the 456 SBSs positive for intersection, the portion of the SMAD2 interval overlapping the corresponding ATAC peak is plotted. On the graph, the SBSs on the x axis are displayed in ascending order according to their percentage of overlap.

6.2.4 SMAD2 binding results in increased chromatin accessibility at both 'baseline off' and 'baseline on' sites

Once I had defined the dataset of consensus ATAC peaks genome-wide, we then set out to determine the sites at which ATAC intensity significantly changed with NODAL/Activin signalling compared to SB-431542. To achieve statistical relevance, the reads from each sample were separated accordingly to the original replicates and used as an input for the DiffReps package. As the output, DiffReps provided with the genomic coordinates of the intervals where the ATAC signal is different between the condition of interest the SB-431542 sample, along with its computed adjusted p-

value (padj) and log_2FC. To increase the stringency of the data obtained we also decided to consider only the DiffReps intervals which had a padj lower then 0.1 and a log_2FC greater than 0.5 as an absolute value. Overall, this analysis aimed to provide answers to two separate questions. First, we wanted to rigorously quantify changes in chromatin accessibility occurring over time at the high confidence set of SMAD2 binding sites. Secondly, we sought to identify genomic regions near NODAL/Activin regulated genes which were not targeted by SMAD2, but which were differentially enriched for ATAC signal in response to the signalling. In the first case, the goal was to pull out all the sites at which SMAD2 binding directly induced chromatin remodelling, in the second case it was to find enhancer-like loci that could be functionally involved in regulating NODAL/Activin transcriptional responses. Here, I will present only the data regarding SMAD2 binding sites, since the rest of the analysis is still on-going.

For the high confidence set of SMAD2 peaks, we identified DiffReps intervals of differential ATAC-seq at 236 sites out of 478, meaning that in at least one signalling condition the chromatin accessibility at these loci changed compared to the SB-431542 state (Figure 6.11, for the full list see Appendix 8.5). These DiffReps positive SBSs were then grouped accordingly to the kinetic category of the genes they were associated with, and the temporal changes in ATAC-seq were displayed next to the time at which the SMAD2 peak was detected at each site (Figure 6.12A).

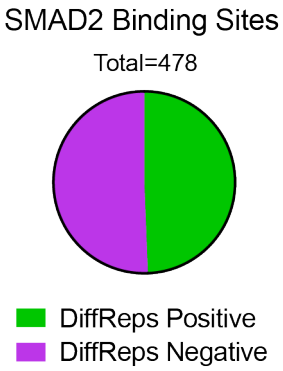


Figure 6.11. NODAL/Activin signalling induced changes in ATAC-seq at 236 out of 478 SBSs. DiffReps analysis was performed to identify intervals of differential ATAC signal relative to SB-431542. After further filtering of the data (see main text), the portion of SBSs which overlapped with a DiffReps interval in at least one signalling condition is displayed in green, whilst the fraction of DiffReps negative SBSs is shown in purple.

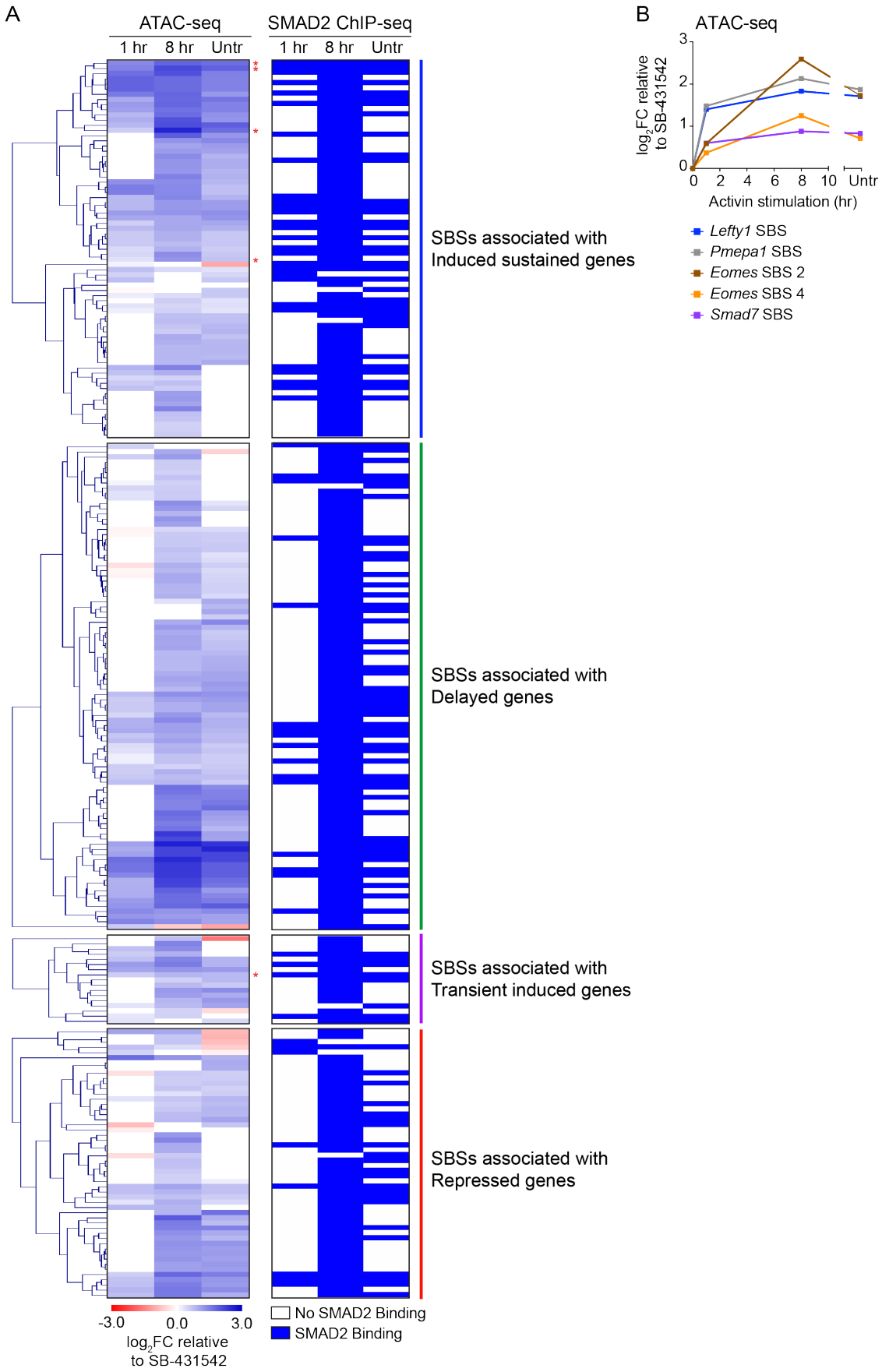
Since I had previously noticed that at some representative genomic regions ATAC-seq levels varied across signalling conditions, I first checked if these differences also resulted from the DiffReps analysis. Indeed, the data displayed in the heatmap confirmed what I first observed by visual inspection and discussed in the section above. For example, ATAC signal at *Lefty1* and *Pmepa1* SBSs

significantly grew upon pathway activation compared to the SB-431542 state, in line with the fact that I had already described SMAD2-dependent chromatin remodelling events occurring at these sites (Figure 6.12B). The ATAC-seq changes over time quantified by DiffReps were in agreement with the IGV screenshots also for the other representative loci previously examined, as validated by the values reported on the heatmap for *Eomes* and *Smad7* SBSs (Figure 6.12B). Once I had acquired confidence in the reliability of the results obtained using the DiffReps approach, I then considered the data displayed in the heatmap from a more global perspective.

Importantly, I noticed that the chromatin accessibility of SBSs tended to generally increase upon NODAL/Activin signalling, regardless of the dynamic of induction/repression of the associated target gene (Figure 6.12A). Moreover, the changes in ATAC-seq correlated well with the time at which SMAD2 binding events occurred, confirming that SMAD complexes directly elicited the chromatin remodelling of its target sites. In fact, for the majority of SBSs no differences in the ATAC-seq levels were detected at 1 hr Actin compared to the SB-431542 condition, but the chromatin accessibility of these loci increased with prolonged Activin signalling at the same time of SMAD2 binding. Nevertheless, when a SMAD2 peak was detected upon acute stimulation, ATAC signal over the corresponding SBSs also increased. Interestingly, in a few cases the changes in ATAC-seq anticipated SMAD2 binding, possibly suggesting NODAL/Activin mediated recruitment of SMAD co-factors at these loci (Figure 6.12A).

Figure 6.12 on next page.

Figure 6.12. SMAD2 binding increases chromatin accessibility compared to SB-431542 at 236 SBSs.
(A) The SMAD2 binding sites positive for a DiffReps interval were divided accordingly to the four kinetics categories of their associated genes. For each group of peaks, the heatmap on the left displays the log_2FC relative to SB-431542 for ATAC-seq at the different time points. The values plotted were computed by DiffReps and reflect the changes measured within the intervals of differential signal identified in each signalling condition. The order of peaks in each group was obtained by hierarchically clustering the data and is presented on the left most part of the figure. For all SBSs, the presence (blue) or absence (white) of a MACS-called SMAD2 peak at each time point is shown in the heatmap on the right. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated. **(B)** The values from (A) are displayed limited to a handful of representative SBSs previously introduced (see Figures 6.4; 6.5; 6.6). These sites are listed in the plot legend accordingly to the order they appear in the heatmap in (A), from top to bottom. The red asterisks on the heatmap in (A) denote their relative positions. Untr, Untreated.


A
ATAC-seq
1 hr 8 hr Untr
SMAD2 ChIP-seq
1 hr 8 hr Untr
SBSs associated with
Induced sustained genes
SBSs associated with
Delayed genes
SBSs associated with
Transient induced genes
SBSs associated with
Repressed genes
-3.0 0.0 3.0
log2FC relative
to SB-431542
No SMAD2 Binding
SMAD2 Binding
B
ATAC-seq
log2FC relative
to SB-431542
0 2 4 6 8 10 Untr
Activin stimulation (hr)
Lefty1 SBS
Pmepa1 SBS
Eomes SBS 2
Eomes SBS 4
Smad7 SBS

In the previous chapter, I showed that SMAD pathway activation induced nucleosome displacement at a subset of sites associated to 'baseline off' genes. I therefore expected to observe a robust increase in ATAC signal at these sites in response to NODAL/Activin signalling. To test the hypothesis, I compared the DiffReps positive SBSs with the DiffReps negative SBSs and noticed that the number of 'baseline off' SBSs in the first group was significantly higher than in the second. Thus, signalling-induced changes in chromatin accessibility were more likely to occur at SBSs associated to 'baseline off' genes rather than at 'baseline on' SBSs (Figure 6.13). This result was not surprising, since I had already reported that in general the latter sites were already nucleosomes depleted in the SB-431542 condition, whilst the 'baseline off' SBSs were in a more closed chromatin state in absence of signalling (see Chapter 5, Figure 5.8). Nevertheless, it was interesting to note that the 'baseline on' SBSs accounted for more than half of the SMAD2 peaks positive for DiffReps intervals, suggesting that many sites already accessible in the SB-431542 condition became even more open upon NODAL/Activin signalling.

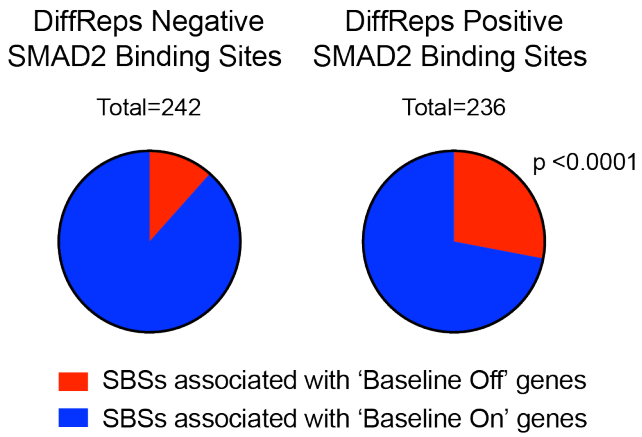


Figure 6.13. The 236 DiffReps positive SBSs are enriched for sites associated with 'baseline off'genes.
The 478 consensus SMAD2 peaks were divided into two groups based on the presence or absence of a DiffReps interval in at least one signalling condition as describe in Figure 6.11. For each group, the percentage of SBSs associated with 'baseline off' (red) or 'baseline on' (blue) genes is indicated. Un-paired Chi-square test was performed on the data using a 95% confidence interval and the resulting p-value is reported on the graph.

In conclusion, DiffReps analysis of ATAC signal over the high confidence set of SMAD2 peaks showed that SMAD2 binding directly increased chromatin accessibility at a large number of target sites. These chromatin remodelling events

did not exclusively occur at loci in a closed conformation, but also occurred at many SBSs already open in absence of signalling. Overall, the findings just described not only confirmed the existence of different modes of SMAD2 chromatin binding, but they also indicated that SMAD complexes generally induce nucleosome displacement at their target sites.

6.2.5 Footprint prediction analysis reveals TF occupancy at representative SMAD2 binding sites in response to NODAL/Activin signalling

Since DNA sequences directly bound by proteins are protected from transposase activity, ATAC-seq has been successfully used to infer loci occupied by DNA binding proteins genome-wide (Buenrostro et al., 2013).This approach is based on the identification of 'footprints', defined as sites where the frequency of Tn5 cuts is relatively lower compared to the surrounding open chromatin regions. I therefore reasoned that carrying out a similar analysis on our ATAC-seq dataset could help us to unveil the network of TFs responsible for recruiting SMAD complexes to chromatin and for shaping the transcriptional responses downstream. Amongst the different computational methods for footprint prediction we decided to use the Wellington tool, which does not require an *a priori* set of motifs and thus provided us with the advantage of identifying also *de novo* TF-DNA interactions (Piper et al., 2013). Indeed, this algorithm had been successfully used for ATAC-seq data by our collaborator Vicky Metzis in James Briscoe lab, at the Francis Crick Institute (Vicky Metzis, unpublished). As before, the results I will present here were generated with the help of Harshil Patel in the Francis Crick Institute BABS facility.

The IGV genome browser screenshot for the *Lefty1* SBS in Figure 6.14 well represents the output of the footprint analysis performed on our ATAC-seq data. First, the Tn5 'cuts' were computed at a single base resolution by aligning the 5' end of each sequencing tag to the forward or reverse DNA strand. Regions of significantly lower cleavage density compared to the flanking sequences were then predicted as footprints. In the case of the *Lefty1* SBS, four deep notches (named footprint (F.) 3 – 6) in the Tn5 signal were detected 1hr after Activin treatment. Notably, two of them well co-localised with FOXH1 motifs, suggesting that FOXH1 was likely bound at both sites. In particular, the second footprint (F. 4) was highly conserved and was occurring right under the SMAD2 peak (Figure 6.14). This confirmed the relevance

of taking into account, when considering possible SMADs recruiting factors, their motif position relative to the SBS summit, as previously hypothesised (see Chapter 5, Figure 5.14). Since I had previously demonstrated that FOXH1 binds to the *Lefty1* SBS in acute and prolonged signalling conditions, but not in the SB-431542 state, I then asked if any footprints were predicted in the samples other than 1 hr Activin.

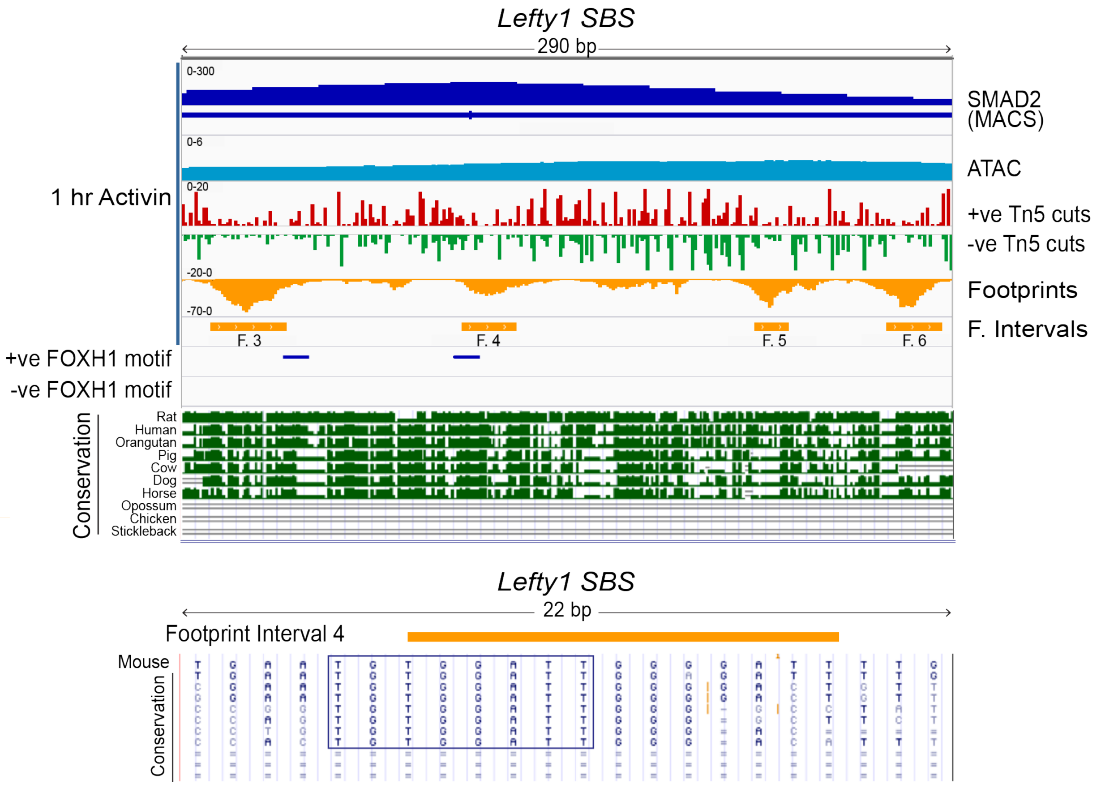


Figure 6.14. Footprints identified at the *Lefty1* SBS co-localise with FOXH1 binding motif sequences.
Top panel, screenshot from the IGV genome browser of the *Lefty1* SBS locus for the 1 hr Activin sample. Displayed are the SMAD2 ChIP-seq and relative peak interval (MACS), alongside with different tracks from the ATAC-seq experiment. In light blue, the coverage obtained considering the ATAC-seq reads for their entire length. In red and green, per nucleotide Tn5 cutting activity. In orange, the Wellington footprint prediction and relative intervals (F. 3–6). DNA sequences matching the FOXH1 motif are denoted in dark blue. The conservation tracks downloaded from UCSC genome browser for the species indicated are also displayed. Bottom panel, zoom in of the region surrounding the footprint indicated. The blue box indicates the mouse sequence matching the FOXH1 motif and its alignment with the species shown in the panel above.

Importantly, when the NODAL/Activin pathway was blocked using SB-431542, no intervals of decreased cutting density were found across the *Lefty1* SBS locus, suggesting that this site was devoid of any bound TFs in absence of signalling

(Figure 6.15). In contrast, the 1 hr Activin footprints described in Figure 6.14 were detected in all the other conditions. However, we noticed that their genomic coordinates varied slightly across samples (eg. F. 3, 4, 5). Since this usually only amount to a few base pairs, I concluded that it was likely to be the result of technical artefacts. We therefore merged the footprints identified in each individual sample to create a consensus list of intervals to use in downstream analyses for identifying temporal changes in footprint occurrences. The loss or the appearance of intervals with prolonged signalling could reflect binding events biologically relevant to shape SMADs transcriptional responses over time. Indeed, in the 8 hr Activin and untreated samples two new footprints (F. 1, 2) were detected at the 5' end of the SMAD2 peak in addition to the four already present at 1 hr Activin (Figure 6.15).

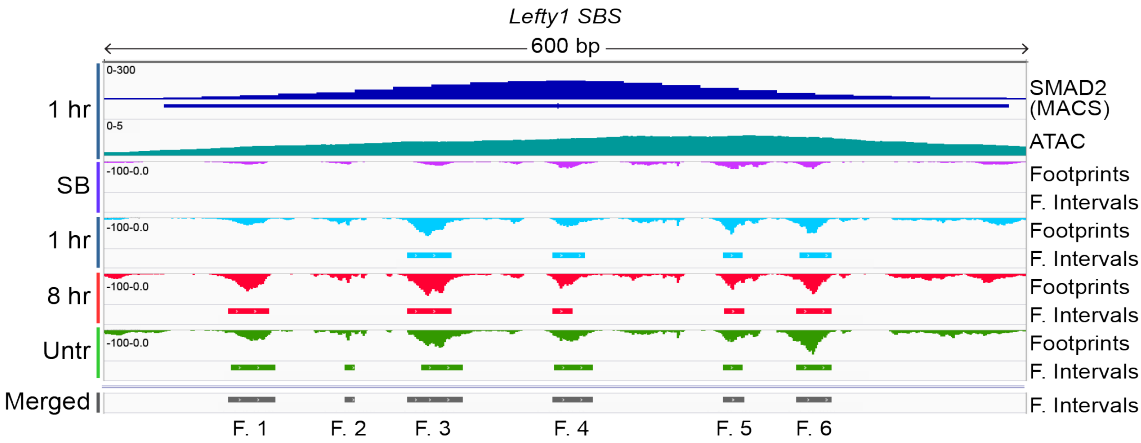


Figure 6.15. At *Lefty1* SBS, footprint signals change in response to NODAL/Activin signalling.
Screenshot from the IGV genome browser of the *Lefty1* SBS locus. Shown are the SMAD2 ChIP-seq and relative peak interval (MACS) together with the ATAC-seq all read coverage for the 1 hr Activin sample. Underneath, the Wellington footprint prediction are displayed for each time point. In grey, the consensus intervals (F. 1–6) generated by merging the individual footprint from each sample.

Overall, the results just described were not unexpected, since I had already characterised FOXH1 binding to the *Lefty1* SBS by ChIP-PCR. However, the screenshots just described worked as crucial positive controls, since they indicated that the footprints predicted using our ATAC-seq data were likely to reflect 'real' TFs binding events. I therefore concluded that considering the intervals identified by Wellington alongside the phylogenetic conservation tracks represented a good strategy to pinpoint the TFs recruiting SMADs to chromatin, other than FOXH1.

Such an approach could for instance be very informative in the case of the *Pmepa1* SBS. As previously shown, SMAD2 binds to this closed site in a FOXH1-independent manner, suggesting the presence of other yet unidentified SMAD2 co-factors at this genomic locus. Indeed, upon acute Activin treatment three footprints were detected at conserved sites in closed proximity of the SMAD2 peak summit (Figure 6.16). Screening experiments of TF candidates able to bind the sequences 'footprinted' and to recruit the SMAD complexes to the *Pmepa1* SBS are currently on-going.

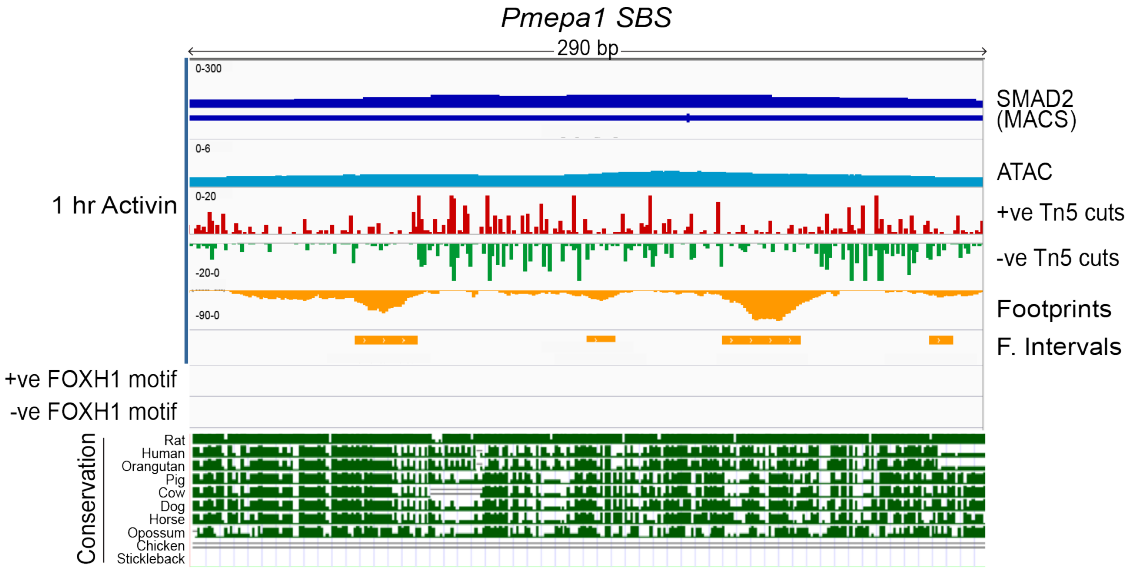


Figure 6.16. Sequence-based analysis of local footprints to pinpoint novel SMAD2 co-factors: the example of the *Pmepa1* SBS.
Screenshot from the IGV genome browser of the *Pmepa1* SBS locus for the 1 hr Activin sample. Displayed are the SMAD2 ChIP-seq and relative peak interval (MACS), alongside with different tracks from the ATAC-seq experiment. Respectively, the ATAC-seq all read coverage, the Tn5 cutting activity at a single base resolution and the Wellington footprint prediction with relative intervals are shown according to the colour coding introduced in Figure 6.13. Note that no sequences matching the FOXH1 motif are present under the SMAD2 peak at this site. At the bottom, the conservation tracks downloaded from UCSC genome browser for the species indicated are also shown.

6.2.6 The four ATAC-seq datasets successfully reveal transcription factor occupancy genome-wide: the CTCF footprint as a paradigm

Having showed that the Wellington algorithm was able to predict specific footprints in response to NODAL/Activin signalling at selected SMAD2 target loci, its efficiency remained to be tested at a genome-wide level. Moreover, it was still possible that the differences across samples described above were due to technical reasons, such as the failure in calling intervals of reduced Tn5 activity for the SB-431542 condition.

We therefore set out to address these questions by investigating the per-nucleotide cut density around all the CTCF recognition sequences in each of the four samples. We chose to focus on CTCF because this TF has been shown to generate footprints at its occupied target sites readily detectable with ATAC-seq data (Buenrostro et al., 2013). Furthermore, I reasoned that CTCF was unlikely to play a role in determining NODAL/Activin transcriptional responses, since it is primarily regarded as a major regulator of 3D chromatin architecture and no evidence suggesting CTCF-SMAD interactions have been reported to date (Phillips and Corces, 2009). To minimise noise in the result, we also decided to consider the cleavage coverage only across those sites with DNA sequences matching the CTCF motif and overlapping with a consensus Wellington interval for at least 1 bp. The same criteria was also employed to investigate the occupancy of other TFs, as discussed further below.

First, Tn5 activity across each potential CTCF binding site was displayed as a 200 bp window TF-centric heatmap for the 1hr Activin sample. In general, the signal sharply dropped either side of the motif, implying those DNA sequences were protected from transposition by chromatin-bound CTCF (Figure 6.17). Indeed, when averaged cleavage profiles were generated, we observed a well-stereotyped footprint comparable to the ones previously obtained from ATAC-seq experiments in other cell lines (Buenrostro et al., 2013). Since the heatmap retained the strand specificity information associated with the per base number of Tn5 insertions, I was also able to appreciate a certain degree of imbalance in the data. In fact, the cuts assigned to the positive strand tended to occur more frequently upstream of the binding site than downstream, and vice versa for the ones aligning to the negative strand (Figure 6.17). This finding suggested that, at least in the case of CTCF, the ‘cut tags’ represented opposite ends of the same DNA fragment spanning the protected region. The result was also consistent with previous analyses performed on DNase-seq datasets (Piper et al., 2013).

We then went on to produce heatmaps and averaged profiles for the other time points as well. To facilitate the comparisons, the CTCF binding sites were ordered in all plots accordingly to the 1 hr Activin sample. Importantly, no readily apparent differences were observed in any case with respect to 1 hr Activin, suggesting that CTCF occupancy genome-wide was not altered by different signalling conditions (Figure 6.17). To further confirm this finding, per base changes in Tn5 activity relative to SB-431542 were computed and displayed as heatmaps. If

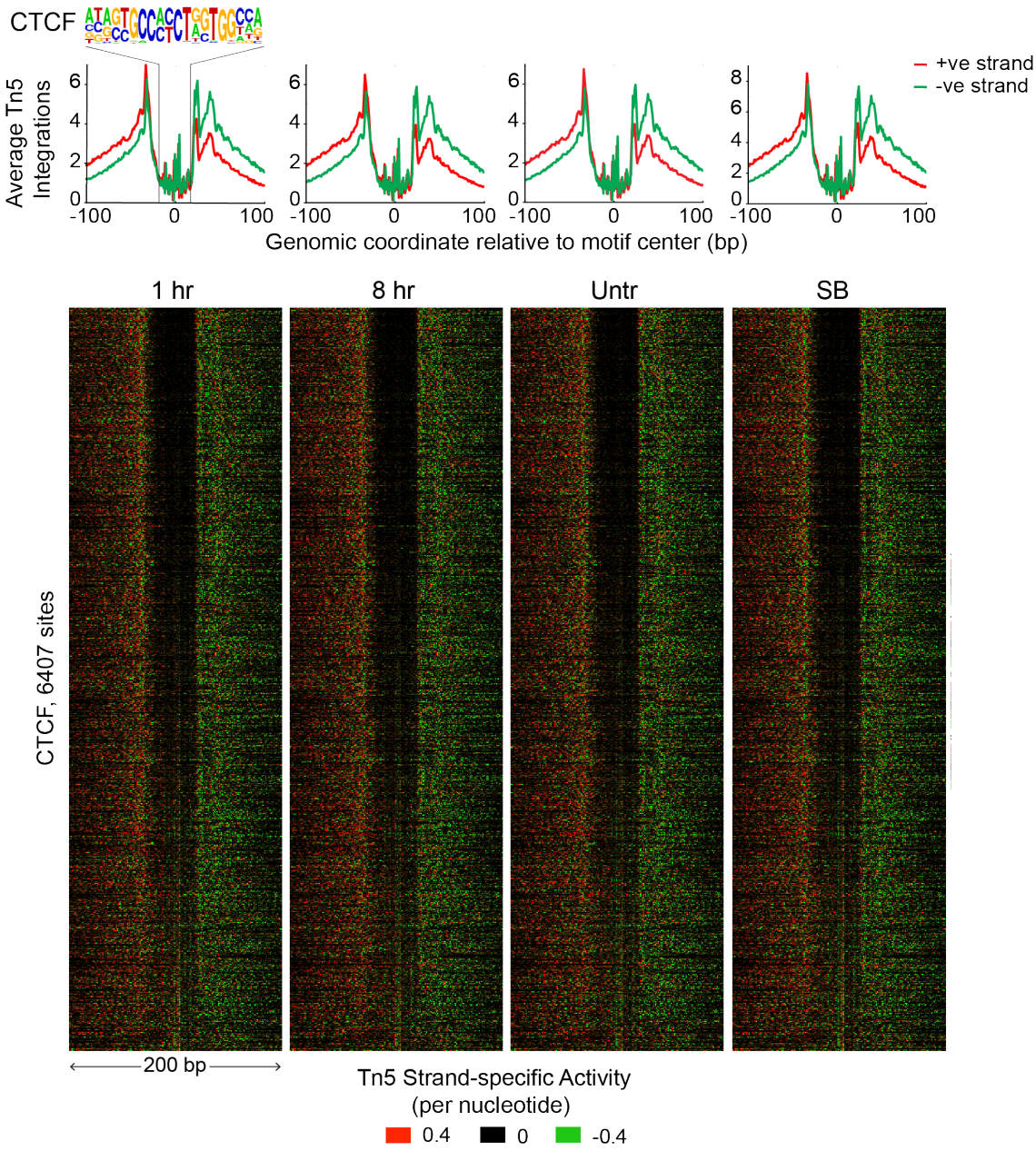


Figure 6.17. ATAC-seq reveals genome-wide CTCF occupancy in all signalling conditions.
For the 4 samples, the heatmaps show the per nucleotide Tn5 cutting activity for each site with a CTCF binding motif intersecting a consensus footprint interval. The data refer to a 200 bp window centred on the CTCF motif. An excess of positive strand Tn5 cuts over the negative strand is shown in grades of red, the opposite is indicated in grades of green. In the 1 hr Activin heatmap, the order of the sites from top to bottom reflects their decreasing Footprint Occupancy Score as calculated by the Wellington algorithm. For all the other samples, sites are instead ordered accordingly to the 1 hr Activin plot, thus each row refers to the same locus moving across the signalling conditions. On top of each heatmap are also displayed the aggregate Tn5 cleavage profiles as obtained from averaging Tn5 integrations in the corresponding samples, with the CTCF motif logo from the Homer database shown on top of the first plot. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated. SB, SB-431542.

NODAL/Activin had directly influenced CTCF binding at a subset of genomic loci, signals reflecting differential footprints would be detected at these sites. However, this was not observed, and for all sites the DNA sequences within the CTCF motif were protected from Tn5 activity with the same efficiency, regardless of SMADs pathway activation. It has to be said that changes were instead detected outside the motif boundaries; however, in the absence of a clear pattern I concluded that they were likely reflecting the stochasticity of Tn5 cuts rather than biologically relevant events (Figure 6.18).

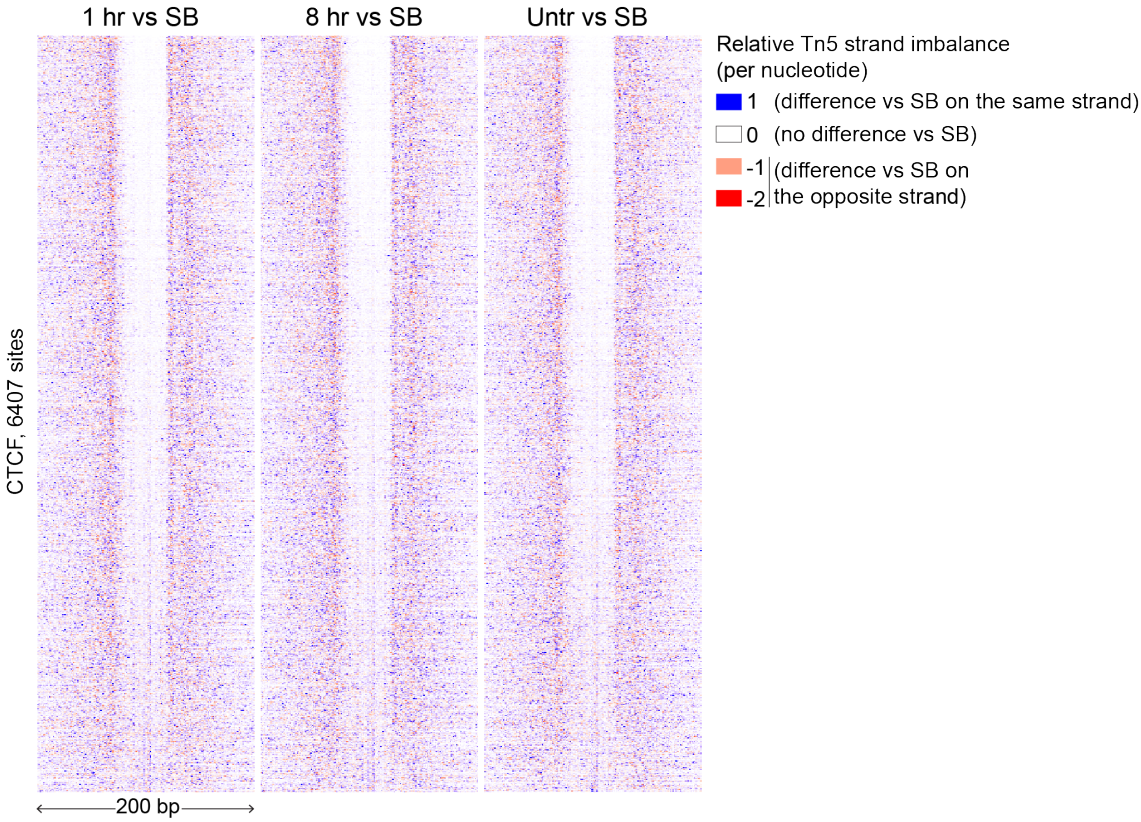


Figure 6.18. Genome-wide CTCF occupancy does not change in response to NODAL/Activin signalling.
For the three on-going signalling samples, the heatmaps display the per nucleotide changes in Tn5 activity relative to the SB-431542 condition over the same sites introduced in Figure 6.17. The order of the sites is also consistent with the plots in Figure 6.17. Grades of blue reflect differences in the number of Tn5 cuts in the presence of NODAL/Activin compared to the SB-431542 sample on the same DNA strand. Grades of red indicate differential Tn5 activity occurring on opposite DNA strands. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated. SB, SB-431542.

Overall, the results here presented demonstrated that CTCF footprints can be successfully detected in all four signalling conditions, thus excluding the presence of technical biases in individual samples. Also, they gave us confidence for using these ATAC-seq datasets to extract changes in the genome-wide occupancy of other TFs over the Activin time course.

6.2.7 Characterisation of footprint changes for 300 known motifs at the SMAD2 binding sites: a novel approach to unveil the NODAL/Activin transcriptional network

Based on the encouraging results described so far, we sought to identify the network of TFs responsible for recruiting SMAD2 to its target sites and shaping the downstream transcriptional responses. To this end, we decided to start by identifying, for each of the 318 known motifs included in the HOMER database, the matching sites overlapping with a consensus ATAC peak. For any given motif, we first asked which fraction of these loci also intersected with a consensus footprint interval, thus defining a motif-specific 'footprint frequency' over ATAC-seq consensus sites. As expected, the results varied depending on the TF considered, suggesting that the efficiency in footprint detection differed between TFs. Not surprisingly, the footprint occurrences did not match the total numbers of motif sites, since the presence of a specific DNA sequence does not necessarily imply the binding of the correspondent TF at that particular locus (For details, see Appendix 8.6).

We then repeated the analysis starting from the motifs which had at least one matching site overlapping with a SMAD2 high confidence consensus peak. For any of the 294 motifs identified, the corresponding motif-specific 'footprint frequency' over SMAD2 consensus sites was obtained, and plotted against its 'footprint frequency' over ATAC consensus sites (Figure 6.19A). In fact, I reasoned that comparing the individual footprint frequency across the two sets of loci would help me to pinpoint the TFs interacting with SMAD complexes. Intuitively, for these factors, the likelihood of being 'footprinted' at a SBS would be expected to be higher than at any other non-functional site harbouring their recognition sequences and overlapping with an ATAC-seq consensus site. Surprisingly, this was the case for the large majority of the 294 motifs which had at least one site overlapping a consensus SMAD2 interval (Figure 6.19A). I therefore decided, by using arbitrary defined cut-offs, to retain only

those motifs which had a low footprint frequency over the ATAC sites ($<\log_{10}(-1.25)$), but an high frequency over the SMAD2 sites ($>\log_{10}(-1)$). As a result, I obtained a 'high SBSs footprint frequency' group of motifs, accounting for 252 of them, amongst which were the ones previously found to be enriched under the SMAD2 peaks by the MEME-ChIP software as described in Figure 5.12 (For the full list, see Appendix 8.7). As expected, the CTCF motif was not included in the high stringency list, thus working as a negative control for the dataset identified (Figure 6.19A). When considering the result, however, it is important to note that the analysis was not weighted for the total number of sites observed for each motif. Indeed, I observed that motifs with just a handful of occurrences, such as HIF1a (4), were 'footprinted' with the same frequency than those having hundreds of loci in common with a SMAD2 interval, as in the case of NANOG (134). Nevertheless, it was reasonable to think that the first were less likely to be relevant for NODAL/Activin transcriptional responses than the second ones.

In order to correct for the analysis biases just discussed, I ranked the motifs in the 'high SMAD2 footprint frequency' group according to the absolute number of Wellington intervals associated with each of them (Figure 6.19B). Importantly, amongst the first 100, there were many motifs recognised by TFs known to interact with SMAD2, such as NANOG, EOMES, POU5F1 and FOXH1 (Beyer et al., 2013, Brown et al., 2011, Chen et al., 1996, Faial et al., 2015, Kunwar et al., 2003, Mullen et al., 2011). Alongside, other motifs whose role in NODAL/Activin signalling has not been fully characterised yet, as in the case of the ZIC and the KLF family of proteins (Figure 6.19C). Overall, the presence of these motifs gave me confidence in using the list as a valid starting point to screen for novel SMAD2 cofactors and/or mechanisms of NODAL/Activin-regulated transcription.

To facilitate the analysis, we decided to visualise how the footprints changed over time at the SMAD2 binding sites. Heatmaps such as the ones described above for CTCF were then generated for the whole set of motifs. Limiting to the ones considered so far, a high degree in 'footprinting efficiency' was observed. A clear footprint was obtained, for instance, in the case of the homeobox PITX1 motif. As shown in Figure 6.20A, for most sites the DNA sequences within this motif were distinctly protected by Tn5 activity compared with the flanking sides. Importantly, the individual footprint signals did not seem to significantly change over time, possibly suggesting that the factors recognising the PITX1 motif were already bound in the

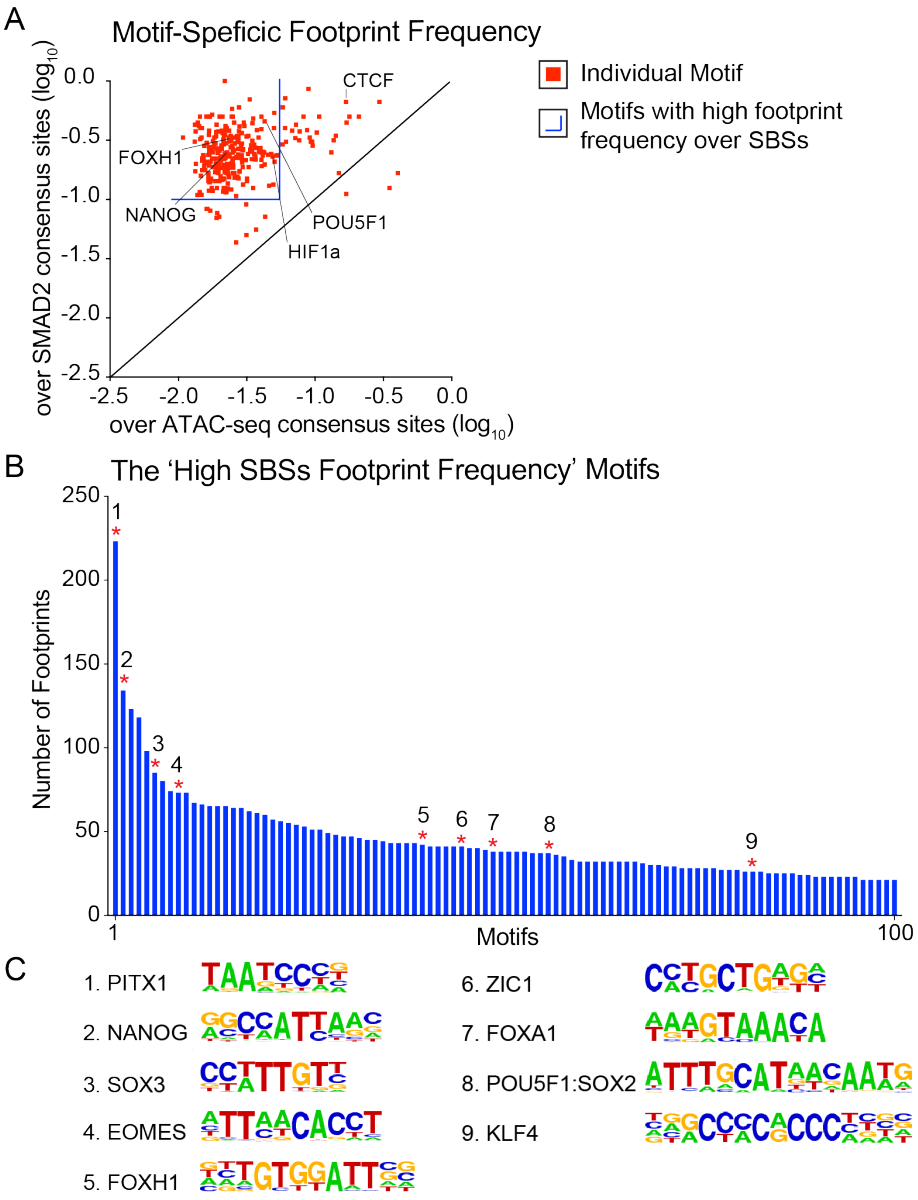


Figure 6.19. Motifs with high 'footprint frequency' at SMAD2 binding sites are likely associated with SMAD2 cofactors.
(A) The 294 known motifs from the Homer database overlapping at least one consensus SMAD2 interval were considered. The individual 'footprint frequency' was defined as the ratio between the number of footprint occurrences over motif occurrences, and calculated referring either to the motif-matching sites intersecting an ATAC-seq consensus interval (x axes), or to the ones under a SMAD2 consensus interval (y axes). The values obtained for each motif were plotted on a Log_2 scale. The blue box delimits the cut-offs chosen to identify the 252 motifs with a 'higher footprint frequency' at the SMAD2 binding sites with respect to the ATAC-seq consensus sites. **(B)** The motifs in the 'high SBSs footprint frequency' group were ranked accordingly to the absolute number of footprints observed under the 478 SMAD2 binding sites and the results for the first 100 motifs are displayed in a decreasing order from left to right. Marked by the red asterisks are the ones for which the motif logo is displayed in **(C)**. Here, the matching TF binding sites are also indicated. The motifs are ordered top to bottom accordingly to what described for (B).

absence of NODAL/Activin signalling (Figure 6.20B). In contrast, in the case of FOXH1, the heatmaps showed a much less defined pattern, due to the fact that at many loci Tn5 was cutting in correspondence of the motif centre (Figure 6.21A). This result was not surprising, since many TFs, amongst which the forkhead family member FOXA1, are known to have a short residence time on the DNA, and thus leave poorly detectable footprints (Goldstein and Hager, 2017, Swinstead et al., 2016a). As a consequence, it became difficult to identify relevant FOXH1 binding events and their dynamics over time (Figure 6.21B). For example, the changes in FOXH1 occupancy at its *Lefty1* SBS target sites were not as evident as previously described, to an extent that they would have gone unnoticed if their biological relevance was not known *a priori* (Figure 6.21).

In summary, I concluded that some TFs leave very clear footprints, whilst others have a low footprint efficiency, presumably because they rapidly cycle on and off chromatin. Therefore, I am currently exploring other complementary methods to identify the TFs responsible for recruiting SMAD2 to specific enhancers and for modulating the transcriptional responses to NODAL/Activin signalling over time.

Figure 6.20 on next page.

Figure 6.20. A clear footprint profile is observed for the PITX1 motif over the SMAD2 binding sites in all signalling conditions.

(A) For the four samples, the heatmaps show the per nucleotide Tn5 cutting activity for each site with a PITX1 binding motif intersecting a consensus footprint interval under a SMAD2 binding site. The data refer to a 200-bp window centred on the PITX1 motif and in each heatmap the sites are ordered accordingly to the Footprint Occupancy Scores obtained for the 1 hr Activin sample, as described in Figure 6.15. The aggregate Tn5 cleavage profile from the 1 hr Activin condition only is also shown, with the PITX1 motif logo from the Homer database displayed on top. Untr, Untreated. SB, SB-431542. **(B)** For the same sites as in (A), the differential heatmaps show changes in the per nucleotide Tn5 activity over time relative to the SB-431542 condition. For further details about the information displayed, see Figure 6.16. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated; SB, SB-431542.

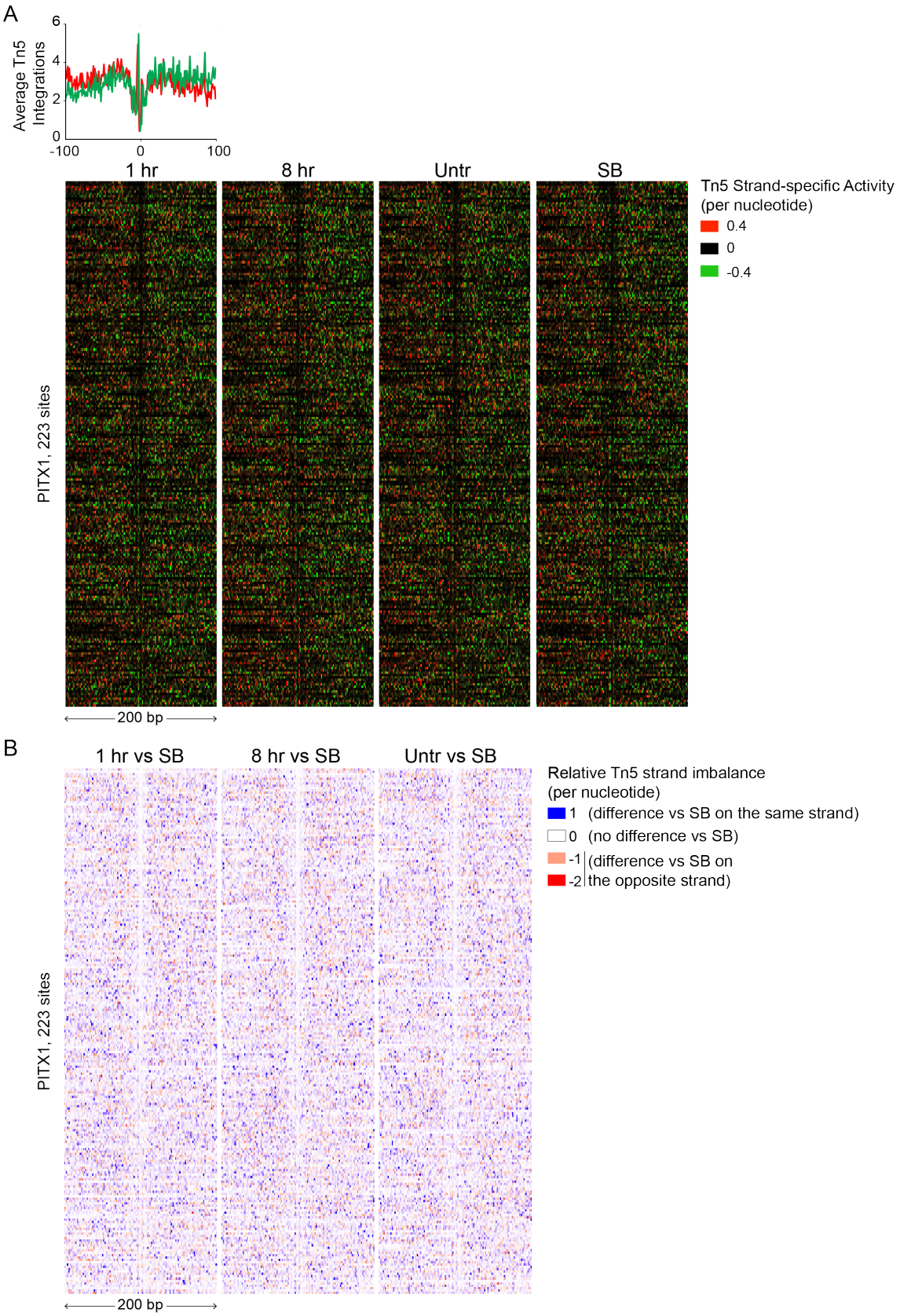


Figure 6.20. See previous page for legend.

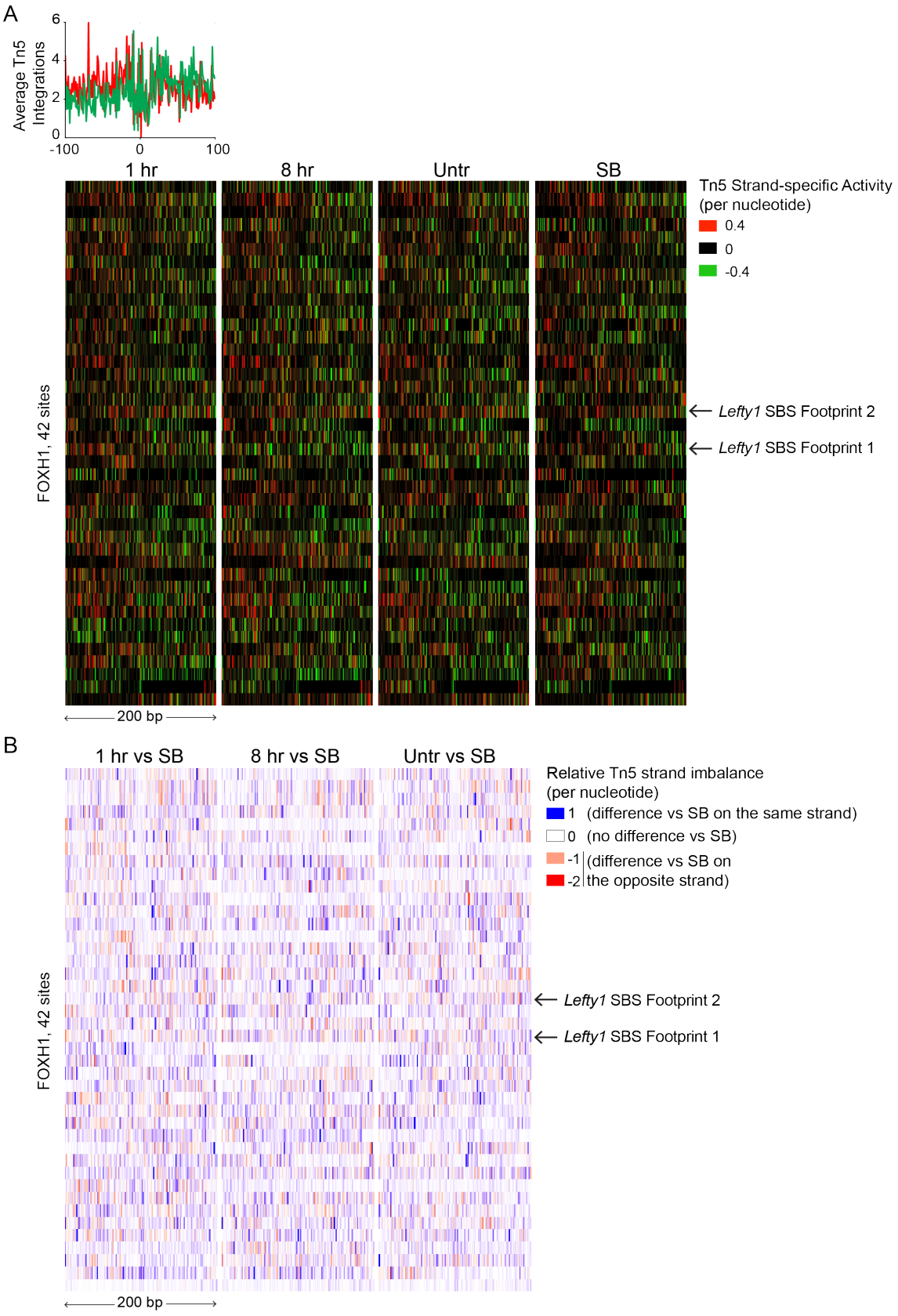


Figure 6.21. See next page for legend.

Figure 6.21 on previous page.

Figure 6.21. Overall FOXH1 occupancy at SMAD2 binding sites is poorly detected by footprint analysis.

(A) The heatmaps display the per nucleotide Tn5 cutting activity for each site with a FOXH1 binding motif intersecting a consensus footprint interval under a SMAD2 binding site. The data refer to a 200-bp window centred on the FOXH1 motif and in each heatmap the sites are ordered accordingly to the Footprint Occupancy Scores obtained for the 1 hr Activin sample, as described in Figure 6.15. The average Tn5 cleavage profile from the 1 hr Activin condition only is also shown, with the FOXH1 motif logo from the Homer database displayed on top. The arrows indicate the FOXH1 sites intersecting the footprint intervals detected in the presence of NODAL/Activin under the Lefty1 SMAD2 binding site. Untr, Untreated. SB, SB-431542. **(B)** For the same sites as in (A), differential heatmaps show changes in the per nucleotide Tn5 activity over time relative to the SB-431542 condition. The arrows refer to the same *Lefty1* SBS footprints introduced in (A). For further details about the information displayed, see Figure 6.15. 1 hr, 1 hr Activin; 8 hr, 8 hr Activin; Untr, Untreated; SB, SB-431542.

6.3 Discussion

6.3.1 Summary of main findings

- The ATAC-seq datasets obtained for the four signalling conditions are highly reproducible. All replicates correlate well with each other and the insert size distribution plots show the expected nucleosome periodicity.
- At representative genomic loci, ATAC signal increases in response to NODAL/Activin at SMAD2 target sites associated with 'baseline off' genes (eg. *Lefty1*, *Pmepa1*), but does not change at 'baseline on' SBSs (eg. *Trh*).
- MACS.2 peak calling analysis identifies regions of chromatin accessibility genome-wide in all ATAC-seq samples. 90% of SBSs intersects an ATAC-seq peak in the SB-431542 state, and this increases to 95% of SBSs in the presence of NODAL/Activin.
- DiffReps analysis confirms the existence of distinct modes of SMAD2 chromatin binding. For 50% of SBSs, chromatin accessibility increases upon SMAD2 binding, and this group is enriched for 'baseline off' SBSs.
- At the major *Lefty1* SBS, the detection of signalling-inducible footprints corresponding to DNA sequences matching the FOXH1 motif confirms the absence of FOXH1 binding in the SB-431542 state.
- The ATAC-seq datasets can be used to infer changes in TF occupancy genome-wide. For CTCF, which functions as a positive control for the technique since it is not involved in NODAL/Activin responses, the characteristic footprint profiles are observed in all samples, and are not affected by NODAL/Activin signalling.
- The footprints detected over SBSs are enriched for motifs recognised by known SMAD2 interactors, but novel SMAD2 co-factors can also be identified with this approach in the future.
- FOXH1 occupancy at SBSs is poorly predictable from footprint heatmaps, but more informative plots can be generated for other motifs such as the PITX1 motif.

6.3.2 ATAC-seq and the changes in chromatin accessibility in response to NODAL/Activin: from the SMAD2 binding sites to future strategies to identify the downstream transcription factor network

In this chapter, I used ATAC-seq data generated in different signalling conditions to explore the changes in chromatin accessibility at SMAD2 target sites over time. I first showed that almost all SBSs are included in regions of detectable ATAC signal already in the SB-431542 state, and this finding at first seems to be at odds with the ChIP-seq results discussed in Chapter 5. There, I concluded that some of these loci were in a closed chromatin conformation in the absence of signalling, being enriched for histone H3 and devoid of any H3K27/H3K9 acetylation. Indeed, by considering only those data, it was not possible to detect such SBSs from the surrounding sequence in the SB-431542 state. However, detectable 'bumps' in ATAC-seq are instead called at these sites in the SB-431542 state, with respect to the background of the flanking genomic regions.

There are several different mechanisms that could explain these hints of ATAC signal detected at the 'latent enhancers' in absence of signalling. It could be speculated that sequences wrapped around nucleosomes marked by histone modifications such as H3K4me1 are in general more accessible to Tn5 compared to the ones spanning equally compacted chromatin, but where these modifications are not present. Alternatively, since the ATAC signal results from sampling chromatin across a population, it is also possible that the 'closed' SBSs are indeed 'opened' in the SB-431542 state in few individual cells, and that these stochastic events are picked by ATAC-seq, but not by less sensitive techniques such as histone H3 ChIP-seq. Nevertheless, it is important to stress that the chromatin accessibility at the 'latent enhancers' is dramatically increased in response to NODAL/Activin signalling by ATAC-seq, in agreement with what was described in Chapter 5. Moreover, the ATAC signal grows upon SMAD2 binding was also at many 'baseline on' sites. Thus, the model previously proposed to describe SMAD2 distinct modes of binding is not only confirmed, but possibly expanded by the data presented in this chapter. It is in fact now evident on a genome-wide level that SMAD2 targets both inactive and active chromatin. At the first sites, SMAD2 induces *de novo* nucleosome displacement, whilst at the second it can further increase their chromatin accessibility.

So far, I have focused on the changes only occurring at the SMAD2 binding sites. However, it could be interesting to explore the full repertoire of enhancers identified in the different signalling conditions, considering in particular those showing differential ATAC-seq levels in response to NODAL/Activin signalling. This point is particularly relevant in the effort of identifying the transcriptional network that contributes to shape SMAD2 responses in prolonged signalling conditions. In fact, it can be possible that these TFs are not only recruited to SMAD2 binding sites, but instead also bind other enhancers associated with SMAD2 target genes. Indeed, I reported a decrease in chromatin accessibility at a SMAD2 independent site near the 'transiently-induced' gene *Smad7*, and this could reflect the binding of a TF responsible for *Smad7* secondary repression. In order to pinpoint similar scenarios, however, it is necessary to complement the ATAC-seq datasets with the information obtained from published experiments investigating the 3D chromatin organisation at a genome-wide level, like Hi-C or Next Generation (NG) Capture-C. Indeed, such strategy has been already employed in other studies (Simon et al., 2017). As a result of this analysis, only those enhancers which physically interact with SMAD2 regulated promoters/TSSs will be taken forward for further screenings. First, their importance for the correct expression of the corresponding SMAD2 target genes will be verified by using a CRISPR/Cas9-mediated approach analogous to that I already employed to test the role of *Lefty1*/*Lefty2* SBSs. Then, in the cases of positive results, the individual ATAC-seq footprints (discussed below) present at those enhancers will be carefully examined to identify the matching TFs. Finally, the latter will be further tested for chromatin binding and functional relevance in the context of the SMAD2 transcriptional responses considered.

6.3.3 ATAC-seq footprinting as a strategy to identify novel SMAD2 cofactors: challenges and future prospects

Other than mapping chromatin accessibility in different signalling conditions, I also propose to use the ATAC-seq data to infer TF occupancy, which in turn depends on the detection of characteristic footprints. I showed that this methodology could be employed to answer both locus-specific and genome-wide questions. In the first case, the goal is to identify the TF binding to a locus of particular interest. I plan, for example, to discover which are the factors required for SMAD2 binding to closed,

FOXH1-independent sites, such as the *Pmepa1* SBS. Such an approach could also be used to find the TF targeting specific enhancers involved in secondary transcriptional responses at SMAD2 target genes, as just discussed for the *Smad7* locus. In contrast, reconstructing the entire regulatory network of TFs downstream of SMAD2 requires a genome-wide perspective. Taking advantage of the HOMER database of known motifs, we aim to locate all the SBSs at which the individual matching DNA sequences are 'footprinted', and to characterise the changes in TF occupancy observed at these sites over time. This analysis in the future could also be extended to all the SMAD2-independent enhancers which are likely to be bound by TFs newly-synthesised in response to the signalling, and thus contribute to establish the programme of gene expression downstream of NODAL/Activin.

It is important to note that both strategies just described crucially rely on two key premises: the detection of footprints and the assignment of footprints to matching TFs. With respect to the first step, multiple aspects, either technical and biological, are currently subject of debate in the transcription field. Due to its recent introduction, ATAC-seq footprint analysis is mostly based on adapting tools originally developed for DNase I data, such as the Wellington algorithm we employed in this work. Indeed, several studies have successfully used other computational methods, such as CENTIPEDE or PIQ, and a consensus 'gold standard' in the data processing workflow is yet to be achieved (Buenrostro et al., 2013, Rendeiro et al., 2016). The different genomic footprinting strategies have been recently reviewed in depth by Vierstra and colleague (Vierstra and Stamatoyannopoulos, 2016). Finally, it is still not known to what extent footprint detection is affected by Tn5 cleavage preferences (Tsompana and Buck, 2014). Indeed, other DNA-cleaving enzymes, such as DNase I, have a non-random sequence specificity, which can generate biases in motif-centric analysis resulting in the so called zigzag signature plots (Goldstein and Hager, 2017).

Besides the existence of technical limitations, it is also becoming clear that many TFs do not leave significant footprints at their binding sites at all. Live cell experiments using fluorescent microscopy techniques have recently revealed that these TFs have a highly dynamic binding behaviour, with residence times at target loci between 5 and 15 seconds (Sung et al., 2016). It is the case, for example, of the glucocorticoid receptor, the TFs OCT2/SOX4 and the pioneer factor FOXA1 (Chen et al., 2014, Sung et al., 2014, Swinstead et al., 2016a); and probably, I speculate,

of FOXH1. As a consequence, only those TFs with long residence time of binding to DNA, such as CTCF, would be able to protect the chromatin from Tn5 activity, generating differences detectable in the ATAC-seq experiments. Indeed, the footprints we described in this chapter are highly factor-dependent, as exemplified by the opposite results obtained with the PITX1 or FOXH1 motifs, respectively.

Provided that the detection of footprints is successfully accomplished, it is then necessary to identify the causative TFs. The challenge of this step is represented by the fact that many TFs recognise highly similar motifs and that, *vice versa*, a single TF can have multiple sequence preferences (Vierstra and Stamatoyannopoulos, 2016). Indeed, examples of such redundancy are reported in the result section. The PITX1 motif, for instance, may be recognised by many other homeobox proteins; and this is the case also for the FOXA1 motif and the KLF4 motif, which are bound by multiple members of the respective families (verified by using the corresponding motif matrixes as input of the TOMTOM suite, data not shown). I also found that the RNA-seq data did not help to narrow-down the list of candidates, since multiple TFs from each family were expressed with similar dynamics.

Considered the limitations of the footprints-motifs based strategy, in order to successfully pinpoint novel SMAD2 cofactors I propose to complement the ATAC-seq datasets with further experiments. In particular, it could be highly informative to perform ChIP for SMAD2 coupled with mass spectrometry, following the recent described SICAP method (Rafiee et al., 2016). SICAP stands for selective isolation of chromatin-associated proteins and it allows the identification of chromatin-bound partners of a bait of interest, which in our case is SMAD2, assayed in the four NODAL/activing signalling conditions. Since DNA proteins are crosslinked by formaldehyde, this technique provides with the advantage of capturing also those TFs which do not leave footprints due to their short binding residence times, such as the FOX proteins discussed earlier. In cases of successful footprinting, SMAD2 ChIP-SICAP will instead facilitate the process of assignment to the matching TF. The screening for candidates recognising a given motif could be simplified by prioritising those TFs interacted with SMAD complexes from the mass spectrometry analyses. For the factors shortlisted, the functional relevance in context of SMAD2 responses will be tested using the CRISPR/Cas9 technology. To that end, I plan to collaborate with the High Throughput Screening Facility at the Francis Crick Institute in order to generate a line of P19 stably expressing inducible Cas9, which will allow us to

perform multiple knock-downs at the same time. This will be essential to test if a particular transcriptional output requires the combinatorial action of distinct TFs, or if the expression of the selected target gene is rescued by TF redundancy mechanisms.

In conclusion, I am confident that combining the ATAC-seq footprint analysis with the experimental approaches just described will enable us to unveil the network of TFs established downstream of NODAL/Activin to modulate the transcriptional responses over time.

Chapter 7. Discussion

7.1 Defining new principles of SMAD-regulated transcription

In my PhD project I have addressed how a signal responsive transcription factor establishes a programme of gene expression which is constantly remodeled over time, and I have unravelled the underlying mechanisms. I used, as a model system, the NODAL/Activin-SMAD2 pathway in mouse embryonic P19 carcinoma cells, which enabled me to study how SMAD2 binds chromatin and regulates transcription in response to acute, sustained and chronic NODAL/Activin signalling from a signal-inhibited baseline. By combining existing RNA-seq and SMAD2 ChIP-seq datasets with ChIP-seq for H3 modifications and total H3, ChIP-seq for different forms of Pol II and ATAC-seq I have delineated the sequence of events that occur from SMAD2 chromatin binding to regulation of gene expression, and determined the underlying molecular mechanisms. Crucially, my findings establish new paradigms for ligand-activated SMAD2-dependent transcription (Figure 7.1).

First, I showed that there are two modes of SMAD2 binding. For a subset of genes, SMAD2 binds to transcriptionally silent non-acetylated chromatin and it induces nucleosome eviction and acetylation of H3 at adjacent nucleosomes with fast kinetics. These sites are likely to be marked by H3K4Me1 and SMAD2 binds them without the requirement for a pioneer factor. In fact, I provided direct evidence that SMAD2 and the co-factor FOXH1 are both detected on chromatin in an inducible manner in response to NODAL/Activin signalling. Moreover, at some target genes I found that SMAD2 binding to inactive chromatin, subsequent nucleosome displacement and histone acetylation all require the remodeling activity of SMARCA4. At other targets, which have a basal level of on-going transcription, SMAD2 binds to pre-acetylated nucleosome-depleted regions where it promotes a further increase in acetylation. In general, SMAD2-induced acetylation is likely mediated by EP300, as this HAT is recruited to chromatin simultaneously with SMAD2. I then demonstrated that SMAD2 regulates Pol II activity via *de novo* recruitment to target promoters, which in turn correlates with further acquisition of histone acetylation at the TSS and in the body of the regulated genes. Importantly, I observed no evidence of SMAD2-induced transcription via a Pol II pause-release mechanism. Finally, I have established that long-term modulation of SMAD2 transcriptional responses requires

continuous NODAL/Activin signalling, providing functional relevance to the observation that SMAD2 remains bound at enhancers even though transcription of target genes is frequently attenuated or terminated during prolonged signalling (Figure 7.1). Overall, my results suggest that SMAD2 sequentially recruits multiple co-factors which in turn modulate the programme of gene expression over time, and footprint analyses of the ATAC-seq dataset are currently on-going to identify the key players of this transcriptional network. Here, I will briefly discuss the more relevant aspects of my study in the context of the transcription and development fields, outlining future directions of research and general implications of my findings.

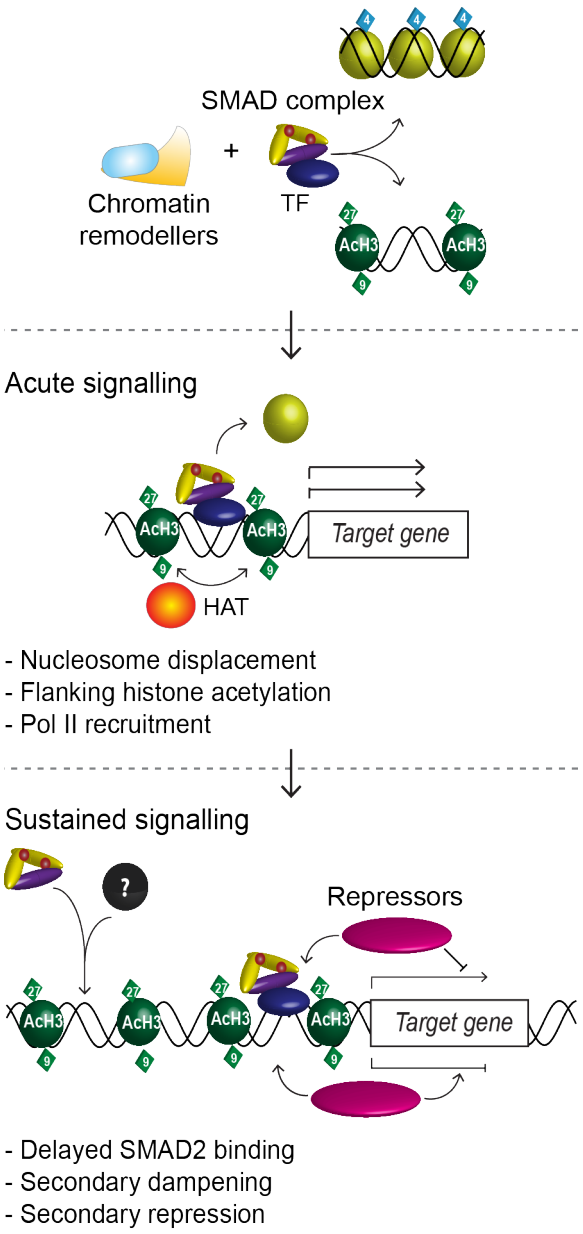


Figure 7.1. A dynamic model for SMAD2-dependent transcription. Depicted are the two modes of SMAD2 binding upon Activin stimulation from the SB-431542 state. SMAD2 can target either acetylated nucleosome-depleted chromatin or closed, non-acetylated loci marked by H3K4Me1. SMAD2 binds together with FOXH1 at some targets or with distinct TFs at others. Chromatin remodellers such as SMARCA4 are also require to mediate SMAD2 binding and nucleosome eviction at closed sites. Once bound, SMAD2-containing complexes locally increase H3K27Ac and H3K9Ac via recruitment of HATs. Upon sustained NODAL/Activin signalling, SMAD2 may be recruited to new targets in a delayed manner or already-bound SMAD2 may recruit repressors to dampen or inhibit transcription. Green diamonds, H3K27Ac and H3K9Ac; blue diamonds, H3K4me1.

7.2 Dynamics of SMAD2-mediated transcription in response to NODAL/Activin signalling

In this thesis, I showed that SMAD2-mediated transcription is gene specific, with different targets being induced or repressed with distinct temporal kinetics in response to NODAL/Activin signalling. Since around a third of genes require protein synthesis for their correct pattern of expression, it is tempting to speculate that the initial transcriptional profiles induced by Activin are subsequently remodelled by factors that are themselves encoded by SMAD2 target genes. Moreover, as on-going NODAL/Activin signalling is also required for long-term responses, it is likely that these factors directly cooperate with SMAD2 on chromatin at later times. During prolonged signalling, SMAD complexes remain bound to sites already occupied after acute stimulation, but new binding events are also observed around many target genes. In this scenario, the newly-synthesized factors could either function as transcriptional repressors recruited to chromatin by already bound SMAD2, or as co-factors mediating secondary SMAD2 binding events to otherwise inaccessible sites. The first hypothesis would provide a mechanistic explanation to the transcription attenuation or transcription repression observed at later times for many genes initially induced in response to NODAL/Activin signaling. In contrast, the second case would account for the delayed SMAD2 peaks observed at genes induced with delayed kinetics (eg. *Eomes, T*), which would ultimately be the result of a 'self-enabling' mechanism similar to what has been previously proposed for TGF-β-mediated repression of the *Id1* gene (Kang et al., 2003). In conclusion, my findings strongly suggest that NODAL/Activin signalling directly establishes a regulatory network which modulates SMAD2 transcriptional responses at later time points.

In the transcription field, TF binding events are often generally thought to linearly correlate with gene expression, and transcriptional termination is seen as the result of loss of TFs from the chromatin template. Importantly, my findings overturn this assumption, as I showed that functional SMAD2 binding does not necessarily follow the transcriptional kinetics of target genes over time. The dynamic binding behaviour of the SMADs was instead less surprising, since there are examples in the literature of SMAD complexes interacting with distinct TFs to target previously un-bound sites in different signalling conditions, for instance during ESC differentiation

(Brown et al., 2011, Faial et al., 2015, Mullen et al., 2011). In this context, however, the retention of SMAD binding at previous targets has not been investigated, and it would be interesting to test if the dampening of transcription orchestrated by SMADs at later time points in P19s is also observed in this system. Currently, I have not yet provided direct evidence for this secondary repression mechanism, neither have I identified the TFs which contribute to regulate SMAD2 transcriptional responses over time. Nevertheless, I showed that the ATAC-seq data can be successfully used to predict TF occupancy at SMAD binding sites on a genome-wide scale, and indeed I detected footprints associated with known SMAD2 interactors at some of these sites. Since this approach will allow me to identify only those SMAD2 cofactors which directly bind DNA for long period of time, I am also planning to complement the footprint-based strategy with ChIP for SMAD2 coupled with mass spectrometry, following the recently described SICAP method (Rafiee et al., 2016). By integrating these results with the RNA-seq datasets, I will finally be able to reveal the network of TFs downstream of NODAL/Activin signalling, and to pinpoint the enzymes or the other non-DNA binding proteins which are recruited to chromatin by SMAD complexes at later time points.

The finding that NODAL/Activin signalling directly orchestrates a programme of gene expression which is constantly modulated over time is also particularly relevant for *in vivo* development. Recent work in zebrafish embryos shows for example that cells which are exposed to NODAL for longer times have the greatest pSMAD2 response and give rise to endoderm, whilst cells which see the ligand for shorter times specify mesoderm mesoderm ((van Boxtel et al., 2015); van Boxtel et al., in revision)). In this context, signal duration rather than amplitude is the major determinant of cell fate specification, proving that the concept of how morphogen gradients function is worth to be reconsidered (Nahmad and Lander, 2011, Cohen et al., 2013, Ashe and Briscoe, 2006, Schier, 2009). Similarly, in P19 the transcriptional output depends more on the time of exposure to NODAL/Activin rather than on the doses of ligands used. Indeed, I observed that ligand concentration has very little effect on determining which genes are activated in response to the signalling pathway. Thus, it is possible to speculate that not just *in vitro* in P19s but also *in vivo* in zebrafish embryos NODAL signalling initiates a cascade of transcriptional events which require time and on-going signalling to be correctly implemented. Performing

time courses of ligand induction in dissociated cells obtained from zebrafish embryos could be highly informative to verify this hypothesis.

Another remarkable result for both the transcription and developmental fields is that SMAD complexes activate transcription of all target genes via a unique mechanism, which is by inducing *de novo* Pol II recruitment. Since Pol II pause-release has been commonly associated with genes that require to be rapidly switched on in response to extracellular stimuli, intuitively it could be expected that the acute induction of NODAL/Activin targets also occurred by triggering the release of paused Pol II from their promoters. Classical examples of signal-induced transcription via such a mechanism are the heat shock genes in *Drosophila*, or the immediate early genes which respond to serum growth factors in mammals (Liu et al., 2015). Nevertheless, in P19s Pol II is *de novo* recruited at the promoters of genes that are rapidly induced in response to NODAL/Activin signalling, and in many cases a high number of transcripts is produced within 1 hr starting from an undetectable mRNA baseline, as measured for the *Lefty1* target gene. Importantly, this observation fits with more recent evidence suggesting that the amount of paused Pol II at the transcription start sites inversely correlates with the rate of mRNA synthesis ((Ehrensberger et al., 2013); Patrick Cramer, personal communication). In light of these studies, it is tempting to speculate that SMAD2 regulates Pol II via *de novo* recruitment rather than via pause-release since only the first mechanism guarantees the production of a high amount of transcripts within a short period of time, which in turn could be essential to quickly establishes the programme of gene expression downstream of NODAL/Activin signalling.

Beside the stress-response genes, Pol II pausing has also been observed at the promoters of poised developmental genes during maintenance of pluripotency (Adelman and Lis, 2012). Indeed, activated SMAD2/3 was shown to trigger Pol II release at NODAL/Activin targets such as *Eomes* in the context of hESC differentiation (Estaras et al., 2015). In P19s, however, I found no paused Pol II at the classical developmental genes induced by the SMAD pathway, suggesting that NODAL/Activin signalling could regulate Pol II activity via different mechanisms in distinct cell types. Alternatively, the presence of paused Pol II in ESCs could somehow be related to the unique chromatin landscape of these cells, where promoters of developmental genes are bivalently marked with both activating and repressive histone modifications (Bernstein et al., 2006, Buecker and Wysocka,

2012). In contrast, in P19s there is very little H3K27me3 at these sites in unstimulated conditions, and it would be interesting to test if SMAD complexes regulate Pol II via *de novo* recruitment also in other cell lines with chromatin characteristics similar to P19s.

7.3 SMAD2-induced chromatin remodelling

By analyzing changes in histone modifications and chromatin accessibility in response to NODAL/Activin signalling, I showed for the first time that at some target loci SMAD2 binds to inactive chromatin. There, it directly induces nucleosome eviction and histone H3 acetylation at flanking sites with rapid kinetics. This result was rather surprising, since it was previously suggested that SMAD2/3 only passively targets pre-acetylated nucleosome-devoid chromatin, where it gets recruited by pre-bound master TFs (Mullen et al., 2011). In that study, however, the authors investigated the chromatin landscape focusing solely on constitutive signalling states. Crucially, I was able to overturn this dogma and uncover an alternative mode of SMAD2 binding because the ChIP-seq and ATAC-seq experiments were performed over a time course of ligand induction starting from a signal-inhibited baseline. My finding that SMAD2 binds and remodels inactive chromatin it is also highly relevant for the transcription field in general. In fact, if the deposition of epigenetic marks is cause or consequence of TF binding it is still a matter of debate (Spitz and Furlong, 2012). Here, I unequivocally showed that transcription factor binding precedes histone modifications and nucleosome displacement.

Since I found no evidence for a pioneer factor stably bound prior to SMAD2 recruitment at silent chromatin, the question of how SMAD complexes recognise these sites in the first place still remains. Importantly, work performed in differentiated macrophages showed that stimulus-activated TFs can bind to inactive loci marked only by H3K4me1, which the authors refer to with the term ‘latent enhancers’ (Ostuni et al., 2013). My data suggest that such latent enhancers could also exist in P19, since I demonstrated that H3K4me1 is present at the *Lefty1* SBS both in the SB-431542 state and after acute and chronic Activin induction. The ATAC-seq experiment also revealed that despite the latent SBSs being clearly occupied by nucleosomes in the absence of signalling, a low level of chromatin accessibility distinguish these sites from flanking genomic regions prior to SMADs binding. It is

tempting to speculate that this could be due to the presence of H3K4me1, as already discussed alongside other possible mechanisms in Section 6.3.2. To better clarify the role of H3K4me1 in the context of SMAD chromatin binding, it would be highly informative to investigate the genome-wide distribution of H3K4me1 in different NODAL/Activin signalling conditions with respect to the ATAC-seq and SMAD2 ChIP seq. If the existence of latent enhancers in P19 was confirmed, it would still remain to be addressed how the H3K4me1 mark is established at these sites in the first place.

In addition to the latent enhancers, I found many examples of SMAD2 target loci that were already nucleosome depleted and flanked by acetylated histone H3, as observed for the *Pou5f1* and *Trh* SBSs. In these cases, the genes exhibit a baseline of expression, indicating that TFs already occupy these sites and likely facilitate SMAD2 recruitment to chromatin. At both latent and active enhancers, SMAD2 binding correlates with an acute induction of H3K27Ac and H3K9Ac at one or two nucleosomes at either side of the SBSs, suggesting that SMAD2 directly recruits HATs at its target sites. Signalling-induced H3K27Ac is likely to be mediated by EP300, since this enzyme is enriched together with SMAD2 at some of the SBSs. The mechanism behind the acquisition of H3K9Ac still needs to be addressed. Possible candidates are the HATs GCN5/PCAF or TIP60, and interactions of SMAD complexes with PCAF and GCN5 have already been reported (Calo and Wysocka, 2013, Ross and Hill, 2008). Since the two histone marks are induced with similar kinetics, it is therefore possible that several distinct HAT complexes are simultaneously recruited by the SMADs to mediate localised H3 acetylation.

7.4 The role of SMAD cooperating factors

Once I established the ability of SMAD2 to bind inactive chromatin, I then investigated the mechanisms underlying it. I first focused on the role of the well characterised SMAD cofactor FOXH1, since it was required for SMAD2 binding to a subset of latent enhancers and for the induction of the corresponding genes (eg. *Lefty1, Lefty2*). Initially, I had hypothesized that FOXH1 would act as a pioneer factor, since it is part of the Forkhead family of TFs whose members frequently function as such (Chen et al., 1996, Randall et al., 2002, Silvestri et al., 2008, Spitz and Furlong, 2012). However, I found that there is no pre-bound FOXH1 at closed SBSs in the

absence of signalling, but it binds together with SMAD2 in a NODAL/Activin-inducible manner. Moreover, FOXH1-SMAD complexes alone are not able to displace nucleosomes are these loci, and require the chromatin remodelling activity of the SWI/SNF ATPase SMARCA4 to do so. SMARCA4 has previously been shown to mediate the induction of a group of TGF-β genes, but which step from SMAD binding to transcription activation depended on SMARCA4 was still unknown (Ross et al., 2006, Xi et al., 2008). Here, I clearly demonstrated that SMARCA4 is a pre-requisite for SMAD containing complexes to bind a subset of unacetylated, nucleosome enriched target sites. Considering the absence of TF pioneering activities and the requirement for chromatin remodelling complexes, I therefore concluded that SMAD2 may bind chromatin through what has recently been termed the dynamic assisted loading mechanism. In this model, TFs binding is achieved cooperatively through rapid interactions with remodelling complexes and the chromatin substrate. As a result, TFs dynamically cycle on and off chromatin, likely explaining why some TFs leave poorly detectable footprints on the DNA, as I observed in the case of FOXH1. Crucially, the dependency on SMARCA4 for SMAD2 binding was locus-specific. For instance, SMARCA4 was not required at the *Pitx2* SBS, even though this locus exhibits some of the same characteristics as the *Lefty1* and *Lefty2* SBSs (FOXH1-dependent and inactive chromatin at the SBS). This suggests that SMAD2 employs distinct chromatin remodelers at different target sites, and that additional factors are involved in their recruitment. Indeed, some evidence already exists for SMADs interacting with another nucleosome remodelling factor beside SMARCA4, that is the subunit of the NURF complex BPTF (Landry et al., 2008). In the future, I propose to take advantage of the SICAP approach to better characterise the full repertoire of chromatin remodelling complexes used by SMAD2, and to test their functional relevance for SMAD-mediated transcriptional responses.

Finally, it is important to highlight that FOXH1 is clearly not the only TF that cooperates with SMAD2 at unacetylated, closed chromatin sites. FOXH1, for instance, it is not required to mediate nucleosome eviction and SMAD2 binding at the upstream enhancer of the *Pmepa1* gene, which must be dependent on a yet unidentified TF. As outlined in more detail in Section 6.3.3, I am confident that analysing changes in TF footprints at SMAD2 binding sites over time and performing ChIP for SMAD2 coupled with mass spectrometry will provide me with a reliable list of candidates to further screen using the CRISPR/Cas9 system.

7.5 Conclusions and findings implications

In this thesis, I delineated for the first time the sequence of events that occur from SMAD2 binding to transcriptional activation, and the mechanisms underlying them. I also demonstrated that NODAL/Activin signalling directly orchestrate a complex programme of gene expression which is constantly remodelled over time. Finally, I laid the foundations to identify the key molecular players which allow cells to correctly execute SMAD2 transcriptional responses.

Considering the broad spectrum of biological processes controlled by NODAL/Activin signalling in both normal and pathological contexts, I am confident that my findings are highly valuable for several areas of research (Caja et al., 2012, Massague, 2012, Pauklin and Vallier, 2015, Wu and Hill, 2009). Determining universal mechanisms of SMAD mediated transcription is crucial to elucidate how NODAL and Activin ligands regulate the balance between pluripotency and differentiation in ES cells, reprogramming in induced pluripotent stem cells (iPS), homeostasis in differentiated cells and how deregulated signalling leads to cancer. NODAL/Activin signalling is now emerging to have prominent roles in both tumour development and metastasis (Wakefield and Hill, 2013). In particular, Activin has recently been shown to promote tumour progression in melanoma, skin and pancreatic cancer, and therefore represents a very attractive target for anticancer therapies (Donovan et al., 2017, Lonardo et al., 2011, Togashi et al., 2015, Antsiferova et al., 2011). Identifying the enzymes and the TFs required for SMAD transcriptional activity could ultimately have a high translational value, by providing a list of candidates which are likely crucial to establish the malignant phenotype driven by aberrant NODAL/Activin signalling.

Chapter 8. Appendices

8.1 List of the high confidence SMAD2 consensus peaks

	Peak information			
Peak ID	**chr**	**peak start**	**peak end**	**Associated gene**
consensus_chr2_167058802_167059178	chr2	167058802	167059178	1500012F01Rik
consensus_chr2_167063163_167063368	chr2	167063163	167063368	1500012F01Rik
consensus_chr6_58595576_58595877	chr6	58595576	58595877	Abcg2
consensus_chr4_126465731_126466196	chr4	126465731	126466196	Ago1
consensus_chr16_8788044_8788566	chr16	8788044	8788566	AK013883
consensus_chr16_8769190_8769359	chr16	8769190	8769359	AK013883
consensus_chr16_8778570_8779080	chr16	8778570	8779080	AK013883
consensus_chr16_8829488_8829637	chr16	8829488	8829637	AK013883
consensus_chr8_4436040_4436209	chr8	4436040	4436209	AK030646
consensus_chr8_4436476_4436884	chr8	4436476	4436884	AK030646
consensus_chr8_4582747_4582992	chr8	4582747	4582992	AK030646
consensus_chr3_89800149_89800499	chr3	89800149	89800499	AK040741
consensus_chr3_89836907_89837218	chr3	89836907	89837218	AK040741
consensus_chr3_89836176_89836827	chr3	89836176	89836827	AK040741
consensus_chr7_81480831_81481312	chr7	81480831	81481312	Ap3b2
consensus_chr7_81497209_81497360	chr7	81497209	81497360	Ap3b2
consensus_chr1_191120954_191121020	chr1	191120954	191121020	Atf3
consensus_chr1_191190073_191190359	chr1	191190073	191190359	Atf3
consensus_chr3_122434049_122434426	chr3	122434049	122434426	Bcar3
consensus_chr3_122414149_122414544	chr3	122414149	122414544	Bcar3
consensus_chr6_3677836_3678529	chr6	3677836	3678529	Calcr
consensus_chr6_3623904_3624150	chr6	3623904	3624150	Calcr
consensus_chr6_127106877_127107026	chr6	127106877	127107026	Ccnd2
consensus_chr10_43551585_43551877	chr10	43551585	43551877	Cd24a
consensus_chr10_44553532_44553939	chr10	44553532	44553939	Cd24a
consensus_chr10_44392015_44392352	chr10	44392015	44392352	Cd24a
consensus_chr10_43481135_43481437	chr10	43481135	43481437	Cd24a
consensus_chr2_104109533_104109799	chr2	104109533	104109799	Cd59a
consensus_chr2_104188245_104188452	chr2	104188245	104188452	Cd59a
consensus_chr2_104122629_104122782	chr2	104122629	104122782	Cd59a
consensus_chr17_29093566_29093896	chr17	29093566	29093896	Cdkn1a
consensus_chr17_29007432_29007640	chr17	29007432	29007640	Cdkn1a
consensus_chr17_29008194_29008596	chr17	29008194	29008596	Cdkn1a
consensus_chr17_29032640_29032874	chr17	29032640	29032874	Cdkn1a
consensus_chr5_147304967_147305067	chr5	147304967	147305067	Cdx2
consensus_chr2_126521591_126521758	chr2	126521591	126521758	Chac1
consensus_chr2_126523532_126523676	chr2	126523532	126523676	Chac1
consensus_chr2_119304534_119304780	chr2	119304534	119304780	Chac1
consensus_chr2_131152139_131152505	chr2	131152139	131152505	Chac1
consensus_chr2_119288567_119288727	chr2	119288567	119288727	Chac1
consensus_chr2_119351226_119351442	chr2	119351226	119351442	Chac1
consensus_chr7_132317113_132317585	chr7	132317113	132317585	Chst15
consensus_chr7_132403936_132404243	chr7	132403936	132404243	Chst15
consensus_chr7_132360099_132360285	chr7	132360099	132360285	Chst15
consensus_chr7_132247588_132247876	chr7	132247588	132247876	Chst15

consensus_chr7_132320651_132320895	chr7	132320651	132320895	Chst15
consensus_chr3_87974317_87974567	chr3	87974317	87974567	Crabp2
consensus_chr3_86566305_86566532	chr3	86566305	86566532	Crabp2
consensus_chr3_87970999_87971247	chr3	87970999	87971247	Crabp2
consensus_chr3_87891547_87891945	chr3	87891547	87891945	Crabp2
consensus_chr6_115252107_115252377	chr6	115252107	115252377	Cxcl12
consensus_chr6_117151734_117152782	chr6	117151734	117152782	Cxcl12
consensus_chr6_115287121_115287449	chr6	115287121	115287449	Cxcl12
consensus_chr6_117180164_117180610	chr6	117180164	117180610	Cxcl12
consensus_chr3_145649846_145649978	chr3	145649846	145649978	Cyr61
consensus_chr3_145690958_145691084	chr3	145690958	145691084	Cyr61
consensus_chr7_3213628_3213837	chr7	3213628	3213837	D7Ertd143e
consensus_chr7_3314282_3314409	chr7	3314282	3314409	D7Ertd143e
consensus_chr7_3206524_3206886	chr7	3206524	3206886	D7Ertd143e
consensus_chr7_3217652_3217973	chr7	3217652	3217973	D7Ertd143e
consensus_chr7_62420131_62420290	chr7	62420131	62420290	D7Ertd715e
consensus_chr12_71369195_71369340	chr12	71369195	71369340	Dact1
consensus_chr12_71369922_71370116	chr12	71369922	71370116	Dact1
consensus_chr12_71296373_71297167	chr12	71296373	71297167	Dact1
consensus_chr12_71376627_71376890	chr12	71376627	71376890	Dact1
consensus_chr12_71301571_71302020	chr12	71301571	71302020	Dact1
consensus_chr10_59957919_59958288	chr10	59957919	59958288	Ddit4
consensus_chr10_59962644_59962885	chr10	59962644	59962885	Ddit4
consensus_chr10_59951664_59951902	chr10	59951664	59951902	Ddit4
consensus_chr10_59988562_59988702	chr10	59988562	59988702	Ddit4
consensus_chr17_15380813_15380866	chr17	15380813	15380866	Dll1
consensus_chr17_15380676_15380801	chr17	15380676	15380801	Dll1
consensus_chr2_170772037_170772607	chr2	170772037	170772607	Dok5
consensus_chr2_170657559_170658226	chr2	170657559	170658226	Dok5
consensus_chr2_170887653_170887870	chr2	170887653	170887870	Dok5
consensus_chr2_170726216_170726396	chr2	170726216	170726396	Dok5
consensus_chr17_26511210_26511798	chr17	26511210	26511798	Dusp1
consensus_chr8_34805012_34805408	chr8	34805012	34805408	Dusp4
consensus_chr8_34732699_34733049	chr8	34732699	34733049	Dusp4
consensus_chr10_99170741_99170917	chr10	99170741	99170917	Dusp6
consensus_chr10_99170969_99171036	chr10	99170969	99171036	Dusp6
consensus_chr10_99261290_99261430	chr10	99261290	99261430	Dusp6
consensus_chrX_106080918_106081259	chrX	106080918	106081259	Efnb1
consensus_chrX_139436747_139436935	chrX	139436747	139436935	Efnb1
consensus_chrX_139345212_139345622	chrX	139345212	139345622	Efnb1
consensus_chrX_99045496_99045800	chrX	99045496	99045800	Efnb1
consensus_chrX_99152773_99152929	chrX	99152773	99152929	Efnb1
consensus_chrX_139317424_139317594	chrX	139317424	139317594	Efnb1
consensus_chrX_145524466_145524856	chrX	145524466	145524856	Efnb1
consensus_chr18_34850208_34850354	chr18	34850208	34850354	Egr1
consensus_chr9_118389946_118390214	chr9	118389946	118390214	Eomes
consensus_chr9_118384752_118384964	chr9	118384752	118384964	Eomes
consensus_chr9_118432551_118433292	chr9	118432551	118433292	Eomes
consensus_chr9_118487571_118487804	chr9	118487571	118487804	Eomes
consensus_chr9_118455982_118456509	chr9	118455982	118456509	Eomes
consensus_chr9_118486783_118487207	chr9	118486783	118487207	Eomes
consensus_chr9_118468200_118468630	chr9	118468200	118468630	Eomes
consensus_chr4_141291432_141292027	chr4	141291432	141292027	Epha2
consensus_chr4_141338992_141339132	chr4	141338992	141339132	Epha2
consensus_chr4_141284266_141284408	chr4	141284266	141284408	Epha2
consensus_chr4_141284518_141284750	chr4	141284518	141284750	Epha2
consensus_chr4_140126698_140126946	chr4	140126698	140126946	Epha2
consensus_chr4_140131460_140131654	chr4	140131460	140131654	Epha2
consensus_chr7_144872813_144872953	chr7	144872813	144872953	Fgf15
consensus_chr7_144898506_144898931	chr7	144898506	144898931	Fgf15

consensus_chr7_144888718_144888877	chr7	144888718	144888877	Fgf15
consensus_chr5_98293510_98293685	chr5	98293510	98293685	Fgf5
consensus_chr5_98310053_98310554	chr5	98310053	98310554	Fgf5
consensus_chr5_98167646_98167808	chr5	98167646	98167808	Fgf5
consensus_chr5_98312032_98312336	chr5	98312032	98312336	Fgf5
consensus_chr5_98283639_98283918	chr5	98283639	98283918	Fgf5
consensus_chr5_98255741_98256081	chr5	98255741	98256081	Fgf5
consensus_chr19_45733065_45733956	chr19	45733065	45733956	Fgf8
consensus_chr19_45745737_45746430	chr19	45745737	45746430	Fgf8
consensus_chr19_45743469_45743630	chr19	45743469	45743630	Fgf8
consensus_chr19_45740924_45741066	chr19	45740924	45741066	Fgf8
consensus_chr19_45783535_45783679	chr19	45783535	45783679	Fgf8
consensus_chr19_45790984_45791161	chr19	45790984	45791161	Fgf8
consensus_chr7_130327950_130328203	chr7	130327950	130328203	Fgfr2
consensus_chr7_130264662_130264897	chr7	130264662	130264897	Fgfr2
consensus_chr7_130252207_130252429	chr7	130252207	130252429	Fgfr2
consensus_chr4_124708779_124709208	chr4	124708779	124709208	Fhl3
consensus_chr17_47887305_47887876	chr17	47887305	47887876	Foxp4
consensus_chr17_47953108_47953293	chr17	47953108	47953293	Foxp4
consensus_chr13_114406548_114406714	chr13	114406548	114406714	Fst
consensus_chr13_114406841_114407585	chr13	114406841	114407585	Fst
consensus_chr5_128596830_128597360	chr5	128596830	128597360	Fzd10
consensus_chr5_128598418_128598625	chr5	128598418	128598625	Fzd10
consensus_chr5_128601637_128601965	chr5	128601637	128601965	Fzd10
consensus_chr5_128600543_128600829	chr5	128600543	128600829	Fzd10
consensus_chr5_128593379_128593636	chr5	128593379	128593636	Fzd10
consensus_chr5_128603194_128603404	chr5	128603194	128603404	Fzd10
consensus_chr5_128579190_128579385	chr5	128579190	128579385	Fzd10
consensus_chr13_51853524_51853585	chr13	51853524	51853585	Gadd45g
consensus_chr13_51846534_51846659	chr13	51846534	51846659	Gadd45g
consensus_chr13_51845953_51846130	chr13	51845953	51846130	Gadd45g
consensus_chr4_146498084_146498487	chr4	146498084	146498487	Gm13247
consensus_chr17_22745725_22746266	chr17	22745725	22746266	Gm16386
consensus_chr10_80050827_80051622	chr10	80050827	80051622	Gpx4
consensus_chr10_80053637_80053790	chr10	80053637	80053790	Gpx4
consensus_chr10_79988724_79988849	chr10	79988724	79988849	Gpx4
consensus_chr5_88679333_88679917	chr5	88679333	88679917	Grsf1
consensus_chr5_88684336_88684662	chr5	88684336	88684662	Grsf1
consensus_chr5_88670674_88670933	chr5	88670674	88670933	Grsf1
consensus_chr12_111664282_111664372	chr12	111664282	111664372	Gsc
consensus_chr12_104466887_104467386	chr12	104466887	104467386	Gsc
consensus_chr12_104470812_104470973	chr12	104470812	104470973	Gsc
consensus_chr12_111672141_111672473	chr12	111672141	111672473	Gsc
consensus_chr12_104381222_104381480	chr12	104381222	104381480	Gsc
consensus_chr12_104473558_104473733	chr12	104473558	104473733	Gsc
consensus_chr6_108696365_108696528	chr6	108696365	108696528	Gt(ROSA)26S
consensus_chr6_108644376_108644559	chr6	108644376	108644559	Gt(ROSA)26S
consensus_chr6_113077039_113077254	chr6	113077039	113077254	Gt(ROSA)26S
consensus_chr15_56742758_56743078	chr15	56742758	56743078	Has2
consensus_chr15_56708427_56708607	chr15	56708427	56708607	Has2
consensus_chr15_56627608_56627951	chr15	56627608	56627951	Has2
consensus_chr15_56626464_56626771	chr15	56626464	56626771	Has2
consensus_chr15_56729756_56730065	chr15	56729756	56730065	Has2
consensus_chr15_54702219_54702825	chr15	54702219	54702825	Has2
consensus_chr16_30155736_30156494	chr16	30155736	30156494	Hes1
consensus_chr16_30129541_30129713	chr16	30129541	30129713	Hes1
consensus_chr16_30160400_30160832	chr16	30160400	30160832	Hes1
consensus_chr16_30103122_30103289	chr16	30103122	30103289	Hes1
consensus_chr16_33605791_33606065	chr16	33605791	33606065	Hes1
consensus_chr16_33845586_33845726	chr16	33845586	33845726	Hes1

consensus_chr16_30062944_30063357	chr16	30062944	30063357	Hes1
consensus_chr16_33737816_33738120	chr16	33737816	33738120	Hes1
consensus_chr16_33844391_33844583	chr16	33844391	33844583	Hes1
consensus_chr11_90356160_90356329	chr11	90356160	90356329	Hlf
consensus_chr11_90371054_90371268	chr11	90371054	90371268	Hlf
consensus_chr11_90391688_90391958	chr11	90391688	90391958	Hlf
consensus_chr11_96341761_96342424	chr11	96341761	96342424	Hoxb1
consensus_chr11_96368741_96368886	chr11	96368741	96368886	Hoxb1
consensus_chr2_152736335_152736401	chr2	152736335	152736401	Id1
consensus_chr2_152735131_152735359	chr2	152735131	152735359	Id1
consensus_chr12_25098494_25098699	chr12	25098494	25098699	Id2
consensus_chr4_136140369_136140581	chr4	136140369	136140581	Id3
consensus_chr4_138453360_138453557	chr4	138453360	138453557	Id3
consensus_chr4_138361509_138362033	chr4	138361509	138362033	Id3
consensus_chr4_135963498_135963678	chr4	135963498	135963678	Id3
consensus_chr4_136031505_136031958	chr4	136031505	136031958	Id3
consensus_chr4_138396228_138396353	chr4	138396228	138396353	Id3
consensus_chr4_138444922_138445231	chr4	138444922	138445231	Id3
consensus_chr4_138418538_138418779	chr4	138418538	138418779	Id3
consensus_chr4_136053302_136053447	chr4	136053302	136053447	Id3
consensus_chr2_30474056_30474199	chr2	30474056	30474199	Ier5l
consensus_chr1_72805935_72806628	chr1	72805935	72806628	Igfbp2
consensus_chr14_27044889_27045392	chr14	27044889	27045392	Il17rd
consensus_chr14_27044304_27044591	chr14	27044304	27044591	Il17rd
consensus_chr14_27039338_27039517	chr14	27039338	27039517	Il17rd
consensus_chr14_25718331_25718953	chr14	25718331	25718953	Il17rd
consensus_chr14_25607054_25607197	chr14	25607054	25607197	Il17rd
consensus_chr14_27096185_27096406	chr14	27096185	27096406	Il17rd
consensus_chr14_27066406_27066795	chr14	27066406	27066795	Il17rd
consensus_chr2_146302378_146302639	chr2	146302378	146302639	Insm1
consensus_chr11_53759684_53760389	chr11	53759684	53760389	Irf1
consensus_chr11_53826945_53827092	chr11	53826945	53827092	Irf1
consensus_chr11_55461054_55461179	chr11	55461054	55461179	Irf1
consensus_chr11_53799980_53800464	chr11	53799980	53800464	Irf1
consensus_chr11_53786762_53787057	chr11	53786762	53787057	Irf1
consensus_chr11_55438305_55438441	chr11	55438305	55438441	Irf1
consensus_chr11_53781310_53781506	chr11	53781310	53781506	Irf1
consensus_chr11_48871235_48871651	chr11	48871235	48871651	Irgm1
consensus_chr11_48818441_48818699	chr11	48818441	48818699	Irgm1
consensus_chr11_48826804_48827096	chr11	48826804	48827096	Irgm1
consensus_chr11_48853597_48853755	chr11	48853597	48853755	Irgm1
consensus_chr11_48800410_48800567	chr11	48800410	48800567	Irgm1
consensus_chr11_48856851_48857129	chr11	48856851	48857129	Irgm1
consensus_chr11_48833788_48834058	chr11	48833788	48834058	Irgm1
consensus_chr11_48834252_48834529	chr11	48834252	48834529	Irgm1
consensus_chr9_58295776_58296231	chr9	58295776	58296231	Islr2
consensus_chr9_58279246_58279681	chr9	58279246	58279681	Islr2
consensus_chr2_155115402_155116063	chr2	155115402	155116063	Itch
consensus_chr2_155074251_155074431	chr2	155074251	155074431	Itch
consensus_chr2_155133455_155133700	chr2	155133455	155133700	Itch
consensus_chr2_155053130_155053272	chr2	155053130	155053272	Itch
consensus_chr4_95134823_95135014	chr4	95134823	95135014	Jun
consensus_chr4_95052482_95052624	chr4	95052482	95052624	Jun
consensus_chr4_95052028_95052213	chr4	95052028	95052213	Jun
consensus_chr4_142166133_142166914	chr4	142166133	142166914	Kazn
consensus_chr4_142197117_142197286	chr4	142197117	142197286	Kazn
consensus_chr4_142140118_142140268	chr4	142140118	142140268	Kazn
consensus_chr9_4283761_4284275	chr9	4283761	4284275	Kbtbd3
consensus_chr9_4226971_4227211	chr9	4226971	4227211	Kbtbd3
consensus_chr17_83694920_83695176	chr17	83694920	83695176	Kcng3

consensus_chr17_83668253_83668509	chr17	83668253	83668509	Kcng3
consensus_chr17_83547624_83548022	chr17	83547624	83548022	Kcng3
consensus_chr17_83633016_83633583	chr17	83633016	83633583	Kcng3
consensus_chr17_83526238_83526490	chr17	83526238	83526490	Kcng3
consensus_chr15_62086921_62087198	chr15	62086921	62087198	Kcnq3
consensus_chr15_40955266_40955765	chr15	40955266	40955765	Klf10
consensus_chr15_41162279_41162512	chr15	41162279	41162512	Klf10
consensus_chr15_38300483_38300665	chr15	38300483	38300665	Klf10
consensus_chr13_5892395_5892635	chr13	5892395	5892635	Klf6
consensus_chr6_145275740_145276846	chr6	145275740	145276846	Kras
consensus_chr6_145275052_145275342	chr6	145275052	145275342	Kras
consensus_chr6_145261829_145262239	chr6	145261829	145262239	Kras
consensus_chr6_144361308_144361462	chr6	144361308	144361462	Kras
consensus_chr6_144454403_144454693	chr6	144454403	144454693	Kras
consensus_chr6_145252001_145252289	chr6	145252001	145252289	Kras
consensus_chr6_144373492_144373661	chr6	144373492	144373661	Kras
consensus_chr6_144251210_144251401	chr6	144251210	144251401	Kras
consensus_chr6_144273631_144273854	chr6	144273631	144273854	Kras
consensus_chr4_98727954_98728510	chr4	98727954	98728510	L1td1
consensus_chr4_98726663_98726945	chr4	98726663	98726945	L1td1
consensus_chr1_180924394_180925384	chr1	180924394	180925384	Lefty1
consensus_chr1_180933629_180934184	chr1	180933629	180934184	Lefty1
consensus_chr1_180934401_180935253	chr1	180934401	180935253	Lefty1
consensus_chr1_181341049_181341199	chr1	181341049	181341199	Lefty1
consensus_chr1_180887934_180889501	chr1	180887934	180889501	Lefty2
consensus_chr1_180889952_180890153	chr1	180889952	180890153	Lefty2
consensus_chr1_180890289_180890577	chr1	180890289	180890577	Lefty2
consensus_chr1_180850818_180850959	chr1	180850818	180850959	Lefty2
consensus_chr5_140583969_140584286	chr5	140583969	140584286	Lfng
consensus_chr5_140611995_140612157	chr5	140611995	140612157	Lfng
consensus_chr2_109917143_109917715	chr2	109917143	109917715	Lgr4
consensus_chr6_30711408_30711749	chr6	30711408	30711749	Mest
consensus_chr11_97661653_97662273	chr11	97661653	97662273	Mllt6
consensus_chr5_37832855_37832998	chr5	37832855	37832998	Msx1
consensus_chr5_37826777_37827028	chr5	37826777	37827028	Msx1
consensus_chr5_37768802_37769120	chr5	37768802	37769120	Msx1
consensus_chr13_53455410_53455576	chr13	53455410	53455576	Msx2
consensus_chr15_65926152_65926942	chr15	65926152	65926942	Myc
consensus_chr6_122702403_122702936	chr6	122702403	122702936	Nanog
consensus_chr6_122703628_122704129	chr6	122703628	122704129	Nanog
consensus_chr6_122715788_122716411	chr6	122715788	122716411	Nanog
consensus_chr6_122707200_122707662	chr6	122707200	122707662	Nanog
consensus_chr6_122609943_122610121	chr6	122609943	122610121	Nanog
consensus_chr6_122608433_122608717	chr6	122608433	122608717	Nanog
consensus_chr6_122624641_122624865	chr6	122624641	122624865	Nanog
consensus_chr19_5802518_5802790	chr19	5802518	5802790	Neat1
consensus_chr13_56222214_56222396	chr13	56222214	56222396	Neurog1
consensus_chr10_61418888_61419589	chr10	61418888	61419589	Nodal
consensus_chr10_61405952_61406482	chr10	61405952	61406482	Nodal
consensus_chr10_61418275_61418474	chr10	61418275	61418474	Nodal
consensus_chr10_61415682_61416344	chr10	61415682	61416344	Nodal
consensus_chr10_61500857_61500996	chr10	61500857	61500996	Nodal
consensus_chr7_30463526_30464004	chr7	30463526	30464004	Nphs1os
consensus_chr7_30464066_30464241	chr7	30464066	30464241	Nphs1os
consensus_chr12_44403632_44404236	chr12	44403632	44404236	Nrcam
consensus_chr12_44327670_44327856	chr12	44327670	44327856	Nrcam
consensus_chr12_44535560_44536334	chr12	44535560	44536334	Nrcam
consensus_chr12_44393882_44394045	chr12	44393882	44394045	Nrcam
consensus_chr14_48754197_48754367	chr14	48754197	48754367	Otx2
consensus_chr14_48753109_48753424	chr14	48753109	48753424	Otx2

consensus_chr14_48700916_48701209	chr14	48700916	48701209	Otx2
consensus_chr14_48662913_48663201	chr14	48662913	48663201	Otx2
consensus_chr14_49155640_49155812	chr14	49155640	49155812	Otx2
consensus_chr14_48654960_48655406	chr14	48654960	48655406	Otx2
consensus_chr14_49089813_49089999	chr14	49089813	49089999	Otx2
consensus_chr14_49275422_49275653	chr14	49275422	49275653	Otx2
consensus_chr14_48653661_48653928	chr14	48653661	48653928	Otx2
consensus_chr14_48707986_48708322	chr14	48707986	48708322	Otx2
consensus_chr14_48709413_48709626	chr14	48709413	48709626	Otx2
consensus_chr18_38210652_38210850	chr18	38210652	38210850	Pcdh1
consensus_chr18_38278836_38278976	chr18	38278836	38278976	Pcdh1
consensus_chr6_4736810_4736968	chr6	4736810	4736968	Peg10
consensus_chr6_4855296_4855436	chr6	4855296	4855436	Peg10
consensus_chr6_4747369_4747494	chr6	4747369	4747494	Peg10
consensus_chr6_4739948_4740139	chr6	4739948	4740139	Peg10
consensus_chr6_122340368_122340910	chr6	122340368	122340910	Phc1
consensus_chr6_122342170_122342533	chr6	122342170	122342533	Phc1
consensus_chr6_122366465_122366649	chr6	122366465	122366649	Phc1
consensus_chr7_143467822_143467974	chr7	143467822	143467974	Phlda2
consensus_chr11_3337600_3338123	chr11	3337600	3338123	Pik3ip1
consensus_chr11_3292790_3292951	chr11	3292790	3292951	Pik3ip1
consensus_chr17_29451085_29451569	chr17	29451085	29451569	Pim1
consensus_chr3_129217178_129217766	chr3	129217178	129217766	Pitx2
consensus_chr3_129222436_129222934	chr3	129222436	129222934	Pitx2
consensus_chr3_129216817_129217096	chr3	129216817	129217096	Pitx2
consensus_chr3_129217914_129218115	chr3	129217914	129218115	Pitx2
consensus_chr3_129214714_129214896	chr3	129214714	129214896	Pitx2
consensus_chr17_83166332_83167152	chr17	83166332	83167152	Pkdcc
consensus_chr17_70755079_70755371	chr17	70755079	70755371	Pkdcc
consensus_chr17_83320515_83320884	chr17	83320515	83320884	Pkdcc
consensus_chr17_83214236_83214528	chr17	83214236	83214528	Pkdcc
consensus_chr17_70900626_70901022	chr17	70900626	70901022	Pkdcc
consensus_chr17_70737468_70737608	chr17	70737468	70737608	Pkdcc
consensus_chr17_70772484_70772902	chr17	70772484	70772902	Pkdcc
consensus_chr10_3851942_3852301	chr10	3851942	3852301	Plekhg1
consensus_chr10_3741224_3742156	chr10	3741224	3742156	Plekhg1
consensus_chr10_3941333_3941499	chr10	3941333	3941499	Plekhg1
consensus_chr2_173334398_173335678	chr2	173334398	173335678	Pmepa1
consensus_chr2_173339932_173340353	chr2	173339932	173340353	Pmepa1
consensus_chr2_173257969_173258380	chr2	173257969	173258380	Pmepa1
consensus_chr2_173238334_173239057	chr2	173238334	173239057	Pmepa1
consensus_chr2_173672536_173672892	chr2	173672536	173672892	Pmepa1
consensus_chr17_35504528_35506015	chr17	35504528	35506015	Pou5f1
consensus_chr17_35503853_35504299	chr17	35503853	35504299	Pou5f1
consensus_chr17_35548650_35548879	chr17	35548650	35548879	Pou5f1
consensus_chr17_35485810_35485956	chr17	35485810	35485956	Pou5f1
consensus_chr12_103564312_103565109	chr12	103564312	103565109	Ppp4r4
consensus_chr12_103599158_103599646	chr12	103599158	103599646	Ppp4r4
consensus_chr12_103600388_103600857	chr12	103600388	103600857	Ppp4r4
consensus_chr12_103511464_103511915	chr12	103511464	103511915	Ppp4r4
consensus_chr13_34741702_34742336	chr13	34741702	34742336	Pxdc1
consensus_chr13_34685053_34685628	chr13	34685053	34685628	Pxdc1
consensus_chr5_142821774_142821827	chr5	142821774	142821827	Radil
consensus_chr5_142812838_142813177	chr5	142812838	142813177	Radil
consensus_chr5_142546212_142546943	chr5	142546212	142546943	Radil
consensus_chr5_142906533_142906819	chr5	142906533	142906819	Radil
consensus_chr5_142800798_142800938	chr5	142800798	142800938	Radil
consensus_chr5_142881399_142881695	chr5	142881399	142881695	Radil
consensus_chr11_98960001_98960202	chr11	98960001	98960202	Rara
consensus_chr11_98925815_98926131	chr11	98925815	98926131	Rara

consensus_chr7_100861981_100862442	chr7	100861981	100862442	Relt
consensus_chr7_82063489_82063814	chr7	82063489	82063814	Rgma
consensus_chr7_73344144_73344284	chr7	73344144	73344284	Rgma
consensus_chr7_73399519_73399721	chr7	73399519	73399721	Rgma
consensus_chr7_82056669_82056859	chr7	82056669	82056859	Rgma
consensus_chr12_8402085_8402511	chr12	8402085	8402511	Rhob
consensus_chr2_168840980_168841183	chr2	168840980	168841183	Sall4
consensus_chr2_168838785_168839141	chr2	168838785	168839141	Sall4
consensus_chr2_168840268_168840521	chr2	168840268	168840521	Sall4
consensus_chr2_168791141_168791428	chr2	168791141	168791428	Sall4
consensus_chr2_168777065_168777396	chr2	168777065	168777396	Sall4
consensus_chr2_168780159_168780342	chr2	168780159	168780342	Sall4
consensus_chr2_168782911_168783312	chr2	168782911	168783312	Sall4
consensus_chr2_168761948_168762206	chr2	168761948	168762206	Sall4
consensus_chr13_94321144_94321833	chr13	94321144	94321833	Scamp1
consensus_chr13_94242174_94242361	chr13	94242174	94242361	Scamp1
consensus_chr19_44474304_44474435	chr19	44474304	44474435	Scd1
consensus_chr5_123136547_123137136	chr5	123136547	123137136	Setd1b
consensus_chr5_123137711_123137959	chr5	123137711	123137959	Setd1b
consensus_chr5_123140997_123141187	chr5	123140997	123141187	Setd1b
consensus_chr5_123138268_123138554	chr5	123138268	123138554	Setd1b
consensus_chr5_123245570_123245742	chr5	123245570	123245742	Setd1b
consensus_chr5_67632834_67634059	chr5	67632834	67634059	Shisa3
consensus_chr5_67696783_67697780	chr5	67696783	67697780	Shisa3
consensus_chr4_155189613_155190201	chr4	155189613	155190201	Ski
consensus_chr4_155249806_155250010	chr4	155249806	155250010	Ski
consensus_chr4_155290432_155291069	chr4	155290432	155291069	Ski
consensus_chr4_155191953_155192422	chr4	155191953	155192422	Ski
consensus_chr3_31087364_31087512	chr3	31087364	31087512	Skil
consensus_chr3_31089484_31089609	chr3	31089484	31089609	Skil
consensus_chr4_119026459_119026723	chr4	119026459	119026723	Slc2a1
consensus_chr4_119108741_119108988	chr4	119108741	119108988	Slc2a1
consensus_chr6_122766285_122766338	chr6	122766285	122766338	Slc2a3
consensus_chr6_122765209_122765420	chr6	122765209	122765420	Slc2a3
consensus_chr6_122742373_122742537	chr6	122742373	122742537	Slc2a3
consensus_chr18_75366741_75367245	chr18	75366741	75367245	Smad7
consensus_chr18_75360801_75361429	chr18	75360801	75361429	Smad7
consensus_chr18_75353497_75353669	chr18	75353497	75353669	Smad7
consensus_chr5_65525609_65526464	chr5	65525609	65526464	Smim14
consensus_chr5_65391226_65391539	chr5	65391226	65391539	Smim14
consensus_chr7_60007871_60008424	chr7	60007871	60008424	Snurf
consensus_chr7_60006072_60006972	chr7	60006072	60006972	Snurf
consensus_chr11_117941914_117942039	chr11	117941914	117942039	Socs3
consensus_chr11_117972755_117973060	chr11	117972755	117973060	Socs3
consensus_chr11_117879734_117879991	chr11	117879734	117879991	Socs3
consensus_chr1_133492580_133492862	chr1	133492580	133492862	Sox13
consensus_chr1_133493152_133493749	chr1	133493152	133493749	Sox13
consensus_chr1_133397576_133398068	chr1	133397576	133398068	Sox13
consensus_chr2_70471866_70472314	chr2	70471866	70472314	Sp5
consensus_chr2_70474647_70475140	chr2	70474647	70475140	Sp5
consensus_chr2_70563254_70563440	chr2	70563254	70563440	Sp5
consensus_chr17_17827567_17828588	chr17	17827567	17828588	Spaca6
consensus_chr17_17827285_17827425	chr17	17827285	17827425	Spaca6
consensus_chr17_17826709_17826997	chr17	17826709	17826997	Spaca6
consensus_chr17_17825762_17826184	chr17	17825762	17826184	Spaca6
consensus_chr1_57890161_57890956	chr1	57890161	57890956	Spats2l
consensus_chr1_59448450_59448696	chr1	59448450	59448696	Spats2l
consensus_chr1_57862138_57862396	chr1	57862138	57862396	Spats2l
consensus_chr2_69222031_69222562	chr2	69222031	69222562	Spc25
consensus_chr2_69150788_69151293	chr2	69150788	69151293	Spc25

consensus_chr18_38573819_38574176	chr18	38573819	38574176	Spry4
consensus_chr18_38635944_38636200	chr18	38635944	38636200	Spry4
consensus_chr17_8376105_8376842	chr17	8376105	8376842	T
consensus_chr17_8377276_8377422	chr17	8377276	8377422	T
consensus_chr17_8375801_8376001	chr17	8375801	8376001	T
consensus_chr17_8389190_8389339	chr17	8389190	8389339	T
consensus_chr17_8408979_8409262	chr17	8408979	8409262	T
consensus_chr5_117121502_117122157	chr5	117121502	117122157	Taok3
consensus_chr5_117135600_117135895	chr5	117135600	117135895	Taok3
consensus_chr5_119706806_119706869	chr5	119706806	119706869	Tbx3
consensus_chr5_119706579_119706794	chr5	119706579	119706794	Tbx3
consensus_chr5_119709831_119709975	chr5	119709831	119709975	Tbx3
consensus_chr5_119613558_119613838	chr5	119613558	119613838	Tbx3
consensus_chr9_110947137_110948383	chr9	110947137	110948383	Tdgf1
consensus_chr9_110946531_110946721	chr9	110946531	110946721	Tdgf1
consensus_chr9_110949302_110949525	chr9	110949302	110949525	Tdgf1
consensus_chr13_40744792_40744854	chr13	40744792	40744854	Tfap2a
consensus_chr9_61324251_61324332	chr9	61324251	61324332	Tle3
consensus_chr9_61293122_61293899	chr9	61293122	61293899	Tle3
consensus_chr9_61339317_61339544	chr9	61339317	61339544	Tle3
consensus_chr9_61445921_61446061	chr9	61445921	61446061	Tle3
consensus_chr9_61385167_61385319	chr9	61385167	61385319	Tle3
consensus_chr9_61361289_61361481	chr9	61361289	61361481	Tle3
consensus_chr9_61368182_61368412	chr9	61368182	61368412	Tle3
consensus_chr9_61367314_61367492	chr9	61367314	61367492	Tle3
consensus_chr9_61367852_61368066	chr9	61367852	61368066	Tle3
consensus_chr9_61372345_61372498	chr9	61372345	61372498	Tle3
consensus_chr9_67105826_67106124	chr9	67105826	67106124	Tpm1
consensus_chr6_92247122_92248225	chr6	92247122	92248225	Trh
consensus_chr6_92246269_92246732	chr6	92246269	92246732	Trh
consensus_chr6_97101730_97101881	chr6	97101730	97101881	Trh
consensus_chr6_97102282_97102637	chr6	97102282	97102637	Trh
consensus_chr6_92364805_92364976	chr6	92364805	92364976	Trh
consensus_chr6_97063168_97063423	chr6	97063168	97063423	Trh
consensus_chr6_92513869_92514072	chr6	92513869	92514072	Trh
consensus_chr6_92469737_92470184	chr6	92469737	92470184	Trh
consensus_chr2_152396579_152396712	chr2	152396579	152396712	Trib3
consensus_chr12_102758889_102759365	chr12	102758889	102759365	Ubr7
consensus_chr12_102758660_102758816	chr12	102758660	102758816	Ubr7
consensus_chr12_102702435_102702575	chr12	102702435	102702575	Ubr7
consensus_chr10_60814586_60815057	chr10	60814586	60815057	Unc5b
consensus_chr10_60793113_60793280	chr10	60793113	60793280	Unc5b
consensus_chr7_98815460_98815615	chr7	98815460	98815615	Wnt11
consensus_chr7_98840585_98841145	chr7	98840585	98841145	Wnt11
consensus_chr7_90457101_90457275	chr7	90457101	90457275	Wnt11
consensus_chr7_98837807_98838124	chr7	98837807	98838124	Wnt11
consensus_chr11_103809172_103809552	chr11	103809172	103809552	Wnt3
consensus_chr11_103810386_103810439	chr11	103810386	103810439	Wnt3
consensus_chr11_103770780_103771467	chr11	103770780	103771467	Wnt3
consensus_chr11_103808652_103808831	chr11	103808652	103808831	Wnt3
consensus_chr11_103772298_103772452	chr11	103772298	103772452	Wnt3
consensus_chr11_103840947_103841240	chr11	103840947	103841240	Wnt3
consensus_chr11_103798585_103798879	chr11	103798585	103798879	Wnt3
consensus_chr18_34555243_34555601	chr18	34555243	34555601	Wnt8a
consensus_chr18_34542264_34542435	chr18	34542264	34542435	Wnt8a
consensus_chr18_34549061_34549615	chr18	34549061	34549615	Wnt8a
consensus_chr18_34551614_34552178	chr18	34551614	34552178	Wnt8a
consensus_chr11_5562956_5563289	chr11	5562956	5563289	Xbp1
consensus_chr17_84149377_84149961	chr17	84149377	84149961	Zfp36l2
consensus_chr17_84152415_84152592	chr17	84152415	84152592	Zfp36l2

consensus_chr17_84177535_84177714	chr17	84177535	84177714	Zfp36l2
consensus_chr11_100527735_100527924	chr11	100527735	100527924	Zfp385c
consensus_chr11_100713769_100714028	chr11	100713769	100714028	Zfp385c
consensus_chr11_100673015_100673393	chr11	100673015	100673393	Zfp385c
consensus_chr11_100674104_100674553	chr11	100674104	100674553	Zfp385c
consensus_chr11_100619643_100619808	chr11	100619643	100619808	Zfp385c
consensus_chr17_22118750_22119104	chr17	22118750	22119104	Zfp947
consensus_chr14_23774570_23774832	chr14	23774570	23774832	Zmiz1
consensus_chr14_25448611_25448840	chr14	25448611	25448840	Zmiz1
consensus_chr14_19585336_19585509	chr14	19585336	19585509	Zmiz1
consensus_chr14_23778040_23778573	chr14	23778040	23778573	Zmiz1
consensus_chr14_23372059_23372254	chr14	23372059	23372254	Zmiz1
consensus_chr14_25377943_25378083	chr14	25377943	25378083	Zmiz1

8.2 List of the high confidence SMAD2 genes

I.S, induced sutained; D., Delayed; T.I, Transiently induced; R., Repressed; On.I, Baseline on induced; Off.I, Baseline off induced; Dir., Direct; Indir., Indirect.

	Gene information (0=NO, 1=YES)							
Gene	**I.S**	**D.**	**T.In**	**R.**	**On.I**	**Off.I**	**Dir.**	**Indir.**
1500012F01Rik	0	1	0	0	1	0	1	0
Abcg2	0	1	0	0	1	0	0	1
Ago1	0	1	0	0	1	0	1	0
AK013883	0	0	1	0	0	1	1	0
AK030646	0	0	1	0	0	1	1	0
AK040741	0	1	0	0	1	0	1	0
Ap3b2	0	1	0	0	1	0	1	0
Atf3	0	0	0	1	0	0	0	1
Bcar3	0	1	0	0	1	0	1	0
Calcr	0	1	0	0	1	0	0	1
Ccnd2	0	1	0	0	1	0	1	0
Cd24a	0	0	0	1	0	0	0	1
Cd59a	0	1	0	0	1	0	1	0
Cdkn1a	0	1	0	0	1	0	1	0
Cdx2	0	0	0	1	0	0	0	1
Chac1	0	0	0	1	0	0	1	0
Chst15	0	1	0	0	0	1	1	0
Crabp2	0	0	0	1	0	0	0	1
Cxcl12	0	0	0	1	0	0	1	0
Cyr61	0	0	0	1	0	0	0	1
D7Ertd143e	0	0	0	1	0	0	0	1
D7Ertd715e	1	0	0	0	1	0	0	1
Dact1	0	1	0	0	1	0	1	0
Ddit4	0	0	0	1	0	0	1	0
Dll1	0	0	0	1	0	0	0	1
Dok5	0	1	0	0	0	1	1	0
Dusp1	0	0	0	1	0	0	0	1
Dusp4	0	0	0	1	0	0	0	1
Dusp6	1	0	0	0	1	0	1	0
Efnb1	0	0	0	1	0	0	0	1
Egr1	0	0	0	1	0	0	1	0
Eomes	1	0	0	0	1	0	1	0
Epha2	1	0	0	0	1	0	1	0
Fgf15	1	0	0	0	1	0	1	0
Fgf5	0	1	0	0	0	1	0	1
Fgf8	1	0	0	0	0	1	1	0
Fgfr2	0	0	0	1	0	0	0	1
Fhl3	0	0	0	1	0	0	1	0
Foxp4	0	0	1	0	1	0	1	0
Fst	0	1	0	0	1	0	0	1
Fzd10	0	0	1	0	1	0	1	0
Gadd45g	0	0	0	1	0	0	0	1
Gm13247	1	0	0	0	0	1	1	0
Gm16386	0	1	0	0	0	1	1	0
Gpx4	0	0	0	1	0	0	0	1
Grsf1	0	1	0	0	1	0	1	0

Gsc	1	0	0	0	0	1	1	0
Gt(ROSA)26Sor	0	0	1	0	1	0	1	0
Has2	1	0	0	0	1	0	1	0
Hes1	0	0	1	0	1	0	0	1
Hlf	0	1	0	0	0	1	1	0
Hoxb1	0	0	0	1	0	0	0	1
Id1	0	0	0	1	0	0	1	0
Id2	0	0	0	1	0	0	0	1
Id3	0	0	0	1	0	0	0	1
Ier5l	0	0	0	1	0	0	0	1
Igfbp2	0	1	0	0	1	0	1	0
Il17rd	0	1	0	0	1	0	1	0
Insm1	0	0	0	1	0	0	0	1
Irf1	1	0	0	0	1	0	1	0
Irgm1	1	0	0	0	1	0	0	1
Islr2	0	0	0	1	0	0	0	1
Itch	0	1	0	0	1	0	1	0
Jun	0	0	0	1	0	0	0	1
Kazn	0	1	0	0	1	0	1	0
Kbtbd3	1	0	0	0	0	1	1	0
Kcng3	0	1	0	0	0	1	0	1
Kcnq3	0	1	0	0	0	1	0	1
Klf10	0	0	0	1	0	0	0	1
Klf6	0	0	0	1	0	0	1	0
Kras	0	1	0	0	1	0	1	0
L1td1	0	1	0	0	1	0	1	0
Lefty1	1	0	0	0	0	1	1	0
Lefty2	1	0	0	0	0	1	1	0
Lfng	0	0	0	1	0	0	0	1
Lgr4	0	1	0	0	1	0	0	1
Mest	0	1	0	0	1	0	0	1
Mllt6	0	1	0	0	1	0	1	0
Msx1	0	0	0	1	0	0	0	1
Msx2	0	0	0	1	0	0	0	1
Myc	1	0	0	0	1	0	1	0
Nanog	1	0	0	0	1	0	1	0
Neat1	0	0	0	1	0	0	0	1
Neurog1	0	0	0	1	0	0	0	1
Nodal	1	0	0	0	1	0	1	0
Nphs1os	1	0	0	0	0	1	1	0
Nrcam	0	1	0	0	1	0	1	0
Otx2	0	0	0	1	0	0	1	0
Pcdh1	0	0	0	1	0	0	0	1
Peg10	0	0	0	1	0	0	1	0
Phc1	0	0	0	1	0	0	1	0
Phlda2	0	0	0	1	0	0	0	1
Pim1	0	0	1	0	1	0	0	1
Pitx2	1	0	0	0	1	0	1	0
Pkdcc	1	0	0	0	1	0	1	0
Plekhg1	0	1	0	0	0	1	1	0
Pmepa1	1	0	0	0	0	1	1	0
Pou5f1	0	1	0	0	1	0	1	0
Ppp4r4	0	1	0	0	1	0	1	0
Pxdc1	0	1	0	0	1	0	1	0
Radil	0	1	0	0	1	0	1	0
Relt	0	1	0	0	1	0	1	0
Rgma	0	0	0	1	0	0	0	1
Rhob	0	0	0	1	0	0	0	1
Sall4	0	0	0	1	0	0	1	0
Scamp1	0	1	0	0	1	0	1	0

Scd1	0	0	0	1	0	0	1	0
Shisa3	1	0	0	0	0	1	1	0
Ski	0	1	0	0	1	0	1	0
Skil	1	0	0	0	1	0	1	0
Slc2a1	0	0	0	1	0	0	1	0
Slc2a3	0	0	0	1	0	0	1	0
Smad7	0	0	1	0	1	0	0	1
Smim14	0	1	0	0	1	0	1	0
Socs3	0	0	0	1	0	0	0	1
Sox13	0	0	0	1	0	0	0	1
Sp5	0	0	1	0	1	0	1	0
Spaca6	1	0	0	0	0	1	1	0
Spats2l	0	1	0	0	0	1	1	0
Spc25	0	1	0	0	1	0	0	1
Spry4	0	0	0	1	0	0	0	1
T	0	1	0	0	1	0	1	0
Taok3	0	1	0	0	1	0	1	0
Tbx3	0	0	0	1	0	0	0	1
Tdgf1	1	0	0	0	0	1	1	0
Tfap2a	0	0	0	1	0	0	0	1
Tle3	0	0	0	1	0	0	0	1
Tpm1	0	1	0	0	1	0	1	0
Trh	0	1	0	0	1	0	0	1
Trib3	0	0	0	1	0	0	0	1
Ubr7	0	1	0	0	1	0	1	0
Unc5b	0	1	0	0	1	0	0	1
Wnt11	1	0	0	0	0	1	1	0
Wnt3	1	0	0	0	0	1	1	0
Wnt8a	0	1	0	0	1	0	1	0
Xbp1	0	1	0	0	1	0	0	1
Zfp36l2	0	0	0	1	0	0	0	1
Zfp385c	0	1	0	0	0	1	1	0
Zfp947	1	0	0	0	0	1	1	0
Zmiz1	0	0	1	0	1	0	1	0

8.3 List of SMAD2 peaks which do not overlap with an ATAC peak in the SB-431542 sample only, or in all samples

SMAD2 Peak ID	Associated gene
consensus_chr5_117121502_117122157	Taok3
consensus_chr4_141284266_141284408	Epha2
consensus_chr12_104473558_104473733	Gsc
consensus_chr11_103808652_103808831	Wnt3
consensus_chr6_122703628_122704129	Nanog
consensus_chr4_141284518_141284750	Epha2
consensus_chr2_126521591_126521758	Chac1
consensus_chr6_4736810_4736968	Peg10
consensus_chr12_103511464_103511915	Ppp4r4
consensus_chr17_83320515_83320884	Pkdcc
consensus_chr7_81497209_81497360	Ap3b2
consensus_chr7_132320651_132320895	Chst15
consensus_chr5_128593379_128593636	Fzd10
consensus_chr2_104109533_104109799	Cd59a
consensus_chr13_34685053_34685628	Pxdc1

	SBSs negative for ATAC-seq in SB-431542

consensus_chr5_128603194_128603404	Fzd10
consensus_chr3_129217914_129218115	Pitx2
consensus_chr6_145275052_145275342	Kras
consensus_chr18_75353497_75353669	Smad7
consensus_chr17_8377276_8377422	T
consensus_chr10_59962644_59962885	Ddit4
consensus_chr6_145252001_145252289	Kras
consensus_chr10_99170969_99171036	Dusp6
consensus_chr18_38278836_38278976	Pcdh1
consensus_chr10_99170741_99170917	Dusp6
consensus_chr4_155191953_155192422	Ski
consensus_chr3_145690958_145691084	Cyr61
consensus_chr6_97063168_97063423	Trh
consensus_chr11_53826945_53827092	Irf1
consensus_chr14_23774570_23774832	Zmiz1
consensus_chr7_132360099_132360285	Chst15
consensus_chr1_191120954_191121020	Atf3
consensus_chr3_89836907_89837218	AK040741
consensus_chr3_89836176_89836827	AK040741
consensus_chr12_71376627_71376890	Dact1
consensus_chr6_92364805_92364976	Trh
consensus_chr10_59988562_59988702	Ddit4

	SBSs negative for ATAC-seq in all samples

8.4 List of SMAD2 peaks which overlap with an ATAC peak

		ATAC peak			
SMAD2 Peak ID	Associated gene	chr	peak start	peak end	% of overlap
consensus_chr12_102702435_102702575	Ubr7	chr12	102702519	102703312	40
consensus_chr3_129222436_129222934	Pitx2	chr3	129222497	129222781	57
consensus_chr14_48662913_48663201	Otx2	chr14	48662024	48663086	60
consensus_chr6_92513869_92514072	Trh	chr6	92513748	92513993	61
consensus_chr5_128601637_128601965	Fzd10	chr5	128598428	128601841	62
consensus_chr17_8376105_8376842	T	chr17	8375237	8376589	66
consensus_chr5_123140997_123141187	Setd1b	chr5	123141060	123142320	67
consensus_chr17_26511210_26511798	Dusp1	chr17	26511272	26511669	68
consensus_chr5_128593379_128593636	Fzd10	chr5	128593461	128593775	68
consensus_chr9_118455982_118456509	Eomes	chr9	118456108	118458352	76
consensus_chr1_180890289_180890577	Lefty2	chr1	180887994	180890522	81
consensus_chr7_60006072_60006972	Snurf	chr7	60006118	60006857	82
consensus_chr16_30062944_30063357	Hes1	chr16	30063016	30065625	83
consensus_chr17_17825762_17826184	Spaca6	chr17	17825835	17829587	83
consensus_chr7_98837807_98838124	Wnt11	chr7	98837860	98839226	83
consensus_chr7_60007871_60008424	Snurf	chr7	60007958	60008941	84
consensus_chr19_45743469_45743630	Fgf8	chr19	45740900	45743607	86
consensus_chr14_48707986_48708322	Otx2	chr14	48708032	48708364	86
consensus_chr10_43481135_43481437	Cd24a	chr10	43481176	43482261	86
consensus_chr11_100674104_100674553	Zfp385c	chr11	100672009	100674493	87
consensus_chr10_44553532_44553939	Cd24a	chr10	44553585	44554082	87
consensus_chr5_117121502_117122157	Taok3	chr5	117119747	117122073	87
consensus_chr9_118432551_118433292	Eomes	chr9	118432646	118433395	87
consensus_chr11_96341761_96342424	Hoxb1	chr11	96340707	96342344	88
consensus_chr6_4736810_4736968	Peg10	chr6	4736828	4740640	89
consensus_chr2_167063163_167063368	1500012F01Rik	chr2	167062411	167063345	89
consensus_chr11_103770780_103771467	Wnt3	chr11	103770853	103774649	89
consensus_chr19_5802518_5802790	Neat1	chr19	5802545	5804185	90
consensus_chr11_53786762_53787057	Irf1	chr11	53786080	53787029	91
consensus_chr2_173334398_173335678	Pmepa1	chr2	173333920	173335557	91
consensus_chr14_25718331_25718953	Il17rd	chr14	25718387	25719897	91
consensus_chr6_122703628_122704129	Nanog	chr6	122701613	122704084	91
consensus_chr16_30155736_30156494	Hes1	chr16	30155242	30156427	91
consensus_chr10_61418888_61419589	Nodal	chr10	61417148	61419533	92
consensus_chr5_67696783_67697780	Shisa3	chr5	67696858	67698206	92
consensus_chr1_180924394_180925384	Lefty1	chr1	180924463	180925518	93
consensus_chr19_45733065_45733956	Fgf8	chr19	45733113	45734147	95
consensus_chr5_128598418_128598625	Fzd10	chr5	128598428	128601841	95
consensus_chr5_142812838_142813177	Radil	chr5	142812851	142815379	96
consensus_chr1_180887934_180889501	Lefty2	chr1	180887994	180890522	96
consensus_chr18_75360801_75361429	Smad7	chr18	75359233	75361405	96
consensus_chr5_128596830_128597360	Fzd10	chr5	128596681	128597340	96
consensus_chr6_115287121_115287449	Cxcl12	chr6	115287130	115287816	97
consensus_chr5_142546212_142546943	Radil	chr5	142546225	142547334	98
consensus_chr5_65525609_65526464	Smim14	chr5	65524932	65526449	98
consensus_chr3_31087364_31087512	Skil	chr3	31087366	31090501	99
consensus_chr3_129217178_129217766	Pitx2	chr3	129215877	129217759	99
consensus_chr2_168761948_168762206	Sall4	chr2	168761951	168762846	99

consensus_chr1_57862138_57862396	Spats2l	chr1	57862028	57862432	100
consensus_chr1_57890161_57890956	Spats2l	chr1	57889876	57891203	100
consensus_chr1_59448450_59448696	Spats2l	chr1	59448434	59449394	100
consensus_chr1_72805935_72806628	Igfbp2	chr1	72804983	72806744	100
consensus_chr1_133397576_133398068	Sox13	chr1	133397330	133398721	100
consensus_chr1_133492580_133492862	Sox13	chr1	133491923	133493867	100
consensus_chr1_133493152_133493749	Sox13	chr1	133491923	133493867	100
consensus_chr1_180850818_180850959	Lefty2	chr1	180850423	180851492	100
consensus_chr1_180889952_180890153	Lefty2	chr1	180887994	180890522	100
consensus_chr1_180933629_180934184	Lefty1	chr1	180933233	180935319	100
consensus_chr1_180934401_180935253	Lefty1	chr1	180933233	180935319	100
consensus_chr1_181341049_181341199	Lefty1	chr1	181339997	181341435	100
consensus_chr1_191190073_191190359	Atf3	chr1	191189718	191190548	100
consensus_chr10_3741224_3742156	Plekhg1	chr10	3740131	3743390	100
consensus_chr10_3851942_3852301	Plekhg1	chr10	3850099	3853950	100
consensus_chr10_3941333_3941499	Plekhg1	chr10	3940953	3942604	100
consensus_chr10_43551585_43551877	Cd24a	chr10	43550943	43551941	100
consensus_chr10_44392015_44392352	Cd24a	chr10	44391667	44392577	100
consensus_chr10_59951664_59951902	Ddit4	chr10	59950680	59953988	100
consensus_chr10_59957919_59958288	Ddit4	chr10	59957262	59959762	100
consensus_chr10_60793113_60793280	Unc5b	chr10	60792573	60794861	100
consensus_chr10_60814586_60815057	Unc5b	chr10	60813755	60816281	100
consensus_chr10_61405952_61406482	Nodal	chr10	61405870	61406572	100
consensus_chr10_61415682_61416344	Nodal	chr10	61414865	61416952	100
consensus_chr10_61418275_61418474	Nodal	chr10	61417148	61419533	100
consensus_chr10_61500857_61500996	Nodal	chr10	61500755	61501181	100
consensus_chr10_79988724_79988849	Gpx4	chr10	79988410	79989345	100
consensus_chr10_80050827_80051622	Gpx4	chr10	80049771	80052372	100
consensus_chr10_80053637_80053790	Gpx4	chr10	80053244	80054031	100
consensus_chr10_99261290_99261430	Dusp6	chr10	99261100	99264726	100
consensus_chr11_3292790_3292951	Pik3ip1	chr11	3291755	3293495	100
consensus_chr11_3337600_3338123	Pik3ip1	chr11	3336799	3338253	100
consensus_chr11_5562956_5563289	Xbp1	chr11	5562112	5564382	100
consensus_chr11_48800410_48800567	Irgm1	chr11	48799535	48800809	100
consensus_chr11_48818441_48818699	Irgm1	chr11	48815688	48819314	100
consensus_chr11_48826804_48827096	Irgm1	chr11	48825950	48827114	100
consensus_chr11_48833788_48834058	Irgm1	chr11	48833712	48835309	100
consensus_chr11_48834252_48834529	Irgm1	chr11	48833712	48835309	100
consensus_chr11_48853597_48853755	Irgm1	chr11	48851435	48859838	100
consensus_chr11_48856851_48857129	Irgm1	chr11	48851435	48859838	100
consensus_chr11_48871235_48871651	Irgm1	chr11	48870851	48872487	100
consensus_chr11_53759684_53760389	Irf1	chr11	53757797	53760748	100
consensus_chr11_53781310_53781506	Irf1	chr11	53780326	53782206	100
consensus_chr11_53799980_53800464	Irf1	chr11	53799751	53800633	100
consensus_chr11_55438305_55438441	Irf1	chr11	55438002	55438542	100
consensus_chr11_55461054_55461179	Irf1	chr11	55460568	55461750	100
consensus_chr11_90356160_90356329	Hlf	chr11	90354435	90357186	100
consensus_chr11_90371054_90371268	Hlf	chr11	90370755	90371432	100
consensus_chr11_90391688_90391958	Hlf	chr11	90389483	90392558	100
consensus_chr11_96368741_96368886	Hoxb1	chr11	96368694	96369348	100
consensus_chr11_97661653_97662273	Mllt6	chr11	97659993	97665349	100
consensus_chr11_98925815_98926131	Rara	chr11	98924372	98927892	100
consensus_chr11_98960001_98960202	Rara	chr11	98958831	98961193	100
consensus_chr11_100527735_100527924	Zfp385c	chr11	100526448	100528699	100
consensus_chr11_100619643_100619808	Zfp385c	chr11	100619037	100620360	100
consensus_chr11_100673015_100673393	Zfp385c	chr11	100672009	100674493	100
consensus_chr11_100713769_100714028	Zfp385c	chr11	100711737	100714917	100
consensus_chr11_103772298_103772452	Wnt3	chr11	103770853	103774649	100
consensus_chr11_103798585_103798879	Wnt3	chr11	103797817	103798977	100
consensus_chr11_103808652_103808831	Wnt3	chr11	103808111	103811114	100

consensus_chr11_103809172_103809552	Wnt3	chr11	103808111	103811114	100
consensus_chr11_103810386_103810439	Wnt3	chr11	103808111	103811114	100
consensus_chr11_103840947_103841240	Wnt3	chr11	103840561	103842147	100
consensus_chr11_117879734_117879991	Socs3	chr11	117878464	117880020	100
consensus_chr11_117941914_117942039	Socs3	chr11	117941456	117942264	100
consensus_chr11_117972755_117973060	Socs3	chr11	117972510	117973116	100
consensus_chr12_8402085_8402511	Rhob	chr12	8401676	8402786	100
consensus_chr12_25098494_25098699	Id2	chr12	25098482	25101755	100
consensus_chr12_44327670_44327856	Nrcam	chr12	44327386	44330755	100
consensus_chr12_44393882_44394045	Nrcam	chr12	44393621	44394819	100
consensus_chr12_44403632_44404236	Nrcam	chr12	44402797	44404782	100
consensus_chr12_44535560_44536334	Nrcam	chr12	44535105	44536663	100
consensus_chr12_71296373_71297167	Dact1	chr12	71296185	71297206	100
consensus_chr12_71301571_71302020	Dact1	chr12	71301442	71302323	100
consensus_chr12_71369195_71369340	Dact1	chr12	71368754	71370157	100
consensus_chr12_71369922_71370116	Dact1	chr12	71368754	71370157	100
consensus_chr12_102758660_102758816	Ubr7	chr12	102756390	102759383	100
consensus_chr12_102758889_102759365	Ubr7	chr12	102756390	102759383	100
consensus_chr12_103511464_103511915	Ppp4r4	chr12	103511319	103512029	100
consensus_chr12_103564312_103565109	Ppp4r4	chr12	103564108	103565426	100
consensus_chr12_103599158_103599646	Ppp4r4	chr12	103598711	103601060	100
consensus_chr12_103600388_103600857	Ppp4r4	chr12	103598711	103601060	100
consensus_chr12_104381222_104381480	Gsc	chr12	104380875	104381557	100
consensus_chr12_104466887_104467386	Gsc	chr12	104466525	104467410	100
consensus_chr12_104470812_104470973	Gsc	chr12	104470594	104474907	100
consensus_chr12_104473558_104473733	Gsc	chr12	104470594	104474907	100
consensus_chr12_111664282_111664372	Gsc	chr12	111663772	111665393	100
consensus_chr12_111672141_111672473	Gsc	chr12	111670257	111672880	100
consensus_chr13_5892395_5892635	Klf6	chr13	5891719	5893070	100
consensus_chr13_34685053_34685628	Pxdc1	chr13	34684935	34686066	100
consensus_chr13_34741702_34742336	Pxdc1	chr13	34741307	34742532	100
consensus_chr13_40744792_40744854	Tfap2a	chr13	40743396	40745157	100
consensus_chr13_51845953_51846130	Gadd45g	chr13	51845539	51849372	100
consensus_chr13_51846534_51846659	Gadd45g	chr13	51845539	51849372	100
consensus_chr13_51853524_51853585	Gadd45g	chr13	51853225	51854077	100
consensus_chr13_53455410_53455576	Msx2	chr13	53454802	53456422	100
consensus_chr13_56222214_56222396	Neurog1	chr13	56220572	56222826	100
consensus_chr13_94242174_94242361	Scamp1	chr13	94241950	94244450	100
consensus_chr13_94321144_94321833	Scamp1	chr13	94320807	94322454	100
consensus_chr13_114406548_114406714	Fst	chr13	114406001	114407668	100
consensus_chr13_114406841_114407585	Fst	chr13	114406001	114407668	100
consensus_chr14_19585336_19585509	Zmiz1	chr14	19584020	19585857	100
consensus_chr14_23372059_23372254	Zmiz1	chr14	23371961	23372555	100
consensus_chr14_23778040_23778573	Zmiz1	chr14	23777648	23778851	100
consensus_chr14_25377943_25378083	Zmiz1	chr14	25377222	25379377	100
consensus_chr14_25448611_25448840	Zmiz1	chr14	25448060	25449180	100
consensus_chr14_25607054_25607197	Il17rd	chr14	25606294	25608461	100
consensus_chr14_27039338_27039517	Il17rd	chr14	27038753	27040127	100
consensus_chr14_27044304_27044591	Il17rd	chr14	27044208	27045475	100
consensus_chr14_27044889_27045392	Il17rd	chr14	27044208	27045475	100
consensus_chr14_27066406_27066795	Il17rd	chr14	27065914	27066921	100
consensus_chr14_27096185_27096406	Il17rd	chr14	27096183	27097188	100
consensus_chr14_48653661_48653928	Otx2	chr14	48653649	48656176	100
consensus_chr14_48654960_48655406	Otx2	chr14	48653649	48656176	100
consensus_chr14_48700916_48701209	Otx2	chr14	48700495	48702332	100
consensus_chr14_48709413_48709626	Otx2	chr14	48709370	48711080	100
consensus_chr14_48753109_48753424	Otx2	chr14	48751040	48754522	100
consensus_chr14_48754197_48754367	Otx2	chr14	48751040	48754522	100
consensus_chr14_49089813_49089999	Otx2	chr14	49089475	49090429	100
consensus_chr14_49155640_49155812	Otx2	chr14	49155478	49155964	100

consensus_chr14_49275422_49275653	Otx2	chr14	49274380	49276437	100
consensus_chr15_38300483_38300665	Klf10	chr15	38298839	38302587	100
consensus_chr15_40955266_40955765	Klf10	chr15	40953701	40955958	100
consensus_chr15_41162279_41162512	Klf10	chr15	41162098	41164383	100
consensus_chr15_54702219_54702825	Has2	chr15	54701955	54703424	100
consensus_chr15_56626464_56626771	Has2	chr15	56626316	56628010	100
consensus_chr15_56627608_56627951	Has2	chr15	56626316	56628010	100
consensus_chr15_56708427_56708607	Has2	chr15	56707695	56708870	100
consensus_chr15_56729756_56730065	Has2	chr15	56728585	56730436	100
consensus_chr15_56742758_56743078	Has2	chr15	56741852	56745176	100
consensus_chr15_62086921_62087198	Kcnq3	chr15	62086614	62088088	100
consensus_chr15_65926152_65926942	Myc	chr15	65926016	65927248	100
consensus_chr16_8769190_8769359	AK013883	chr16	8768721	8771400	100
consensus_chr16_8778570_8779080	AK013883	chr16	8778530	8779246	100
consensus_chr16_8788044_8788566	AK013883	chr16	8787538	8789184	100
consensus_chr16_8829488_8829637	AK013883	chr16	8829075	8831183	100
consensus_chr16_30103122_30103289	Hes1	chr16	30101918	30104856	100
consensus_chr16_30129541_30129713	Hes1	chr16	30129277	30130500	100
consensus_chr16_30160400_30160832	Hes1	chr16	30160362	30161131	100
consensus_chr16_33605791_33606065	Hes1	chr16	33605724	33607534	100
consensus_chr16_33737816_33738120	Hes1	chr16	33736837	33738143	100
consensus_chr16_33844391_33844583	Hes1	chr16	33843688	33845850	100
consensus_chr16_33845586_33845726	Hes1	chr16	33843688	33845850	100
consensus_chr17_8375801_8376001	T	chr17	8375237	8376589	100
consensus_chr17_8389190_8389339	T	chr17	8388337	8390106	100
consensus_chr17_8408979_8409262	T	chr17	8408429	8409598	100
consensus_chr17_15380676_15380801	Dll1	chr17	15380446	15382510	100
consensus_chr17_15380813_15380866	Dll1	chr17	15380446	15382510	100
consensus_chr17_17826709_17826997	Spaca6	chr17	17825835	17829587	100
consensus_chr17_17827285_17827425	Spaca6	chr17	17825835	17829587	100
consensus_chr17_17827567_17828588	Spaca6	chr17	17825835	17829587	100
consensus_chr17_22118750_22119104	Zfp947	chr17	22118228	22119603	100
consensus_chr17_22745725_22746266	Gm16386	chr17	22745473	22746599	100
consensus_chr17_29007432_29007640	Cdkn1a	chr17	29007240	29008723	100
consensus_chr17_29008194_29008596	Cdkn1a	chr17	29007240	29008723	100
consensus_chr17_29032640_29032874	Cdkn1a	chr17	29031781	29033022	100
consensus_chr17_29093566_29093896	Cdkn1a	chr17	29093344	29095973	100
consensus_chr17_29451085_29451569	Pim1	chr17	29450754	29451942	100
consensus_chr17_35485810_35485956	Pou5f1	chr17	35483966	35487054	100
consensus_chr17_35503853_35504299	Pou5f1	chr17	35503828	35506249	100
consensus_chr17_35504528_35506015	Pou5f1	chr17	35503828	35506249	100
consensus_chr17_35548650_35548879	Pou5f1	chr17	35547554	35549968	100
consensus_chr17_47887305_47887876	Foxp4	chr17	47887256	47889086	100
consensus_chr17_47953108_47953293	Foxp4	chr17	47952272	47953578	100
consensus_chr17_70737468_70737608	Pkdcc	chr17	70737082	70739641	100
consensus_chr17_70755079_70755371	Pkdcc	chr17	70754951	70756461	100
consensus_chr17_70772484_70772902	Pkdcc	chr17	70771856	70773179	100
consensus_chr17_70900626_70901022	Pkdcc	chr17	70900552	70901130	100
consensus_chr17_83166332_83167152	Pkdcc	chr17	83164032	83167933	100
consensus_chr17_83214236_83214528	Pkdcc	chr17	83213686	83216196	100
consensus_chr17_83320515_83320884	Pkdcc	chr17	83320431	83321105	100
consensus_chr17_83526238_83526490	Kcng3	chr17	83525681	83529500	100
consensus_chr17_83547624_83548022	Kcng3	chr17	83546129	83549785	100
consensus_chr17_83633016_83633583	Kcng3	chr17	83630735	83633623	100
consensus_chr17_83668253_83668509	Kcng3	chr17	83666608	83668699	100
consensus_chr17_83694920_83695176	Kcng3	chr17	83694322	83695258	100
consensus_chr17_84149377_84149961	Zfp36l2	chr17	84147313	84150553	100
consensus_chr17_84152415_84152592	Zfp36l2	chr17	84152235	84152874	100
consensus_chr17_84177535_84177714	Zfp36l2	chr17	84177426	84177980	100
consensus_chr18_34542264_34542435	Wnt8a	chr18	34541156	34542799	100

consensus_chr18_34549061_34549615	Wnt8a	chr18	34548872	34552584	100
consensus_chr18_34551614_34552178	Wnt8a	chr18	34548872	34552584	100
consensus_chr18_34555243_34555601	Wnt8a	chr18	34554793	34555744	100
consensus_chr18_34850208_34850354	Egr1	chr18	34849269	34850676	100
consensus_chr18_38210652_38210850	Pcdh1	chr18	38209610	38213731	100
consensus_chr18_38573819_38574176	Spry4	chr18	38572117	38576547	100
consensus_chr18_38635944_38636200	Spry4	chr18	38635854	38637692	100
consensus_chr18_75366741_75367245	Smad7	chr18	75365921	75368411	100
consensus_chr19_44474304_44474435	Scd1	chr19	44474122	44474595	100
consensus_chr19_45740924_45741066	Fgf8	chr19	45740900	45743607	100
consensus_chr19_45745737_45746430	Fgf8	chr19	45745363	45750389	100
consensus_chr19_45783535_45783679	Fgf8	chr19	45782545	45784654	100
consensus_chr19_45790984_45791161	Fgf8	chr19	45790615	45791209	100
consensus_chr2_30474056_30474199	Ier5l	chr2	30473155	30474581	100
consensus_chr2_69150788_69151293	Spc25	chr2	69150553	69151468	100
consensus_chr2_69222031_69222562	Spc25	chr2	69221863	69223136	100
consensus_chr2_70471866_70472314	Sp5	chr2	70471283	70477901	100
consensus_chr2_70474647_70475140	Sp5	chr2	70471283	70477901	100
consensus_chr2_70563254_70563440	Sp5	chr2	70561022	70563846	100
consensus_chr2_104109533_104109799	Cd59a	chr2	104109492	104110041	100
consensus_chr2_104122629_104122782	Cd59a	chr2	104122400	104123708	100
consensus_chr2_104188245_104188452	Cd59a	chr2	104187774	104188705	100
consensus_chr2_109917143_109917715	Lgr4	chr2	109916698	109918893	100
consensus_chr2_119288567_119288727	Chac1	chr2	119287479	119289247	100
consensus_chr2_119304534_119304780	Chac1	chr2	119304426	119305283	100
consensus_chr2_119351226_119351442	Chac1	chr2	119350520	119351620	100
consensus_chr2_126521591_126521758	Chac1	chr2	126519973	126521947	100
consensus_chr2_126523532_126523676	Chac1	chr2	126523207	126523974	100
consensus_chr2_131152139_131152505	Chac1	chr2	131151546	131152948	100
consensus_chr2_146302378_146302639	Insm1	chr2	146301662	146302929	100
consensus_chr2_152396579_152396712	Trib3	chr2	152395922	152399535	100
consensus_chr2_152735131_152735359	Id1	chr2	152734849	152740023	100
consensus_chr2_152736335_152736401	Id1	chr2	152734849	152740023	100
consensus_chr2_155053130_155053272	Itch	chr2	155053114	155053323	100
consensus_chr2_155074251_155074431	Itch	chr2	155073839	155075462	100
consensus_chr2_155115402_155116063	Itch	chr2	155115035	155116126	100
consensus_chr2_155133455_155133700	Itch	chr2	155133057	155134465	100
consensus_chr2_167058802_167059178	1500012F01Rik	chr2	167058106	167060006	100
consensus_chr2_168777065_168777396	Sall4	chr2	168776250	168777795	100
consensus_chr2_168780159_168780342	Sall4	chr2	168779890	168781024	100
consensus_chr2_168782911_168783312	Sall4	chr2	168782765	168783614	100
consensus_chr2_168791141_168791428	Sall4	chr2	168790762	168792239	100
consensus_chr2_168838785_168839141	Sall4	chr2	168837189	168843142	100
consensus_chr2_168840268_168840521	Sall4	chr2	168837189	168843142	100
consensus_chr2_168840980_168841183	Sall4	chr2	168837189	168843142	100
consensus_chr2_170657559_170658226	Dok5	chr2	170657440	170659601	100
consensus_chr2_170726216_170726396	Dok5	chr2	170726200	170726550	100
consensus_chr2_170772037_170772607	Dok5	chr2	170771407	170773225	100
consensus_chr2_170887653_170887870	Dok5	chr2	170887081	170888701	100
consensus_chr2_173238334_173239057	Pmepa1	chr2	173237991	173240321	100
consensus_chr2_173257969_173258380	Pmepa1	chr2	173257634	173258907	100
consensus_chr2_173339932_173340353	Pmepa1	chr2	173339148	173340992	100
consensus_chr2_173672536_173672892	Pmepa1	chr2	173672159	173672987	100
consensus_chr3_31089484_31089609	Skil	chr3	31087366	31090501	100
consensus_chr3_86566305_86566532	Crabp2	chr3	86565934	86566991	100
consensus_chr3_87891547_87891945	Crabp2	chr3	87891392	87892001	100
consensus_chr3_87970999_87971247	Crabp2	chr3	87970207	87971612	100
consensus_chr3_87974317_87974567	Crabp2	chr3	87972894	87975806	100
consensus_chr3_89800149_89800499	AK040741	chr3	89799229	89800600	100
consensus_chr3_122414149_122414544	Bcar3	chr3	122412766	122414592	100

consensus_chr3_122434049_122434426	Bcar3	chr3	122432400	122434568	100
consensus_chr3_129214714_129214896	Pitx2	chr3	129213081	129215112	100
consensus_chr3_129216817_129217096	Pitx2	chr3	129215877	129217759	100
consensus_chr3_145649846_145649978	Cyr61	chr3	145649274	145653887	100
consensus_chr4_95052028_95052213	Jun	chr4	95050047	95054042	100
consensus_chr4_95052482_95052624	Jun	chr4	95050047	95054042	100
consensus_chr4_95134823_95135014	Jun	chr4	95133713	95136474	100
consensus_chr4_98726663_98726945	L1td1	chr4	98725404	98728926	100
consensus_chr4_98727954_98728510	L1td1	chr4	98725404	98728926	100
consensus_chr4_119026459_119026723	Slc2a1	chr4	119025307	119027088	100
consensus_chr4_119108741_119108988	Slc2a1	chr4	119108251	119109392	100
consensus_chr4_124708779_124709208	Fhl3	chr4	124708476	124711141	100
consensus_chr4_126465731_126466196	Ago1	chr4	126464871	126466321	100
consensus_chr4_135963498_135963678	Id3	chr4	135963103	135964702	100
consensus_chr4_136031505_136031958	Id3	chr4	136030923	136032124	100
consensus_chr4_136053302_136053447	Id3	chr4	136052900	136054761	100
consensus_chr4_136140369_136140581	Id3	chr4	136140261	136140998	100
consensus_chr4_138361509_138362033	Id3	chr4	138360515	138362411	100
consensus_chr4_138396228_138396353	Id3	chr4	138394321	138397087	100
consensus_chr4_138418538_138418779	Id3	chr4	138418185	138418793	100
consensus_chr4_138444922_138445231	Id3	chr4	138444679	138445445	100
consensus_chr4_138453360_138453557	Id3	chr4	138452979	138456607	100
consensus_chr4_140126698_140126946	Epha2	chr4	140126449	140127014	100
consensus_chr4_140131460_140131654	Epha2	chr4	140130986	140132487	100
consensus_chr4_141284266_141284408	Epha2	chr4	141282355	141285605	100
consensus_chr4_141284518_141284750	Epha2	chr4	141282355	141285605	100
consensus_chr4_141291432_141292027	Epha2	chr4	141291358	141292214	100
consensus_chr4_141338992_141339132	Epha2	chr4	141337599	141339620	100
consensus_chr4_142140118_142140268	Kazn	chr4	142137147	142140462	100
consensus_chr4_142166133_142166914	Kazn	chr4	142164994	142167035	100
consensus_chr4_142197117_142197286	Kazn	chr4	142196444	142197502	100
consensus_chr4_146498084_146498487	Gm13247	chr4	146497103	146499323	100
consensus_chr4_155189613_155190201	Ski	chr4	155188466	155190543	100
consensus_chr4_155249806_155250010	Ski	chr4	155249008	155250758	100
consensus_chr4_155290432_155291069	Ski	chr4	155289990	155291081	100
consensus_chr5_37768802_37769120	Msx1	chr5	37768620	37770174	100
consensus_chr5_37826777_37827028	Msx1	chr5	37823179	37827573	100
consensus_chr5_37832855_37832998	Msx1	chr5	37832294	37833194	100
consensus_chr5_65391226_65391539	Smim14	chr5	65390857	65392108	100
consensus_chr5_67632834_67634059	Shisa3	chr5	67632823	67634698	100
consensus_chr5_88670674_88670933	Grsf1	chr5	88670665	88670937	100
consensus_chr5_88679333_88679917	Grsf1	chr5	88679240	88679958	100
consensus_chr5_88684336_88684662	Grsf1	chr5	88684149	88685990	100
consensus_chr5_98167646_98167808	Fgf5	chr5	98166046	98173370	100
consensus_chr5_98255741_98256081	Fgf5	chr5	98252668	98256396	100
consensus_chr5_98283639_98283918	Fgf5	chr5	98282761	98285930	100
consensus_chr5_98293510_98293685	Fgf5	chr5	98292479	98294020	100
consensus_chr5_98310053_98310554	Fgf5	chr5	98309386	98311115	100
consensus_chr5_98312032_98312336	Fgf5	chr5	98312015	98312593	100
consensus_chr5_117135600_117135895	Taok3	chr5	117134664	117136208	100
consensus_chr5_119613558_119613838	Tbx3	chr5	119613345	119614308	100
consensus_chr5_119706579_119706794	Tbx3	chr5	119706391	119707092	100
consensus_chr5_119706806_119706869	Tbx3	chr5	119706391	119707092	100
consensus_chr5_119709831_119709975	Tbx3	chr5	119709752	119711089	100
consensus_chr5_123136547_123137136	Setd1b	chr5	123135534	123139649	100
consensus_chr5_123137711_123137959	Setd1b	chr5	123135534	123139649	100
consensus_chr5_123138268_123138554	Setd1b	chr5	123135534	123139649	100
consensus_chr5_123245570_123245742	Setd1b	chr5	123245319	123246053	100
consensus_chr5_128579190_128579385	Fzd10	chr5	128579072	128579770	100
consensus_chr5_128600543_128600829	Fzd10	chr5	128598428	128601841	100

consensus_chr5_140583969_140584286	Lfng	chr5	140582972	140584613	100
consensus_chr5_140611995_140612157	Lfng	chr5	140610714	140612967	100
consensus_chr5_142800798_142800938	Radil	chr5	142800136	142801229	100
consensus_chr5_142821774_142821827	Radil	chr5	142821384	142822094	100
consensus_chr5_142881399_142881695	Radil	chr5	142880992	142882070	100
consensus_chr5_142906533_142906819	Radil	chr5	142905629	142907127	100
consensus_chr5_147304967_147305067	Cdx2	chr5	147304118	147307572	100
consensus_chr6_3623904_3624150	Calcr	chr6	3623322	3625185	100
consensus_chr6_3677836_3678529	Calcr	chr6	3677571	3678854	100
consensus_chr6_4739948_4740139	Peg10	chr6	4736828	4740640	100
consensus_chr6_4747369_4747494	Peg10	chr6	4746688	4749541	100
consensus_chr6_4855296_4855436	Peg10	chr6	4855205	4855691	100
consensus_chr6_30711408_30711749	Mest	chr6	30710979	30712531	100
consensus_chr6_58595576_58595877	Abcg2	chr6	58594345	58597093	100
consensus_chr6_92246269_92246732	Trh	chr6	92245906	92248473	100
consensus_chr6_92247122_92248225	Trh	chr6	92245906	92248473	100
consensus_chr6_92469737_92470184	Trh	chr6	92468353	92470294	100
consensus_chr6_97101730_97101881	Trh	chr6	97101287	97102919	100
consensus_chr6_97102282_97102637	Trh	chr6	97101287	97102919	100
consensus_chr6_108644376_108644559	Gt(ROSA)26Sor	chr6	108642104	108644746	100
consensus_chr6_108696365_108696528	Gt(ROSA)26Sor	chr6	108695302	108696669	100
consensus_chr6_113077039_113077254	Gt(ROSA)26Sor	chr6	113075907	113078241	100
consensus_chr6_115252107_115252377	Cxcl12	chr6	115251322	115253114	100
consensus_chr6_117151734_117152782	Cxcl12	chr6	117150743	117152797	100
consensus_chr6_117180164_117180610	Cxcl12	chr6	117180017	117181031	100
consensus_chr6_122340368_122340910	Phc1	chr6	122338680	122343286	100
consensus_chr6_122342170_122342533	Phc1	chr6	122338680	122343286	100
consensus_chr6_122366465_122366649	Phc1	chr6	122365895	122368078	100
consensus_chr6_122608433_122608717	Nanog	chr6	122607822	122610232	100
consensus_chr6_122609943_122610121	Nanog	chr6	122607822	122610232	100
consensus_chr6_122624641_122624865	Nanog	chr6	122624292	122624990	100
consensus_chr6_122702403_122702936	Nanog	chr6	122701613	122704084	100
consensus_chr6_122707200_122707662	Nanog	chr6	122706980	122709094	100
consensus_chr6_122715788_122716411	Nanog	chr6	122715027	122717032	100
consensus_chr6_122742373_122742537	Slc2a3	chr6	122741967	122745207	100
consensus_chr6_122765209_122765420	Slc2a3	chr6	122764395	122766409	100
consensus_chr6_122766285_122766338	Slc2a3	chr6	122764395	122766409	100
consensus_chr6_127106877_127107026	Ccnd2	chr6	127106776	127107096	100
consensus_chr6_144251210_144251401	Kras	chr6	144250568	144253027	100
consensus_chr6_144273631_144273854	Kras	chr6	144272929	144273970	100
consensus_chr6_144361308_144361462	Kras	chr6	144360822	144362007	100
consensus_chr6_144373492_144373661	Kras	chr6	144372954	144373762	100
consensus_chr6_144454403_144454693	Kras	chr6	144453102	144455119	100
consensus_chr6_145261829_145262239	Kras	chr6	145261501	145262394	100
consensus_chr6_145275740_145276846	Kras	chr6	145275595	145277427	100
consensus_chr7_3206524_3206886	D7Ertd143e	chr7	3206269	3207272	100
consensus_chr7_3213628_3213837	D7Ertd143e	chr7	3211686	3214281	100
consensus_chr7_3217652_3217973	D7Ertd143e	chr7	3217447	3217978	100
consensus_chr7_3314282_3314409	D7Ertd143e	chr7	3314025	3314538	100
consensus_chr7_30463526_30464004	Nphs1os	chr7	30463332	30464333	100
consensus_chr7_30464066_30464241	Nphs1os	chr7	30463332	30464333	100
consensus_chr7_62420131_62420290	D7Ertd715e	chr7	62419180	62420975	100
consensus_chr7_73344144_73344284	Rgma	chr7	73343916	73344738	100
consensus_chr7_73399519_73399721	Rgma	chr7	73398095	73400761	100
consensus_chr7_81480831_81481312	Ap3b2	chr7	81479797	81481530	100
consensus_chr7_81497209_81497360	Ap3b2	chr7	81496860	81497553	100
consensus_chr7_82056669_82056859	Rgma	chr7	82056576	82057552	100
consensus_chr7_82063489_82063814	Rgma	chr7	82062770	82065596	100
consensus_chr7_90457101_90457275	Wnt11	chr7	90456200	90457902	100
consensus_chr7_98815460_98815615	Wnt11	chr7	98813946	98817445	100

consensus_chr7_98840585_98841145	Wnt11	chr7	98840577	98842221	100
consensus_chr7_100861981_100862442	Relt	chr7	100861960	100863725	100
consensus_chr7_130252207_130252429	Fgfr2	chr7	130251722	130252953	100
consensus_chr7_130264662_130264897	Fgfr2	chr7	130263623	130267312	100
consensus_chr7_130327950_130328203	Fgfr2	chr7	130327437	130328844	100
consensus_chr7_132247588_132247876	Chst15	chr7	132247149	132248023	100
consensus_chr7_132317113_132317585	Chst15	chr7	132315323	132317716	100
consensus_chr7_132320651_132320895	Chst15	chr7	132320547	132320937	100
consensus_chr7_132403936_132404243	Chst15	chr7	132403424	132404503	100
consensus_chr7_143467822_143467974	Phlda2	chr7	143467514	143468865	100
consensus_chr7_144872813_144872953	Fgf15	chr7	144872147	144873516	100
consensus_chr7_144888718_144888877	Fgf15	chr7	144887523	144889451	100
consensus_chr7_144898506_144898931	Fgf15	chr7	144898431	144899150	100
consensus_chr8_4436040_4436209	AK030646	chr8	4435732	4437266	100
consensus_chr8_4436476_4436884	AK030646	chr8	4435732	4437266	100
consensus_chr8_4582747_4582992	AK030646	chr8	4582001	4583143	100
consensus_chr8_34732699_34733049	Dusp4	chr8	34732495	34733704	100
consensus_chr8_34805012_34805408	Dusp4	chr8	34804723	34809969	100
consensus_chr9_4226971_4227211	Kbtbd3	chr9	4226480	4227505	100
consensus_chr9_4283761_4284275	Kbtbd3	chr9	4283348	4284535	100
consensus_chr9_58279246_58279681	Islr2	chr9	58278880	58281939	100
consensus_chr9_58295776_58296231	Islr2	chr9	58295742	58296522	100
consensus_chr9_61293122_61293899	Tle3	chr9	61293078	61294065	100
consensus_chr9_61324251_61324332	Tle3	chr9	61324037	61325110	100
consensus_chr9_61339317_61339544	Tle3	chr9	61339114	61340107	100
consensus_chr9_61361289_61361481	Tle3	chr9	61360627	61361760	100
consensus_chr9_61367314_61367492	Tle3	chr9	61367276	61372593	100
consensus_chr9_61367852_61368066	Tle3	chr9	61367276	61372593	100
consensus_chr9_61368182_61368412	Tle3	chr9	61367276	61372593	100
consensus_chr9_61372345_61372498	Tle3	chr9	61367276	61372593	100
consensus_chr9_61385167_61385319	Tle3	chr9	61383983	61386657	100
consensus_chr9_61445921_61446061	Tle3	chr9	61445653	61446747	100
consensus_chr9_67105826_67106124	Tpm1	chr9	67104729	67107126	100
consensus_chr9_110946531_110946721	Tdgf1	chr9	110946041	110948893	100
consensus_chr9_110947137_110948383	Tdgf1	chr9	110946041	110948893	100
consensus_chr9_110949302_110949525	Tdgf1	chr9	110949271	110949585	100
consensus_chr9_118384752_118384964	Eomes	chr9	118384635	118385117	100
consensus_chr9_118389946_118390214	Eomes	chr9	118389295	118390784	100
consensus_chr9_118468200_118468630	Eomes	chr9	118468086	118469648	100
consensus_chr9_118486783_118487207	Eomes	chr9	118486360	118490086	100
consensus_chr9_118487571_118487804	Eomes	chr9	118486360	118490086	100
consensus_chrX_99045496_99045800	Efnb1	chrX	99045412	99045903	100
consensus_chrX_99152773_99152929	Efnb1	chrX	99152145	99153740	100
consensus_chrX_106080918_106081259	Efnb1	chrX	106080543	106081706	100
consensus_chrX_139317424_139317594	Efnb1	chrX	139317333	139318099	100
consensus_chrX_139345212_139345622	Efnb1	chrX	139344862	139346096	100
consensus_chrX_139436747_139436935	Efnb1	chrX	139436324	139437066	100
consensus_chrX_145524466_145524856	Efnb1	chrX	145524379	145525062	100

8.5 List of SMAD2 peaks with differential ATAC signal compared to the SB-4431542 sample

Act. 1hr, Activin 1hr; Act. 8 hr, Activin 8 hr; Untr, Untreated.

		SMAD2 ChIP-seq (0=NO, 1=YES)		
SMAD2 Peak ID	**Associated gene**	**Act. 1hr**	**Act. 8 hr**	**Untr.**
consensus_chr1_180924394_180925384	Lefty1	1	1	1
consensus_chr2_173334398_173335678	Pmepa1	1	1	1
consensus_chr2_173257969_173258380	Pmepa1	1	1	1
consensus_chr3_129216817_129217096	Pitx2	0	1	0
consensus_chr3_129217178_129217766	Pitx2	1	1	1
consensus_chr3_129217914_129218115	Pitx2	0	1	0
consensus_chr3_129222436_129222934	Pitx2	1	1	1
consensus_chr15_54702219_54702825	Has2	0	1	1
consensus_chr17_83166332_83167152	Pkdcc	1	1	1
consensus_chr2_173672536_173672892	Pmepa1	0	1	0
consensus_chr12_104466887_104467386	Gsc	0	1	1
consensus_chr9_110949302_110949525	Tdgf1	0	1	0
consensus_chr17_83320515_83320884	Pkdcc	0	1	0
consensus_chr9_118432551_118433292	Eomes	0	1	0
consensus_chr19_45733065_45733956	Fgf8	1	1	1
consensus_chr11_53786762_53787057	Irf1	0	1	0
consensus_chr15_56729756_56730065	Has2	0	1	0
consensus_chr2_173672536_173672892	Pmepa1	0	1	0
consensus_chr7_98840585_98841145	Wnt11	0	1	1
consensus_chr7_30463526_30464004	Nphs1os	1	1	1
consensus_chr7_30464066_30464241	Nphs1os	0	1	0
consensus_chr7_144888718_144888877	Fgf15	0	1	0
consensus_chr9_118455982_118456509	Eomes	0	1	0
consensus_chr11_53781310_53781506	Irf1	0	1	0
consensus_chr4_141284266_141284408	Epha2	0	1	0
consensus_chr4_141284518_141284750	Epha2	0	1	0
consensus_chr6_122703628_122704129	Nanog	1	1	0
consensus_chr1_180933629_180934184	Lefty1	1	1	1
consensus_chr1_180934401_180935253	Lefty1	1	1	1
consensus_chr2_173238334_173239057	Pmepa1	1	1	1
consensus_chr9_4226971_4227211	Kbtbd3	0	1	0
consensus_chr15_65926152_65926942	Myc	1	1	1
consensus_chr2_173339932_173340353	Pmepa1	1	1	1
consensus_chr11_103808652_103808831	Wnt3	0	1	0
consensus_chr11_103809172_103809552	Wnt3	1	1	1
consensus_chr12_104473558_104473733	Gsc	0	1	0
consensus_chr9_110947137_110948383	Tdgf1	1	1	1
consensus_chr10_61405952_61406482	Nodal	1	1	1
consensus_chr9_118468200_118468630	Eomes	0	1	0
consensus_chr11_3337600_3338123	Pik3ip1	1	1	1
consensus_chr10_61418888_61419589	Nodal	1	1	1
consensus_chr12_111664282_111664372	Gsc	1	0	0
consensus_chr5_67696783_67697780	Shisa3	1	1	1
consensus_chr10_61418275_61418474	Nodal	0	1	0
consensus_chr3_31089484_31089609	Skil	0	0	1
consensus_chr15_56708427_56708607	Has2	0	1	0
consensus_chr11_53759684_53760389	Irf1	0	1	1

consensus_chr5_67696783_67697780	Shisa3	1	1	1
consensus_chr6_122715788_122716411	Nanog	1	1	1
consensus_chr10_61415682_61416344	Nodal	0	1	1
consensus_chr11_55438305_55438441	Irf1	0	0	1
consensus_chr11_53799980_53800464	Irf1	0	1	1
consensus_chr3_31087364_31087512	Skil	0	1	0
consensus_chr4_140131460_140131654	Epha2	0	1	0
consensus_chr17_70900626_70901022	Pkdcc	0	1	0
consensus_chr1_180889952_180890153	Lefty2	0	1	0
consensus_chr1_180890289_180890577	Lefty2	0	1	0
consensus_chr4_146498084_146498487	Gm13247	0	1	1
consensus_chr7_144872813_144872953	Fgf15	0	1	0
consensus_chr1_180887934_180889501	Lefty2	1	1	1
consensus_chr11_103770780_103771467	Wnt3	1	1	1
consensus_chr17_17825762_17826184	Spaca6	0	1	0
consensus_chr5_67632834_67634059	Shisa3	1	1	1
consensus_chr9_4283761_4284275	Kbtbd3	1	1	1
consensus_chr11_103798585_103798879	Wnt3	0	1	0
consensus_chr19_45745737_45746430	Fgf8	1	1	1
consensus_chr19_45790984_45791161	Fgf8	0	1	0
consensus_chr9_118384752_118384964	Eomes	0	1	0
consensus_chr11_103772298_103772452	Wnt3	0	1	0
consensus_chr19_45743469_45743630	Fgf8	0	1	0
consensus_chr12_104470812_104470973	Gsc	0	1	0
consensus_chr19_45740924_45741066	Fgf8	0	1	0
consensus_chr9_118487571_118487804	Eomes	0	1	0
consensus_chr5_142812838_142813177	Radil	1	1	1
consensus_chr5_88679333_88679917	Grsf1	0	1	1
consensus_chr2_170887653_170887870	Dok5	0	1	0
consensus_chr11_100673015_100673393	Zfp385c	0	1	1
consensus_chr12_71369922_71370116	Dact1	0	1	0
consensus_chr13_114406841_114407585	Fst	0	1	0
consensus_chr5_65525609_65526464	Smim14	1	1	1
consensus_chr6_92247122_92248225	Trh	1	1	1
consensus_chr10_3741224_3742156	Plekhg1	0	0	1
consensus_chr6_97101730_97101881	Trh	0	1	0
consensus_chr6_97102282_97102637	Trh	0	1	1
consensus_chr12_71301571_71302020	Dact1	0	1	0
consensus_chr17_8389190_8389339	T	0	1	0
consensus_chr12_71369195_71369340	Dact1	0	1	0
consensus_chr2_155053130_155053272	Itch	0	1	0
consensus_chr2_170726216_170726396	Dok5	0	1	0
consensus_chr12_44393882_44394045	Nrcam	0	1	0
consensus_chr11_90391688_90391958	Hlf	0	1	0
consensus_chr11_97661653_97662273	Mllt6	1	1	1
consensus_chr17_83633016_83633583	Kcng3	0	1	1
consensus_chr4_142140118_142140268	Kazn	0	1	0
consensus_chr14_27066406_27066795	Il17rd	0	1	1
consensus_chr5_98310053_98310554	Fgf5	0	1	1
consensus_chr2_167058802_167059178	1500012F01Rik	0	1	0
consensus_chr12_44327670_44327856	Nrcam	0	1	0
consensus_chr5_98255741_98256081	Fgf5	0	1	1
consensus_chr6_144361308_144361462	Kras	0	1	0
consensus_chr12_103600388_103600857	Ppp4r4	0	1	1
consensus_chr17_8408979_8409262	T	0	1	0
consensus_chr6_145261829_145262239	Kras	0	1	1
consensus_chr11_90371054_90371268	Hlf	0	1	0
consensus_chr17_35504528_35506015	Pou5f1	1	1	1
consensus_chr3_89836176_89836827	AK040741	0	1	1
consensus_chr5_98283639_98283918	Fgf5	0	1	0

consensus_chr13_34741702_34742336	Pxdc1	0	1	1
consensus_chr17_83547624_83548022	Kcng3	0	1	0
consensus_chr12_71369195_71369340	Dact1	0	1	0
consensus_chr2_104188245_104188452	Cd59a	0	1	0
consensus_chr14_25718331_25718953	Il17rd	0	1	1
consensus_chr3_122414149_122414544	Bcar3	0	1	0
consensus_chr10_3851942_3852301	Plekhg1	0	1	1
consensus_chr4_142197117_142197286	Kazn	0	1	0
consensus_chr5_142881399_142881695	Radil	0	1	0
consensus_chr6_3623904_3624150	Calcr	0	1	1
consensus_chr12_44403632_44404236	Nrcam	0	1	1
consensus_chr13_94242174_94242361	Scamp1	0	1	0
consensus_chr15_62086921_62087198	Kcnq3	0	1	0
consensus_chr6_92469737_92470184	Trh	0	1	1
consensus_chr13_94321144_94321833	Scamp1	0	1	1
consensus_chr12_103599158_103599646	Ppp4r4	0	1	1
consensus_chr14_27044304_27044591	Il17rd	0	1	1
consensus_chr14_27044889_27045392	Il17rd	0	1	1
consensus_chr2_170657559_170658226	Dok5	0	1	1
consensus_chr5_117135600_117135895	Taok3	0	1	0
consensus_chr1_57890161_57890956	Spats2l	1	1	1
consensus_chr5_142546212_142546943	Radil	1	1	1
consensus_chr4_155189613_155190201	Ski	1	1	1
consensus_chr7_81497209_81497360	Ap3b2	0	1	0
consensus_chr12_102758889_102759365	Ubr7	1	1	1
consensus_chr2_69150788_69151293	Spc25	0	1	1
consensus_chr6_145275052_145275342	Kras	0	1	0
consensus_chr6_145275740_145276846	Kras	1	1	1
consensus_chr11_90356160_90356329	Hlf	0	1	0
consensus_chr14_27096185_27096406	Il17rd	0	1	0
consensus_chr3_122434049_122434426	Bcar3	1	1	1
consensus_chr7_132317113_132317585	Chst15	1	1	1
consensus_chr18_34549061_34549615	Wnt8a	0	1	1
consensus_chr18_34555243_34555601	Wnt8a	0	1	0
consensus_chr18_34551614_34552178	Wnt8a	0	1	1
consensus_chr5_98312032_98312336	Fgf5	0	1	0
consensus_chr6_30711408_30711749	Mest	0	1	0
consensus_chr11_100674104_100674553	Zfp385c	0	1	1
consensus_chr17_83694920_83695176	Kcng3	0	1	0
consensus_chr12_71296373_71297167	Dact1	0	1	0
consensus_chr6_144373492_144373661	Kras	0	1	0
consensus_chr5_88684336_88684662	Grsf1	0	1	0
consensus_chr7_132320651_132320895	Chst15	0	1	1
consensus_chr12_44535560_44536334	Nrcam	0	1	1
consensus_chr13_34685053_34685628	Pxdc1	0	1	1
consensus_chr2_155115402_155116063	Itch	1	1	1
consensus_chr4_155290432_155291069	Ski	0	1	1
consensus_chr17_8375801_8376001	T	0	1	0
consensus_chr17_8376105_8376842	T	1	1	1
consensus_chr5_117121502_117122157	Taok3	1	1	1
consensus_chr12_103511464_103511915	Ppp4r4	0	1	0
consensus_chr12_103564312_103565109	Ppp4r4	0	1	1
consensus_chr2_104109533_104109799	Cd59a	0	1	0
consensus_chr17_22745725_22746266	Gm16386	0	1	1
consensus_chr5_88670674_88670933	Grsf1	0	1	0
consensus_chr6_92513869_92514072	Trh	0	1	0
consensus_chr3_89800149_89800499	AK040741	1	1	1
consensus_chr1_57862138_57862396	Spats2l	0	1	0
consensus_chr7_132403936_132404243	Chst15	0	1	0
consensus_chr9_67105826_67106124	Tpm1	0	1	1

consensus_chr16_30129541_30129713	Hes1	0	1	0
consensus_chr18_75353497_75353669	Smad7	0	1	0
consensus_chr16_33605791_33606065	Hes1	0	1	0
consensus_chr2_70474647_70475140	Sp5	1	1	1
consensus_chr16_8788044_8788566	AK013883	0	1	1
consensus_chr16_30155736_30156494	Hes1	1	1	1
consensus_chr5_128593379_128593636	Fzd10	0	1	0
consensus_chr18_75360801_75361429	Smad7	1	1	1
consensus_chr14_23778040_23778573	Zmiz1	0	1	1
consensus_chr17_47953108_47953293	Foxp4	0	1	0
consensus_chr16_30160400_30160832	Hes1	0	1	0
consensus_chr16_33737816_33738120	Hes1	0	1	0
consensus_chr6_108644376_108644559	Gt(ROSA)26Sor	0	1	0
consensus_chr8_4436040_4436209	AK030646	0	0	1
consensus_chr16_8778570_8779080	AK013883	0	1	0
consensus_chr5_128596830_128597360	Fzd10	1	1	1
consensus_chr8_4436476_4436884	AK030646	0	1	1
consensus_chr13_5892395_5892635	Klf6	0	1	0
consensus_chr11_96368741_96368886	Hoxb1	0	1	0
consensus_chr5_147304967_147305067	Cdx2	0	0	0
consensus_chr9_58295776_58296231	Islr2	1	1	1
consensus_chr9_61324251_61324332	Tle3	1	0	0
consensus_chr10_44392015_44392352	Cd24a	0	1	0
consensus_chr2_131152139_131152505	Chac1	0	1	0
consensus_chr2_131152139_131152505	Chac1	0	1	0
consensus_chr10_59957919_59958288	Ddit4	0	1	1
consensus_chr2_119304534_119304780	Chac1	0	1	0
consensus_chr2_168840980_168841183	Sall4	0	1	1
consensus_chr3_86566305_86566532	Crabp2	0	1	0
consensus_chr4_138453360_138453557	Id3	0	1	0
consensus_chr1_191190073_191190359	Atf3	0	1	0
consensus_chr14_48653661_48653928	Otx2	0	1	1
consensus_chr14_49275422_49275653	Otx2	0	1	0
consensus_chr4_136031505_136031958	Id3	0	1	1
consensus_chr9_58279246_58279681	Islr2	0	1	1
consensus_chr7_130252207_130252429	Fgfr2	0	1	1
consensus_chrX_99152773_99152929	Efnb1	0	1	0
consensus_chr17_84152415_84152592	Zfp36l2	0	1	0
consensus_chr5_119613558_119613838	Tbx3	0	1	0
consensus_chr15_40955266_40955765	Klf10	1	1	1
consensus_chr7_73344144_73344284	Rgma	0	1	0
consensus_chr6_122366465_122366649	Phc1	0	0	0
consensus_chr2_168761948_168762206	Sall4	0	0	0
consensus_chr5_37768802_37769120	Msx1	0	1	0
consensus_chr5_37832855_37832998	Msx1	0	0	0
consensus_chr7_130327950_130328203	Fgfr2	0	1	1
consensus_chrX_139317424_139317594	Efnb1	0	1	0
consensus_chr11_96341761_96342424	Hoxb1	1	1	1
consensus_chr3_87891547_87891945	Crabp2	0	1	1
consensus_chr4_138361509_138362033	Id3	0	1	1
consensus_chr9_61339317_61339544	Tle3	0	1	1
consensus_chrX_139436747_139436935	Efnb1	0	1	0
consensus_chr10_43481135_43481437	Cd24a	0	1	0
consensus_chr14_49155640_49155812	Otx2	0	1	0
consensus_chr6_4855296_4855436	Peg10	0	1	0
consensus_chr14_48707986_48708322	Otx2	0	1	1
consensus_chr14_49089813_49089999	Otx2	0	1	0
consensus_chr2_168782911_168783312	Sall4	0	1	1
consensus_chr4_138418538_138418779	Id3	0	1	0
consensus_chr18_38278836_38278976	Pcdh1	0	1	0

consensus_chr2_126521591_126521758	Chac1	0	1	0
consensus_chr6_115287121_115287449	Cxcl12	0	1	0
consensus_chr2_126523532_126523676	Chac1	0	1	0
consensus_chr7_82063489_82063814	Rgma	0	1	0
consensus_chr6_117151734_117152782	Cxcl12	1	1	1
consensus_chr17_26511210_26511798	Dusp1	1	1	1
consensus_chr9_61293122_61293899	Tle3	1	1	1
consensus_chrX_139345212_139345622	Efnb1	0	1	0
consensus_chrX_145524466_145524856	Efnb1	0	1	1

8.6 List of motif-specific footprint frequencies over ATAC peaks or SMAD2 peaks

HOMER motif	sites in genome	sites in ATAC	sites in ATAC footprint	sites in SMAD2 peak	sites in SMAD2 footprint	ATAC footprint frequency (log)	SMAD2 footprint frequency (log)
pitx1	10016069	964354	24599	862	223	-1.59	-0.59
nanog	9834198	890053	19746	542	134	-1.65	-0.61
ar-half	7209449	767353	11199	417	66	-1.84	-0.80
nkx6.1	6805717	475797	7357	257	73	-1.81	-0.55
scl	6560719	851282	18584	649	118	-1.66	-0.74
tgif1	6437945	618756	9557	339	65	-1.81	-0.72
smad3	5518624	664998	12461	609	123	-1.73	-0.69
nkx2.1	5226505	535707	9566	303	67	-1.75	-0.66
isl1	4463906	363780	5690	213	43	-1.81	-0.69
crx	4442929	416197	8075	365	98	-1.71	-0.57
ptf1a	4213562	514295	11203	383	65	-1.66	-0.77
tbx5	3982484	445723	8486	336	60	-1.72	-0.75
nkx3.1	3963673	368769	5505	215	38	-1.83	-0.75
eomes	3811022	355764	6031	277	73	-1.77	-0.58
pr	3797953	354708	5598	192	40	-1.80	-0.68
bapx1	3767912	366053	5759	181	32	-1.80	-0.75
foxo1	3738563	304294	5818	208	56	-1.72	-0.57
erra	3721067	437760	9684	329	74	-1.66	-0.65
nkx2.5	3699452	378341	6356	207	40	-1.77	-0.71
tata	3603286	270543	6364	142	54	-1.63	-0.42
sox3	3527699	332270	7452	285	85	-1.65	-0.53
sox6	3355608	288321	6192	220	64	-1.67	-0.54
foxa1.lncap	3345033	235320	3958	148	38	-1.77	-0.59
lhx3	3288189	246652	4662	182	62	-1.72	-0.47
sox10	3235939	312961	7398	258	80	-1.63	-0.51
olig2	3157698	328345	5856	211	25	-1.75	-0.93
nkx2.2	2916586	317317	5761	199	44	-1.74	-0.66
lhx2	2887322	202500	4228	134	51	-1.68	-0.42
cdx2	2852278	163565	2363	62	20	-1.84	-0.49
nf1-half	2835353	334796	7592	166	33	-1.64	-0.70
bmal	2670245	229723	3949	91	19	-1.76	-0.68
hoxd13	2620504	154671	2586	80	28	-1.78	-0.46
foxa1.mcf7	2573821	179789	3058	116	32	-1.77	-0.56
heb	2559526	343681	7934	281	53	-1.64	-0.72
pit1	2541918	175989	1902	69	23	-1.97	-0.48
myb	2541040	254419	5444	132	29	-1.67	-0.66
ews-erg	2513320	263206	6923	183	41	-1.58	-0.65
gata3	2482027	197454	2947	113	24	-1.83	-0.67
pu1-irf	2377597	226788	5447	145	45	-1.62	-0.51
smad4	2341193	283129	5278	321	61	-1.73	-0.72
meis1	2281654	263212	4889	159	27	-1.73	-0.77
lhx1	2254526	165265	3484	132	48	-1.68	-0.44
znf711	2234430	317331	8894	250	43	-1.55	-0.76
myba	2198357	236529	5571	159	38	-1.63	-0.62
erg	2194254	260344	8403	202	49	-1.49	-0.62
bcl6	2157378	203869	5933	123	29	-1.54	-0.63
hoxb13	2157238	154206	2589	84	32	-1.77	-0.42
bmyb	2127165	211669	5117	149	41	-1.62	-0.56

pdx1	2119981	150041	2292	67	15	-1.82	-0.65
znf263	2058671	322601	7251	214	38	-1.65	-0.75
zfx	2053936	291645	8176	219	38	-1.55	-0.76
foxl2	2005409	137860	2228	106	32	-1.79	-0.52
tgif2	1945861	214368	3553	135	21	-1.78	-0.81
gata4	1941236	158308	2404	114	27	-1.82	-0.63
nfatc1	1889112	148972	3751	67	13	-1.60	-0.71
rbpj1	1885981	215432	5905	161	43	-1.56	-0.57
gsc	1877638	173854	3716	199	64	-1.67	-0.49
klf14	1855699	357306	27323	310	98	-1.12	-0.50
ehf	1824055	201525	6856	157	45	-1.47	-0.54
fox-ebox	1773130	155837	2579	97	23	-1.78	-0.63
stat6.2	1767501	165967	4175	108	28	-1.60	-0.59
tbet	1712834	168414	2841	134	36	-1.77	-0.57
limb	1685083	104053	1860	61	19	-1.75	-0.51
foxm1	1669864	120812	2186	69	19	-1.74	-0.56
runx1	1668328	176268	3020	87	7	-1.77	-1.09
foxa2	1659119	136559	2286	83	23	-1.78	-0.56
nfy	1650937	182001	6653	78	18	-1.44	-0.64
znf416	1642776	203491	5093	155	32	-1.60	-0.69
sox15	1630290	160634	4091	171	57	-1.59	-0.48
ascl1	1627717	223102	6040	180	28	-1.57	-0.81
e47	1621432	191552	4073	155	18	-1.67	-0.94
ets1	1599701	180180	6218	138	35	-1.46	-0.60
sox4	1595594	163521	4367	137	47	-1.57	-0.46
mef2b	1592140	108300	2020	61	18	-1.73	-0.53
smad2	1588487	205815	4061	256	51	-1.70	-0.70
spdef	1580016	185721	5493	142	37	-1.53	-0.58
etv1	1575152	193995	8047	162	41	-1.38	-0.60
sox2	1545096	158262	4620	179	65	-1.53	-0.44
rxr	1469871	193479	4300	148	17	-1.65	-0.94
stat4	1408167	126376	3627	64	17	-1.54	-0.58
e2a	1402823	198205	4674	193	30	-1.63	-0.81
mitf	1391635	139407	4157	64	17	-1.53	-0.58
otx2	1365341	131859	3178	160	55	-1.62	-0.46
npas2	1332011	132933	3270	69	14	-1.61	-0.69
prdm1	1318613	133172	2870	62	14	-1.67	-0.65
runx2	1307223	125822	1855	64	7	-1.83	-0.96
atoh1	1279449	153255	3515	121	15	-1.64	-0.91
hoxa9	1278431	90551	1529	43	6	-1.77	-0.86
ap2gamma	1249004	206360	7399	190	46	-1.45	-0.62
chr	1217444	89564	2482	61	23	-1.56	-0.42
cebp	1212856	95546	2266	67	21	-1.62	-0.50
ap4	1197289	174000	3720	138	27	-1.67	-0.71
etv2	1187964	136454	4701	119	26	-1.46	-0.66
zic	1185001	158147	4448	183	41	-1.55	-0.65
znf189	1177088	134356	2931	86	21	-1.66	-0.61
cebp-ap1	1174249	102547	2207	72	25	-1.67	-0.46
tead4	1163080	126353	3263	101	30	-1.59	-0.53
tead	1124814	97079	2659	77	26	-1.56	-0.47
runx	1119633	110689	1527	55	6	-1.86	-0.96
sox9	1113476	119873	3242	117	39	-1.57	-0.48
pparg	1109796	145895	3742	126	21	-1.59	-0.78
gata2	1105918	97969	1463	72	16	-1.83	-0.65
fli	1082602	132332	7656	110	32	-1.24	-0.54
mafa	1077506	138493	2809	65	12	-1.69	-0.73
38261	1076815	89936	2519	104	43	-1.55	-0.38
tcf12	1063012	169023	4263	151	28	-1.60	-0.73
elf5	1054974	120640	4698	91	23	-1.41	-0.60
foxh1	1036239	83074	1613	170	42	-1.71	-0.61

znf467	1029950	183982	6756	172	47	-1.44	-0.56
runx1m	1024977	102885	1786	60	9	-1.76	-0.82
neurod1	1007076	118116	2248	86	6	-1.72	-1.16
hnf6	989462	67427	1214	32	6	-1.74	-0.73
srf	985021	78885	1630	49	14	-1.68	-0.54
hoxc9	984443	67778	1187	24	3	-1.76	-0.90
klf5	981358	205080	19694	158	45	-1.02	-0.55
gfi1b	978754	91288	1767	57	11	-1.71	-0.71
gata	977236	83899	1117	58	10	-1.88	-0.76
lrh1	974052	124742	2672	131	41	-1.67	-0.50
max	966926	92280	2691	41	9	-1.54	-0.66
ap1	958429	117983	2047	43	11	-1.76	-0.59
phox2a	948826	49820	742	25	10	-1.83	-0.40
tcf21	935736	132862	2882	120	18	-1.66	-0.82
gabp	922505	116349	5557	103	25	-1.32	-0.61
ebf-old	915936	155490	4550	151	31	-1.53	-0.69
bzip-irf	911889	66163	1253	35	13	-1.72	-0.43
tcf4	904892	77833	2109	105	32	-1.57	-0.52
atf3	903419	112971	2004	40	11	-1.75	-0.56
38991	897486	79810	2402	72	32	-1.52	-0.35
klf10	884635	104749	4490	71	14	-1.37	-0.71
esrrb	864116	98229	2165	77	23	-1.66	-0.52
stat3.il23	857762	90395	3030	57	15	-1.47	-0.58
foxp1	837909	60119	1201	62	16	-1.70	-0.59
stat6	829381	69617	1855	44	10	-1.57	-0.64
mef2c	821404	49259	968	24	9	-1.71	-0.43
mef2a	808982	54578	965	25	13	-1.75	-0.28
hif1b	807672	91114	4881	76	18	-1.27	-0.63
myog	798668	126159	2569	96	15	-1.69	-0.81
atf1	798174	75943	2397	50	17	-1.50	-0.47
batf	787497	98250	1653	34	7	-1.77	-0.69
37530	765658	56082	1708	60	25	-1.52	-0.38
tead2	736469	76585	1916	56	15	-1.60	-0.57
fra1	714037	90170	1646	30	9	-1.74	-0.52
zbtb18	713678	79735	1332	59	7	-1.78	-0.93
pit1-long	712691	41855	553	14	6	-1.88	-0.37
pu1	711817	84548	1959	73	12	-1.64	-0.78
myod	704400	116699	3213	93	14	-1.56	-0.82
klf4	703567	125652	6143	113	26	-1.31	-0.64
znf415	702697	101175	2427	70	13	-1.62	-0.73
znf264	669961	77223	1989	59	9	-1.59	-0.82
klf9	662528	116102	8331	100	33	-1.14	-0.48
ews-fli1	657733	74383	3360	63	15	-1.35	-0.62
nmyc	654103	71276	2896	42	8	-1.39	-0.72
brn1	643319	57645	2035	58	28	-1.45	-0.32
nr5a2	634093	81747	1987	99	32	-1.61	-0.49
myf5	619152	91968	1833	70	10	-1.70	-0.85
egr1	610169	105119	6446	101	31	-1.21	-0.51
grhl2	609736	52605	1243	39	9	-1.63	-0.64
irf4	585966	51019	1399	33	11	-1.56	-0.48
hnf4	581622	63317	1424	43	5	-1.65	-0.93
zbtb12	568496	53817	1469	25	6	-1.56	-0.62
es	568410	102175	3955	136	37	-1.41	-0.57
tcf3	546969	48932	1430	71	21	-1.53	-0.53
ap2	537778	100623	4161	102	24	-1.38	-0.63
p65	532145	69854	2179	56	12	-1.51	-0.67
atf7	530685	54474	1891	32	10	-1.46	-0.51
arnt	530307	61563	2839	46	6	-1.34	-0.88
cmyc	521898	53280	1226	22	3	-1.64	-0.87
stat5	498175	44635	1418	20	3	-1.50	-0.82

stat3	475575	50008	1839	23	5	-1.43	-0.66
six1	474218	40857	930	20	5	-1.64	-0.60
rfx5	457774	50354	1751	25	11	-1.46	-0.36
elk1	455869	63483	5812	69	23	-1.04	-0.48
fxr	439455	55795	1279	52	9	-1.64	-0.76
jun-cre	431028	45325	1472	29	10	-1.49	-0.46
prdm9	430181	55644	1523	31	7	-1.56	-0.65
ets-distal	426373	44647	938	41	6	-1.68	-0.83
stat1	423898	36466	1190	16	3	-1.49	-0.73
mafF	422477	33903	546	12	1	-1.79	-1.08
usf1	417169	47187	2389	24	5	-1.30	-0.68
oct4-sox2	416667	44556	1905	81	37	-1.37	-0.34
elk4	410819	58888	6069	60	22	-0.99	-0.44
atf4	399146	32723	719	20	6	-1.66	-0.52
p63	392234	43463	981	24	4	-1.65	-0.78
prdm14	385726	40017	830	41	11	-1.68	-0.57
hsf	381677	33914	849	12	3	-1.60	-0.60
dmrt1	380325	28084	389	18	8	-1.86	-0.35
pax3-fkhr	372352	28097	623	13	5	-1.65	-0.41
staf	370620	50156	2035	37	8	-1.39	-0.67
rbpj1-ebox	368125	41999	1046	27	6	-1.60	-0.65
pax5	366513	47362	1301	29	9	-1.56	-0.51
tra	356491	47580	1159	33	4	-1.61	-0.92
pgr	350428	27932	395	12	5	-1.85	-0.38
usf2	346381	31668	1584	15	6	-1.30	-0.40
brachyury	335474	36433	832	25	6	-1.64	-0.62
fosl2	334381	56308	1105	18	3	-1.71	-0.78
pax8	333253	44772	1104	32	9	-1.61	-0.55
elf1	328690	50756	4718	52	16	-1.03	-0.51
dmrt6	327908	25021	329	13	4	-1.88	-0.51
tlx	320280	45624	1143	40	10	-1.60	-0.60
atf2	317164	36154	1466	23	9	-1.39	-0.41
hif2a	314950	36589	1311	31	9	-1.45	-0.54
are	314056	32324	527	12	1	-1.79	-1.08
ere	311062	30459	567	14	2	-1.73	-0.85
pu1-irf8	296017	25025	618	12	7	-1.61	-0.23
mef2d	295246	18774	415	12	5	-1.66	-0.38
chop	290085	25034	549	17	6	-1.66	-0.45
maz	287824	40358	1620	40	10	-1.40	-0.60
gre-raw	283057	27316	403	12	4	-1.83	-0.48
nf1	281758	34554	992	21	6	-1.54	-0.54
pbx3	280658	32658	864	28	3	-1.58	-0.97
tcf7l2	279223	26608	767	44	16	-1.54	-0.44
spib	276661	29664	933	20	1	-1.50	-1.30
cMyc.lncap	270913	34353	2649	24	9	-1.11	-0.43
tbx20	269994	36193	696	13	1	-1.72	-1.11
oct4-sox17	269273	19315	322	8	4	-1.78	-0.30
clock	268260	30620	2084	21	8	-1.17	-0.42
irf1	266591	20129	668	10	7	-1.48	-0.15
nur77	265708	25197	694	13	3	-1.56	-0.64
nfat-ap1	260330	23557	497	20	3	-1.68	-0.82
hnf1	259795	14538	233	2	1	-1.80	-0.30
jun-ap1	256439	47274	943	14	1	-1.70	-1.15
rorc	256042	24706	332	15	3	-1.87	-0.70
bhlhe40	249161	28047	2036	21	9	-1.14	-0.37
srebp1a	247032	31549	453	26	3	-1.84	-0.94
mafk	242025	37935	779	14	3	-1.69	-0.67
znf322	241944	40372	926	25	6	-1.64	-0.62
cebp-cebp	241506	17615	353	12	6	-1.70	-0.30
gata3.dr4	240608	12690	170	8	3	-1.87	-0.43

pknox1	240294	28215	887	22	2	-1.50	-1.04
batf3-irf8	236536	16510	303	9	3	-1.74	-0.48
gata3.ir3	231264	17312	306	9	2	-1.75	-0.65
ets	223002	35459	3376	39	12	-1.02	-0.51
pax7	217108	12085	195	5	0	-1.79	N.A
tbox-smad	206242	26143	441	50	12	-1.77	-0.62
rfx1	200908	26773	1349	13	7	-1.30	-0.27
vdr	196163	25596	549	16	2	-1.67	-0.90
are-fox	192940	12284	203	5	0	-1.78	N.A
hoxb4	187122	17298	333	9	2	-1.72	-0.65
hsf1	186601	18422	649	8	1	-1.45	-0.90
hoxa2	184953	13921	200	8	4	-1.84	-0.30
znf136	182892	11342	179	2	0	-1.80	N.A
znf519	180652	30037	886	32	6	-1.53	-0.73
eklf	176294	28396	1226	14	1	-1.36	-1.15
egr2	176057	29207	2556	13	5	-1.06	-0.41
ebf	163243	27538	874	28	4	-1.50	-0.85
znf675	162826	16629	425	9	4	-1.59	-0.35
cre	156642	21364	1756	14	7	-1.09	-0.30
hif1a	154678	18656	973	17	4	-1.28	-0.63
bach2	153033	28026	610	9	2	-1.66	-0.65
rfx2	152197	21230	1234	10	6	-1.24	-0.22
boris	151644	44710	7607	27	12	-0.77	-0.35
znf317	151518	18892	362	8	1	-1.72	-0.90
p53	149340	12683	184	5	2	-1.84	-0.40
gata-scl	146199	16232	320	11	2	-1.71	-0.74
nfkb	141801	14488	543	13	5	-1.43	-0.41
reverb	140881	16785	278	7	1	-1.78	-0.85
gli	139238	19934	330	12	2	-1.78	-0.78
gata3.dr8	138954	10334	176	8	4	-1.77	-0.30
gre	136233	14391	247	7	1	-1.77	-0.85
tcfcp2l1	134210	18497	670	11	3	-1.44	-0.56
gata3.ir4	130386	9766	157	10	3	-1.79	-0.52
srebp2	129705	17176	351	13	3	-1.69	-0.64
pax6	124645	10705	242	5	2	-1.65	-0.40
jund	123860	14298	862	10	4	-1.22	-0.40
pax5-short	123448	15499	365	13	2	-1.63	-0.81
e2f4	120032	25868	3669	26	9	-0.85	-0.46
isre	119615	8913	247	5	3	-1.56	-0.22
ctcf	115851	37991	6399	12	8	-0.77	-0.18
znf692	111300	17723	469	23	1	-1.58	-1.36
tr4	108606	13228	319	10	2	-1.62	-0.70
xbox	107658	15094	914	7	5	-1.22	-0.15
pbx1	107106	14203	423	15	2	-1.53	-0.88
zfp809	105417	19844	983	15	2	-1.31	-0.88
dmc1	100382	8654	185	4	0	-1.67	N.A
rarg	97325	9348	178	14	6	-1.72	-0.37
e2f6	96222	27480	3761	38	12	-0.86	-0.50
p50	77416	12915	478	18	1	-1.43	-1.26
brn2	76190	5993	107	4	1	-1.75	-0.60
sp1	73046	33611	9936	30	20	-0.53	-0.18
irf2	72068	5873	238	4	2	-1.39	-0.30
ets-runx	70145	10164	433	5	0	-1.37	N.A
ets-ebox	68801	9576	225	5	0	-1.63	N.A
znf165	68046	12011	509	8	2	-1.37	-0.60
yy1	64857	6880	739	2	1	-0.97	-0.30
nf1-fox	64718	5372	116	3	0	-1.67	N.A
pax7-longest	64368	3383	32	0	0	-2.02	N.A
pax7-long	63186	4189	88	4	2	-1.68	-0.30
rfx3	61630	11856	1060	7	6	-1.05	-0.07

nfe2	61199	11644	273	6	2	-1.63	-0.48
bach1	60524	11167	274	4	2	-1.61	-0.30
zscan22	58251	9892	394	11	3	-1.40	-0.56
znf382	55811	3437	75	1	1	-1.66	0.00
e2f1	55599	15143	2258	18	3	-0.83	-0.78
gfy	54830	7918	673	8	0	-1.07	N.A
znf669	51996	6225	240	9	3	-1.41	-0.48
lxre	50743	6130	154	6	1	-1.60	-0.78
ebox	50258	7263	953	4	1	-0.88	-0.60
znf41	48800	3331	81	2	0	-1.61	N.A
nrf2	44639	8914	223	4	2	-1.60	-0.30
gfy-staf	39001	5863	1044	6	3	-0.75	-0.30
nrf1	38990	8182	2866	8	1	-0.46	-0.90
oct.ir1	35399	2587	56	2	0	-1.66	N.A
ebna1	28849	2917	62	2	0	-1.67	N.A
e2f7	28821	7874	1332	9	1	-0.77	-0.95
nrf	25170	5405	2176	6	1	-0.40	-0.78
dr5	17437	2568	69	2	0	-1.57	N.A
e2f	16477	3881	811	2	1	-0.68	-0.30
t1isre	15228	1108	44	0	0	-1.40	N.A
oct.ir	15140	1214	41	0	0	-1.47	N.A
ctcf-mys	14967	3863	293	0	0	-1.12	N.A
p53-myc	13539	2278	37	0	0	-1.79	N.A
znf16	5479	846	31	1	0	-1.44	N.A
zbtb33	5214	1747	546	1	0	-0.51	N.A
znf528	4939	527	12	0	0	-1.64	N.A
rest	3887	1864	291	0	0	-0.81	N.A
zfp3	1687	188	2	1	0	-1.97	N.A
gfx	1481	557	288	0	0	-0.29	N.A

8.7 List of the 'SBS high footprint frequency' motifs

HOMER motif	sites in SMAD2 peak	ATAC footprint frequency (log)	SMAD2 footprint frequency (log)
pitx1	223	-1.59	-0.59
nanog	134	-1.65	-0.61
smad3	123	-1.73	-0.69
scl	118	-1.66	-0.74
crx	98	-1.71	-0.57
sox3	85	-1.65	-0.53
sox10	80	-1.63	-0.51
erra	74	-1.66	-0.65
eomes	73	-1.77	-0.58
nkx6.1	73	-1.81	-0.55
nkx2.1	67	-1.75	-0.66
ar-half	66	-1.84	-0.80
ptf1a	65	-1.66	-0.77
sox2	65	-1.53	-0.44
tgif1	65	-1.81	-0.72
gsc	64	-1.67	-0.49
sox6	64	-1.67	-0.54
lhx3	62	-1.72	-0.47
smad4	61	-1.73	-0.72
tbx5	60	-1.72	-0.75
sox15	57	-1.59	-0.48
foxo1	56	-1.72	-0.57
otx2	55	-1.62	-0.46
tata	54	-1.63	-0.42
heb	53	-1.64	-0.72
lhx2	51	-1.68	-0.42
smad2	51	-1.70	-0.70
erg	49	-1.49	-0.62
lhx1	48	-1.68	-0.44
sox4	47	-1.57	-0.46
znf467	47	-1.44	-0.56
ap2gamma	46	-1.45	-0.62
ehf	45	-1.47	-0.54
pu1-irf	45	-1.62	-0.51
nkx2.2	44	-1.74	-0.66
38261	43	-1.55	-0.38
isl1	43	-1.81	-0.69
rbpj1	43	-1.56	-0.57
znf711	43	-1.55	-0.76
foxh1	42	-1.71	-0.61
bmyb	41	-1.62	-0.56
etv1	41	-1.38	-0.60
ews-erg	41	-1.58	-0.65
lrh1	41	-1.67	-0.50
zic	41	-1.55	-0.65
nkx2.5	40	-1.77	-0.71
pr	40	-1.80	-0.68
sox9	39	-1.57	-0.48

foxa1.lncap	38	-1.77	-0.59
myba	38	-1.63	-0.62
nkx3.1	38	-1.83	-0.75
zfx	38	-1.55	-0.76
znf263	38	-1.65	-0.75
es	37	-1.41	-0.57
oct4-sox2	37	-1.37	-0.34
spdef	37	-1.53	-0.58
tbet	36	-1.77	-0.57
ets1	35	-1.46	-0.60
nf1-half	33	-1.64	-0.70
38991	32	-1.52	-0.35
bapx1	32	-1.80	-0.75
foxa1.mcf7	32	-1.77	-0.56
foxl2	32	-1.79	-0.52
hoxb13	32	-1.77	-0.42
nr5a2	32	-1.61	-0.49
tcf4	32	-1.57	-0.52
znf416	32	-1.60	-0.69
ebf-old	31	-1.53	-0.69
e2a	30	-1.63	-0.81
tead4	30	-1.59	-0.53
bcl6	29	-1.54	-0.63
myb	29	-1.67	-0.66
ascl1	28	-1.57	-0.81
brn1	28	-1.45	-0.32
hoxd13	28	-1.78	-0.46
stat6.2	28	-1.60	-0.59
tcf12	28	-1.60	-0.73
ap4	27	-1.67	-0.71
gata4	27	-1.82	-0.63
meis1	27	-1.73	-0.77
etv2	26	-1.46	-0.66
klf4	26	-1.31	-0.64
tead	26	-1.56	-0.47
37530	25	-1.52	-0.38
cebp-ap1	25	-1.67	-0.46
gabp	25	-1.32	-0.61
olig2	25	-1.75	-0.93
ap2	24	-1.38	-0.63
gata3	24	-1.83	-0.67
chr	23	-1.56	-0.42
elf5	23	-1.41	-0.60
esrrb	23	-1.66	-0.52
fox-ebox	23	-1.78	-0.63
foxa2	23	-1.78	-0.56
pit1	23	-1.97	-0.48
cebp	21	-1.62	-0.50
pparg	21	-1.59	-0.78
tcf3	21	-1.53	-0.53
tgif2	21	-1.78	-0.81
znf189	21	-1.66	-0.61
cdx2	20	-1.84	-0.49
bmal	19	-1.76	-0.68
foxm1	19	-1.74	-0.56
limb	19	-1.75	-0.51
e47	18	-1.67	-0.94
hif1b	18	-1.27	-0.63
mef2b	18	-1.73	-0.53
nfy	18	-1.44	-0.64

tcf21	18	-1.66	-0.82
atf1	17	-1.50	-0.47
mitf	17	-1.53	-0.58
rxr	17	-1.65	-0.94
stat4	17	-1.54	-0.58
foxp1	16	-1.70	-0.59
gata2	16	-1.83	-0.65
tcf7l2	16	-1.54	-0.44
atoh1	15	-1.64	-0.91
ews-fli1	15	-1.35	-0.62
myog	15	-1.69	-0.81
pdx1	15	-1.82	-0.65
stat3.il23	15	-1.47	-0.58
tead2	15	-1.60	-0.57
klf10	14	-1.37	-0.71
myod	14	-1.56	-0.82
npas2	14	-1.61	-0.69
prdm1	14	-1.67	-0.65
srf	14	-1.68	-0.54
bzip-irf	13	-1.72	-0.43
mef2a	13	-1.75	-0.28
nfatc1	13	-1.60	-0.71
znf415	13	-1.62	-0.73
mafa	12	-1.69	-0.73
p65	12	-1.51	-0.67
pu1	12	-1.64	-0.78
tbox-smad	12	-1.77	-0.62
ap1	11	-1.76	-0.59
atf3	11	-1.75	-0.56
gfi1b	11	-1.71	-0.71
irf4	11	-1.56	-0.48
prdm14	11	-1.68	-0.57
rfx5	11	-1.46	-0.36
atf7	10	-1.46	-0.51
gata	10	-1.88	-0.76
jun-cre	10	-1.49	-0.46
maz	10	-1.40	-0.60
myf5	10	-1.70	-0.85
phox2a	10	-1.83	-0.40
stat6	10	-1.57	-0.64
tlx	10	-1.60	-0.60
atf2	9	-1.39	-0.41
fra1	9	-1.74	-0.52
fxr	9	-1.64	-0.76
grhl2	9	-1.63	-0.64
hif2a	9	-1.45	-0.54
max	9	-1.54	-0.66
mef2c	9	-1.71	-0.43
pax5	9	-1.56	-0.51
pax8	9	-1.61	-0.55
runx1m	9	-1.76	-0.82
znf264	9	-1.59	-0.82
dmrt1	8	-1.86	-0.35
nmyc	8	-1.39	-0.72
staf	8	-1.39	-0.67
batf	7	-1.77	-0.69
irf1	7	-1.48	-0.15
prdm9	7	-1.56	-0.65
pu1-irf8	7	-1.61	-0.23
rfx1	7	-1.30	-0.27

runx2	7	-1.83	-0.96
zbtb18	7	-1.78	-0.93
arnt	6	-1.34	-0.88
atf4	6	-1.66	-0.52
brachyury	6	-1.64	-0.62
cebp-cebp	6	-1.70	-0.30
chop	6	-1.66	-0.45
ets-distal	6	-1.68	-0.83
hnf6	6	-1.74	-0.73
hoxa9	6	-1.77	-0.86
nf1	6	-1.54	-0.54
pit1-long	6	-1.88	-0.37
rarg	6	-1.72	-0.37
rbpj1-ebox	6	-1.60	-0.65
runx	6	-1.86	-0.96
usf2	6	-1.30	-0.40
zbtb12	6	-1.56	-0.62
znf322	6	-1.64	-0.62
znf519	6	-1.53	-0.73
hnf4	5	-1.65	-0.93
mef2d	5	-1.66	-0.38
nfkb	5	-1.43	-0.41
pax3-fkhr	5	-1.65	-0.41
pgr	5	-1.85	-0.38
six1	5	-1.64	-0.60
stat3	5	-1.43	-0.66
usf1	5	-1.30	-0.68
dmrt6	4	-1.88	-0.51
ebf	4	-1.50	-0.85
gata3.dr8	4	-1.77	-0.30
gre-raw	4	-1.83	-0.48
hif1a	4	-1.28	-0.63
hoxa2	4	-1.84	-0.30
oct4-sox17	4	-1.78	-0.30
p63	4	-1.65	-0.78
tra	4	-1.61	-0.92
znf675	4	-1.59	-0.35
batf3-irf8	3	-1.74	-0.48
cmyc	3	-1.64	-0.87
fosl2	3	-1.71	-0.78
gata3.dr4	3	-1.87	-0.43
gata3.ir4	3	-1.79	-0.52
hoxc9	3	-1.76	-0.90
hsf	3	-1.60	-0.60
isre	3	-1.56	-0.22
mafk	3	-1.69	-0.67
nfat-ap1	3	-1.68	-0.82
nur77	3	-1.56	-0.64
pbx3	3	-1.58	-0.97
rorc	3	-1.87	-0.70
srebp1a	3	-1.84	-0.94
srebp2	3	-1.69	-0.64
stat1	3	-1.49	-0.73
stat5	3	-1.50	-0.82
tcfcp2l1	3	-1.44	-0.56
znf669	3	-1.41	-0.48
zscan22	3	-1.40	-0.56
bach1	2	-1.61	-0.30
bach2	2	-1.66	-0.65
ere	2	-1.73	-0.85

gata-scl	2	-1.71	-0.74
gata3.ir3	2	-1.75	-0.65
gli	2	-1.78	-0.78
hoxb4	2	-1.72	-0.65
irf2	2	-1.39	-0.30
nfe2	2	-1.63	-0.48
nrf2	2	-1.60	-0.30
p53	2	-1.84	-0.40
pax5-short	2	-1.63	-0.81
pax6	2	-1.65	-0.40
pax7-long	2	-1.68	-0.30
pbx1	2	-1.53	-0.88
tr4	2	-1.62	-0.70
vdr	2	-1.67	-0.90
zfp809	2	-1.31	-0.88
znf165	2	-1.37	-0.60
brn2	1	-1.75	-0.60
gre	1	-1.77	-0.85
hnf1	1	-1.80	-0.30
hsf1	1	-1.45	-0.90
lxre	1	-1.60	-0.78
reverb	1	-1.78	-0.85
znf317	1	-1.72	-0.90
znf382	1	-1.66	0.00

Reference List

ADELMAN, K. & LIS, J. T. 2012. Promoter-proximal pausing of RNA polymerase II: emerging roles in metazoans. *Nat Rev Genet,* 13**,** 720-31.

ADEY, A., MORRISON, H. G., ASAN, XUN, X., KITZMAN, J. O., TURNER, E. H., STACKHOUSE, B., MACKENZIE, A. P., CARUCCIO, N. C., ZHANG, X. & SHENDURE, J. 2010. Rapid, low-input, low-bias construction of shotgun fragment libraries by high-density in vitro transposition. *Genome Biol,* 11**,** R119.

AKIZU, N., ESTARAS, C., GUERRERO, L., MARTI, E. & MARTINEZ-BALBAS, M. A. 2010. H3K27me3 regulates BMP activity in developing spinal cord. *Development,* 137**,** 2915-25.

ALARCON, C., ZAROMYTIDOU, A. I., XI, Q., GAO, S., YU, J., FUJISAWA, S., BARLAS, A., MILLER, A. N., MANOVA-TODOROVA, K., MACIAS, M. J., SAPKOTA, G., PAN, D. & MASSAGUE, J. 2009. Nuclear CDKs drive Smad transcriptional activation and turnover in BMP and TGF-beta pathways. *Cell,* 139**,** 757-69.

ANDERS, S. & HUBER, W. 2010. Differential expression analysis for sequence count data. *Genome Biol,* 11**,** R106.

ANTSIFEROVA, M., HUBER, M., MEYER, M., PIWKO-CZUCHRA, A., RAMADAN, T., MACLEOD, A. S., HAVRAN, W. L., DUMMER, R., HOHL, D. & WERNER, S. 2011. Activin enhances skin tumourigenesis and malignant progression by inducing a pro-tumourigenic immune cell response. *Nat Commun,* 2**,** 576.

ANTSIFEROVA, M. & WERNER, S. 2012. The bright and the dark sides of activin in wound healing and cancer. *J Cell Sci,* 125**,** 3929-37.

ARAGON, E., GOERNER, N., XI, Q., GOMES, T., GAO, S., MASSAGUE, J. & MACIAS, M. J. 2012. Structural basis for the versatile interactions of Smad7 with regulator WW domains in TGF-beta Pathways. *Structure,* 20**,** 1726-36.

ARAGON, E., GOERNER, N., ZAROMYTIDOU, A. I., XI, Q., ESCOBEDO, A., MASSAGUE, J. & MACIAS, M. J. 2011. A Smad action turnover switch operated by WW domain readers of a phosphoserine code. *Genes Dev,* 25**,** 1275-88.

ARNOLD, S. J. & ROBERTSON, E. J. 2009. Making a commitment: cell lineage allocation and axis patterning in the early mouse embryo. *Nat Rev Mol Cell Biol,* 10**,** 91-103.

ASHE, H. L. & BRISCOE, J. 2006. The interpretation of morphogen gradients. *Development,* 133**,** 385-94.

ATTISANO, L. & WRANA, J. L. 2013. Signal integration in TGF-beta, WNT, and Hippo pathways. *F1000Prime Rep,* 5**,** 17.

AVERY, S., ZAFARANA, G., GOKHALE, P. J. & ANDREWS, P. W. 2010. The role of SMAD4 in human embryonic stem cell self-renewal and stem cell fate. *Stem Cells,* 28**,** 863-73.

BANNISTER, A. J. & KOUZARIDES, T. 2011. Regulation of chromatin by histone modifications. *Cell Res,* 21**,** 381-95.

BAROZZI, I., SIMONATTO, M., BONIFACIO, S., YANG, L., ROHS, R., GHISLETTI, S. & NATOLI, G. 2014. Coregulation of transcription factor binding and nucleosome occupancy through DNA features of mammalian enhancers. *Mol Cell,* 54**,** 844-857.

BEDDINGTON, R. S. & ROBERTSON, E. J. 1999. Axis development and early asymmetry in mammals. *Cell,* 96**,** 195-209.

BERGERON-SANDOVAL, L. P., SAFAEE, N. & MICHNICK, S. W. 2016. Mechanisms and Consequences of Macromolecular Phase Separation. *Cell,* 165**,** 1067-1079.

BERNSTEIN, B. E., MIKKELSEN, T. S., XIE, X., KAMAL, M., HUEBERT, D. J., CUFF, J., FRY, B., MEISSNER, A., WERNIG, M., PLATH, K., JAENISCH, R., WAGSCHAL, A., FEIL, R., SCHREIBER, S. L. & LANDER, E. S. 2006. A bivalent

chromatin structure marks key developmental genes in embryonic stem cells. *Cell,* 125**,** 315-26.

BERTERO, A., MADRIGAL, P., GALLI, A., HUBNER, N. C., MORENO, I., BURKS, D., BROWN, S., PEDERSEN, R. A., GAFFNEY, D., MENDJAN, S., PAUKLIN, S. & VALLIER, L. 2015. Activin/nodal signaling and NANOG orchestrate human embryonic stem cell fate decisions by controlling the H3K4me3 chromatin mark. *Genes Dev,* 29**,** 702-17.

BEYER, T. A., WEISS, A., KHOMCHUK, Y., HUANG, K., OGUNJIMI, A. A., VARELAS, X. & WRANA, J. L. 2013. Switch enhancers interpret TGF-beta and Hippo signaling to control cell fate in human embryonic stem cells. *Cell Rep,* 5**,** 1611-24.

BOYER, L. A., LEE, T. I., COLE, M. F., JOHNSTONE, S. E., LEVINE, S. S., ZUCKER, J. P., GUENTHER, M. G., KUMAR, R. M., MURRAY, H. L., JENNER, R. G., GIFFORD, D. K., MELTON, D. A., JAENISCH, R. & YOUNG, R. A. 2005. Core transcriptional regulatory circuitry in human embryonic stem cells. *Cell,* 122**,** 947-56.

BRENNAN, J., LU, C. C., NORRIS, D. P., RODRIGUEZ, T. A., BEDDINGTON, R. S. & ROBERTSON, E. J. 2001. Nodal signalling in the epiblast patterns the early mouse embryo. *Nature,* 411**,** 965-9.

BRONS, I. G., SMITHERS, L. E., TROTTER, M. W., RUGG-GUNN, P., SUN, B., CHUVA DE SOUSA LOPES, S. M., HOWLETT, S. K., CLARKSON, A., AHRLUND-RICHTER, L., PEDERSEN, R. A. & VALLIER, L. 2007. Derivation of pluripotent epiblast stem cells from mammalian embryos. *Nature,* 448**,** 191-5.

BROWN, S., TEO, A., PAUKLIN, S., HANNAN, N., CHO, C. H., LIM, B., VARDY, L., DUNN, N. R., TROTTER, M., PEDERSEN, R. & VALLIER, L. 2011. Activin/Nodal signaling controls divergent transcriptional networks in human embryonic stem cells and in endoderm progenitors. *Stem Cells,* 29**,** 1176-85.

BRUCE, D. L. & SAPKOTA, G. P. 2012. Phosphatases in SMAD regulation. *FEBS Lett,* 586**,** 1897-905.

BUECKER, C. & WYSOCKA, J. 2012. Enhancers as information integration hubs in development: lessons from genomics. *Trends Genet,* 28**,** 276-84.

BUENROSTRO, J. D., GIRESI, P. G., ZABA, L. C., CHANG, H. Y. & GREENLEAF, W. J. 2013. Transposition of native chromatin for fast and sensitive epigenomic profiling of open chromatin, DNA-binding proteins and nucleosome position. *Nat Methods,* 10**,** 1213-8.

BUENROSTRO, J. D., WU, B., CHANG, H. Y. & GREENLEAF, W. J. 2015. ATAC-seq: A Method for Assaying Chromatin Accessibility Genome-Wide. *Curr Protoc Mol Biol,* 109**,** 21 29 1-9.

CAJA, L., KAHATA, K. & MOUSTAKAS, A. 2012. Context-dependent action of transforming growth factor beta family members on normal and cancer stem cells. *Curr Pharm Des,* 18**,** 4072-86.

CALO, E. & WYSOCKA, J. 2013. Modification of enhancer chromatin: what, how, and why? *Mol Cell,* 49**,** 825-37.

CHACKO, B. M., QIN, B. Y., TIWARI, A., SHI, G., LAM, S., HAYWARD, L. J., DE CAESTECKER, M. & LIN, K. 2004. Structural basis of heteromeric smad protein assembly in TGF-beta signaling. *Mol Cell,* 15**,** 813-23.

CHAN, S. S. & KYBA, M. 2013. What is a Master Regulator? *J Stem Cell Res Ther,* 3.

CHARNEY, R. M., FOROUZMAND, E., CHO, J. S., CHEUNG, J., PARAISO, K. D., YASUOKA, Y., TAKAHASHI, S., TAIRA, M., BLITZ, I. L., XIE, X. & CHO, K. W. 2017. Foxh1 Occupies cis-Regulatory Modules Prior to Dynamic Transcription Factor Interactions Controlling the Mesendoderm Gene Program. *Dev Cell,* 40**,** 595-607 e4.

CHEN, J., ZHANG, Z., LI, L., CHEN, B. C., REVYAKIN, A., HAJJ, B., LEGANT, W., DAHAN, M., LIONNET, T., BETZIG, E., TJIAN, R. & LIU, Z. 2014. Single-molecule dynamics of enhanceosome assembly in embryonic stem cells. *Cell,* 156**,** 1274-1285.

CHEN, X., RUBOCK, M. J. & WHITMAN, M. 1996. A transcriptional partner for MAD proteins in TGF-beta signalling. *Nature,* 383**,** 691-6.

CHEN, Y. G., HATA, A., LO, R. S., WOTTON, D., SHI, Y., PAVLETICH, N. & MASSAGUE, J. 1998. Determinants of specificity in TGF-beta signal transduction. *Genes Dev,* 12**,** 2144-52.

CHNG, Z., TEO, A., PEDERSEN, R. A. & VALLIER, L. 2010. SIP1 mediates cell-fate decisions between neuroectoderm and mesendoderm in human pluripotent stem cells. *Cell Stem Cell,* 6**,** 59-70.

CIABRELLI, F. & CAVALLI, G. 2015. Chromatin-driven behavior of topologically associating domains. *J Mol Biol,* 427**,** 608-25.

CODA, D. M., GAARENSTROOM, T., EAST, P., PATEL, H., MILLER, D. S., LOBLEY, A., MATTHEWS, N., STEWART, A. & HILL, C. S. 2017. Distinct modes of SMAD2 chromatin binding and remodeling shape the transcriptional response to NODAL/Activin signaling. *Elife,* 6.

COHEN, M., BRISCOE, J. & BLASSBERG, R. 2013. Morphogen interpretation: the transcriptional logic of neural tube patterning. *Curr Opin Genet Dev,* 23**,** 423-8.

COMPE, E. & EGLY, J. M. 2012. TFIIH: when transcription met DNA repair. *Nat Rev Mol Cell Biol,* 13**,** 343-54.

CRAMER, P. 2004. RNA polymerase II structure: from core to functional complexes. *Curr Opin Genet Dev,* 14**,** 218-26.

CREYGHTON, M. P., CHENG, A. W., WELSTEAD, G. G., KOOISTRA, T., CAREY, B. W., STEINE, E. J., HANNA, J., LODATO, M. A., FRAMPTON, G. M., SHARP, P. A., BOYER, L. A., YOUNG, R. A. & JAENISCH, R. 2010. Histone H3K27ac separates active from poised enhancers and predicts developmental state. *Proc Natl Acad Sci U S A,* 107**,** 21931-6.

DAHLE, O., KUMAR, A. & KUEHN, M. R. 2010. Nodal signaling recruits the histone demethylase Jmjd3 to counteract polycomb-mediated repression at target genes. *Sci Signal,* 3**,** ra48.

DEHEUNINCK, J. & LUO, K. 2009. Ski and SnoN, potent negative regulators of TGF-beta signaling. *Cell Res,* 19**,** 47-57.

DEKKER, J., MARTI-RENOM, M. A. & MIRNY, L. A. 2013. Exploring the three-dimensional organization of genomes: interpreting chromatin interaction data. *Nat Rev Genet,* 14**,** 390-403.

DENNY, S. K., YANG, D., CHUANG, C. H., BRADY, J. J., LIM, J. S., GRUNER, B. M., CHIOU, S. H., SCHEP, A. N., BARAL, J., HAMARD, C., ANTOINE, M., WISLEZ, M., KONG, C. S., CONNOLLY, A. J., PARK, K. S., SAGE, J., GREENLEAF, W. J. & WINSLOW, M. M. 2016. Nfib Promotes Metastasis through a Widespread Increase in Chromatin Accessibility. *Cell,* 166**,** 328-342.

DESCOSTES, N., HEIDEMANN, M., SPINELLI, L., SCHULLER, R., MAQBOOL, M. A., FENOUIL, R., KOCH, F., INNOCENTI, C., GUT, M., GUT, I., EICK, D. & ANDRAU, J. C. 2014. Tyrosine phosphorylation of RNA polymerase II CTD is associated with antisense promoter transcription and active enhancers in mammalian cells. *Elife,* 3**,** e02105.

DI GUGLIELMO, G. M., LE ROY, C., GOODFELLOW, A. F. & WRANA, J. L. 2003. Distinct endocytic pathways regulate TGF-beta receptor signalling and turnover. *Nat Cell Biol,* 5**,** 410-21.

DIXON, J. R., SELVARAJ, S., YUE, F., KIM, A., LI, Y., SHEN, Y., HU, M., LIU, J. S. & REN, B. 2012. Topological domains in mammalian genomes identified by analysis of chromatin interactions. *Nature,* 485**,** 376-80.

DONOVAN, P., DUBEY, O. A., KALLIOINEN, S., ROGERS, K. W., MUEHLETHALER, K., MULLER, P., RIMOLDI, D. & CONSTAM, D. B. 2017. Paracrine Activin-A signaling promotes melanoma growth and metastasis through immune evasion. *J Invest Dermatol*.

EGLOFF, S., DIENSTBIER, M. & MURPHY, S. 2012. Updating the RNA polymerase CTD code: adding gene-specific layers. *Trends Genet,* 28**,** 333-41.

EHRENSBERGER, A. H., KELLY, G. P. & SVEJSTRUP, J. Q. 2013. Mechanistic interpretation of promoter-proximal peaks and RNAPII density maps. *Cell,* 154**,** 713-5.

ESNAULT, C., STEWART, A., GUALDRINI, F., EAST, P., HORSWELL, S., MATTHEWS, N. & TREISMAN, R. 2014. Rho-actin signaling to the MRTF coactivators dominates the immediate transcriptional response to serum in fibroblasts. *Genes Dev,* 28**,** 943-58.

ESTARAS, C., AKIZU, N., GARCIA, A., BELTRAN, S., DE LA CRUZ, X. & MARTINEZ-BALBAS, M. A. 2012. Genome-wide analysis reveals that Smad3 and JMJD3 HDM co-activate the neural developmental program. *Development,* 139**,** 2681-91.

ESTARAS, C., BENNER, C. & JONES, K. A. 2015. SMADs and YAP Compete to Control Elongation of beta-Catenin:LEF-1-Recruited RNAPII during hESC Differentiation. *Mol Cell,* 58**,** 780-793.

FAIAL, T., BERNARDO, A. S., MENDJAN, S., DIAMANTI, E., ORTMANN, D., GENTSCH, G. E., MASCETTI, V. L., TROTTER, M. W., SMITH, J. C. & PEDERSEN, R. A. 2015. Brachyury and SMAD signalling collaboratively orchestrate distinct mesoderm and endoderm gene regulatory networks in differentiating human embryonic stem cells. *Development,* 142**,** 2121-35.

FEI, T., XIA, K., LI, Z., ZHOU, B., ZHU, S., CHEN, H., ZHANG, J., CHEN, Z., XIAO, H., HAN, J. D. & CHEN, Y. G. 2010. Genome-wide mapping of SMAD target genes reveals the role of BMP signaling in embryonic stem cell fate determination. *Genome Res,* 20**,** 36-44.

FENG, X. H., ZHANG, Y., WU, R. Y. & DERYNCK, R. 1998. The tumor suppressor Smad4/DPC4 and transcriptional adaptor CBP/p300 are coactivators for smad3 in TGF-beta-induced transcriptional activation. *Genes Dev,* 12**,** 2153-63.

FISHER, C. L. & FISHER, A. G. 2011. Chromatin states in pluripotent, differentiated, and reprogrammed cells. *Curr Opin Genet Dev,* 21**,** 140-6.

FLAVAHAN, W. A., DRIER, Y., LIAU, B. B., GILLESPIE, S. M., VENTEICHER, A. S., STEMMER-RACHAMIMOV, A. O., SUVA, M. L. & BERNSTEIN, B. E. 2016. Insulator dysfunction and oncogene activation in IDH mutant gliomas. *Nature,* 529**,** 110-4.

FUDA, N. J., ARDEHALI, M. B. & LIS, J. T. 2009. Defining mechanisms that regulate RNA polymerase II transcription in vivo. *Nature,* 461**,** 186-92.

FUDA, N. J. & LIS, J. T. 2013. A new player in Pol II pausing. *EMBO J,* 32**,** 1796-8.

FUKAYA, T., LIM, B. & LEVINE, M. 2016. Enhancer Control of Transcriptional Bursting. *Cell,* 166**,** 358-368.

GAARENSTROOM, T. & HILL, C. S. 2014. TGF-beta signaling to chromatin: how Smads regulate transcription during self-renewal and differentiation. *Semin Cell Dev Biol,* 32**,** 107-18.

GALVIN, K. E., TRAVIS, E. D., YEE, D., MAGNUSON, T. & VIVIAN, J. L. 2010. Nodal signaling regulates the bone morphogenic protein pluripotency pathway in mouse embryonic stem cells. *J Biol Chem,* 285**,** 19747-56.

GAO, S., ALARCON, C., SAPKOTA, G., RAHMAN, S., CHEN, P. Y., GOERNER, N., MACIAS, M. J., ERDJUMENT-BROMAGE, H., TEMPST, P. & MASSAGUE, J. 2009. Ubiquitin ligase Nedd4L targets activated Smad2/3 to limit TGF-beta signaling. *Mol Cell,* 36**,** 457-68.

GERMAIN, S., HOWELL, M., ESSLEMONT, G. M. & HILL, C. S. 2000. Homeodomain and winged-helix transcription factors recruit activated Smads to distinct promoter elements via a common Smad interaction motif. *Genes Dev,* 14**,** 435-51.

GOLDSTEIN, I. & HAGER, G. L. 2017. Dynamic enhancer function in the chromatin context. *Wiley Interdiscip Rev Syst Biol Med*.

GOUMANS, M. J., VALDIMARSDOTTIR, G., ITOH, S., LEBRIN, F., LARSSON, J., MUMMERY, C., KARLSSON, S. & TEN DIJKE, P. 2003. Activin receptor-like kinase (ALK)1 is an antagonistic mediator of lateral TGFbeta/ALK5 signaling. *Mol Cell,* 12**,** 817-28.

GRAY, L. T., YAO, Z., NGUYEN, T. N., KIM, T. K., ZENG, H. & TASIC, B. 2017. Layer-specific chromatin accessibility landscapes reveal regulatory networks in adult mouse visual cortex. *Elife,* 6.

GRONROOS, E., KINGSTON, I. J., RAMACHANDRAN, A., RANDALL, R. A., VIZAN, P. & HILL, C. S. 2012. Transforming growth factor beta inhibits bone morphogenetic protein-induced transcription through novel phosphorylated Smad1/5-Smad3 complexes. *Mol Cell Biol,* 32**,** 2904-16.

GUALDRINI, F., ESNAULT, C., HORSWELL, S., STEWART, A., MATTHEWS, N. & TREISMAN, R. 2016. SRF Co-factors Control the Balance between Cell Proliferation and Contractility. *Mol Cell,* 64**,** 1048-1061.

HAGOS, E. G. & DOUGAN, S. T. 2007. Time-dependent patterning of the mesoderm and endoderm by Nodal signals in zebrafish. *BMC Dev Biol,* 7**,** 22.

HARIKUMAR, A. & MESHORER, E. 2015. Chromatin remodeling and bivalent histone modifications in embryonic stem cells. *EMBO Rep,* 16**,** 1609-19.

HEINZ, S., BENNER, C., SPANN, N., BERTOLINO, E., LIN, Y. C., LASLO, P., CHENG, J. X., MURRE, C., SINGH, H. & GLASS, C. K. 2010. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol Cell,* 38**,** 576-89.

HELDIN, C. H. & MOUSTAKAS, A. 2016. Signaling Receptors for TGF-beta Family Members. *Cold Spring Harb Perspect Biol,* 8.

HENIKOFF, S. & SHILATIFARD, A. 2011. Histone modification: cause or cog? *Trends Genet,* 27**,** 389-96.

HILL, C. S. 2016. Transcriptional Control by the SMADs. *Cold Spring Harb Perspect Biol,* 8.

HNISZ, D., ABRAHAM, B. J., LEE, T. I., LAU, A., SAINT-ANDRE, V., SIGOVA, A. A., HOKE, H. A. & YOUNG, R. A. 2013. Super-enhancers in the control of cell identity and disease. *Cell,* 155**,** 934-47.

HNISZ, D., SHRINIVAS, K., YOUNG, R. A., CHAKRABORTY, A. K. & SHARP, P. A. 2017. A Phase Separation Model for Transcriptional Control. *Cell,* 169**,** 13-23.

HOWELL, M., INMAN, G. J. & HILL, C. S. 2002. A novel Xenopus Smad-interacting forkhead transcription factor (XFast-3) cooperates with XFast-1 in regulating gastrulation movements. *Development,* 129**,** 2823-34.

HUMINIECKI, L., GOLDOVSKY, L., FREILICH, S., MOUSTAKAS, A., OUZOUNIS, C. & HELDIN, C. H. 2009. Emergence, development and diversification of the TGF-beta signalling pathway within the animal kingdom. *BMC Evol Biol,* 9**,** 28.

INMAN, G. J. & HILL, C. S. 2002. Stoichiometry of active smad-transcription factor complexes on DNA. *J Biol Chem,* 277**,** 51008-16.

ITOH, F., WATABE, T. & MIYAZONO, K. 2014. Roles of TGF-beta family signals in the fate determination of pluripotent stem cells. *Semin Cell Dev Biol,* 32**,** 98-106.

ITOH, S. & TEN DIJKE, P. 2007. Negative regulation of TGF-beta receptor/Smad signal transduction. *Curr Opin Cell Biol,* 19**,** 176-84.

JANKNECHT, R., WELLS, N. J. & HUNTER, T. 1998. TGF-beta-stimulated cooperation of smad proteins with the coactivators CBP/p300. *Genes Dev,* 12**,** 2114-9.

JI, X., DADON, D. B., POWELL, B. E., FAN, Z. P., BORGES-RIVERA, D., SHACHAR, S., WEINTRAUB, A. S., HNISZ, D., PEGORARO, G., LEE, T. I., MISTELI, T., JAENISCH, R. & YOUNG, R. A. 2016. 3D Chromosome Regulatory Landscape of Human Pluripotent Cells. *Cell Stem Cell,* 18**,** 262-75.

JONKERS, I., KWAK, H. & LIS, J. T. 2014. Genome-wide dynamics of Pol II elongation and its interplay with promoter proximal pausing, chromatin, and exons. *Elife,* 3**,** e02407.

JUVEN-GERSHON, T. & KADONAGA, J. T. 2010. Regulation of gene expression via the core promoter and the basal transcriptional machinery. *Dev Biol,* 339**,** 225-9.

KANG, Y., CHEN, C. R. & MASSAGUE, J. 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**,** 915-26.

KARMODIYA, K., KREBS, A. R., OULAD-ABDELGHANI, M., KIMURA, H. & TORA, L. 2012. H3K9 and H3K14 acetylation co-occur at many gene regulatory elements, while H3K14ac marks a subset of inactive inducible promoters in mouse embryonic stem cells. *BMC Genomics,* 13**,** 424.

KAROLCHIK, D., HINRICHS, A. S., FUREY, T. S., ROSKIN, K. M., SUGNET, C. W., HAUSSLER, D. & KENT, W. J. 2004. The UCSC Table Browser data retrieval tool. *Nucleic Acids Res,* 32**,** D493-6.

KATO, Y., HABAS, R., KATSUYAMA, Y., NAAR, A. M. & HE, X. 2002. A component of the ARC/Mediator complex required for TGF beta/Nodal signalling. *Nature,* 418**,** 641-6.

KENT, W. J., ZWEIG, A. S., BARBER, G., HINRICHS, A. S. & KAROLCHIK, D. 2010. BigWig and BigBed: enabling browsing of large distributed datasets. *Bioinformatics,* 26**,** 2204-7.

KIM, S. W., YOON, S. J., CHUONG, E., OYOLU, C., WILLS, A. E., GUPTA, R. & BAKER, J. 2011. Chromatin and transcriptional signatures for Nodal signaling during endoderm formation in hESCs. *Dev Biol,* 357**,** 492-504.

KOBA, M. & KONOPA, J. 2005. [Actinomycin D and its mechanisms of action]. *Postepy Hig Med Dosw (Online),* 59**,** 290-8.

KOUZARIDES, T. 2007. Chromatin modifications and their function. *Cell,* 128**,** 693-705.

KUMAR, A., NOVOSELOV, V., CELESTE, A. J., WOLFMAN, N. M., TEN DIJKE, P. & KUEHN, M. R. 2001. Nodal signaling uses activin and transforming growth factor-beta receptor-regulated Smads. *J Biol Chem,* 276**,** 656-61.

KUNWAR, P. S., ZIMMERMAN, S., BENNETT, J. T., CHEN, Y., WHITMAN, M. & SCHIER, A. F. 2003. Mixer/Bon and FoxH1/Sur have overlapping and divergent roles in Nodal signaling and mesendoderm induction. *Development,* 130**,** 5589-99.

KUROKAWA, M., MITANI, K., IRIE, K., MATSUYAMA, T., TAKAHASHI, T., CHIBA, S., YAZAKI, Y., MATSUMOTO, K. & HIRAI, H. 1998. The oncoprotein Evi-1 represses TGF-beta signalling by inhibiting Smad3. *Nature,* 394**,** 92-6.

LABBE, E., SILVESTRI, C., HOODLESS, P. A., WRANA, J. L. & ATTISANO, L. 1998. Smad2 and Smad3 positively and negatively regulate TGF beta-dependent transcription through the forkhead DNA-binding protein FAST2. *Mol Cell,* 2**,** 109-20.

LANDRY, J., SHAROV, A. A., PIAO, Y., SHAROVA, L. V., XIAO, H., SOUTHON, E., MATTA, J., TESSAROLLO, L., ZHANG, Y. E., KO, M. S., KUEHN, M. R., YAMAGUCHI, T. P. & WU, C. 2008. Essential role of chromatin remodeling protein Bptf in early mouse embryos and embryonic stem cells. *PLoS Genet,* 4**,** e1000241.

LANGE, M., DEMAJO, S., JAIN, P. & DI CROCE, L. 2011. Combinatorial assembly and function of chromatin regulatory complexes. *Epigenomics,* 3**,** 567-80.

LEE, K. L., LIM, S. K., ORLOV, Y. L., YIT LE, Y., YANG, H., ANG, L. T., POELLINGER, L. & LIM, B. 2011. Graded Nodal/Activin signaling titrates conversion of quantitative phospho-Smad2 levels into qualitative embryonic stem cell fate decisions. *PLoS Genet,* 7**,** e1002130.

LEE, T. I. & YOUNG, R. A. 2013. Transcriptional regulation and its misregulation in disease. *Cell,* 152**,** 1237-51.

LEVINE, M. 2011. Paused RNA polymerase II as a developmental checkpoint. *Cell,* 145**,** 502-11.

LEVY, L. & HILL, C. S. 2005. Smad4 dependency defines two classes of transforming growth factor b (TGF-b) target genes and distinguishes TGF-b-induced epithelial-mesenchymal transition from its antiproliferative and migratory responses. *Mol Cell Biol,* 25**,** 8108-25.

LEVY, L., HOWELL, M., DAS, D., HARKIN, S., EPISKOPOU, V. & HILL, C. S. 2007. Arkadia activates Smad3/Smad4-dependent transcription by triggering signal-induced SnoN degradation. *Mol Cell Biol,* 27**,** 6068-83.

LI, B., CAREY, M. & WORKMAN, J. L. 2007. The role of chromatin during transcription. *Cell,* 128**,** 707-19.

LI, H. & DURBIN, R. 2009. Fast and accurate short read alignment with Burrows-Wheeler transform. *Bioinformatics,* 25**,** 1754-60.

LI, H., HANDSAKER, B., WYSOKER, A., FENNELL, T., RUAN, J., HOMER, N., MARTH, G., ABECASIS, G., DURBIN, R. & GENOME PROJECT DATA PROCESSING, S. 2009. The Sequence Alignment/Map format and SAMtools. *Bioinformatics,* 25**,** 2078-9.

LIAO, Y., SMYTH, G. K. & SHI, W. 2014. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics,* 30**,** 923-30.

LIU, D., BLACK, B. L. & DERYNCK, R. 2001. TGF-beta inhibits muscle differentiation through functional repression of myogenic transcription factors by Smad3. *Genes Dev,* 15**,** 2950-66.

LIU, X., KRAUS, W. L. & BAI, X. 2015. Ready, pause, go: regulation of RNA polymerase II pausing and release by cellular signaling pathways. *Trends Biochem Sci,* 40**,** 516-25.

LONARDO, E., HERMANN, P. C., MUELLER, M. T., HUBER, S., BALIC, A., MIRANDA-LORENZO, I., ZAGORAC, S., ALCALA, S., RODRIGUEZ-ARABAOLAZA, I., RAMIREZ, J. C., TORRES-RUIZ, R., GARCIA, E., HIDALGO, M., CEBRIAN, D. A., HEUCHEL, R., LOHR, M., BERGER, F., BARTENSTEIN, P., AICHER, A. & HEESCHEN, C. 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**,** 433-46.

LONG, H. K., PRESCOTT, S. L. & WYSOCKA, J. 2016. Ever-Changing Landscapes: Transcriptional Enhancers in Development and Evolution. *Cell,* 167**,** 1170-1187.

LU, C. C., BRENNAN, J. & ROBERTSON, E. J. 2001. From fertilization to gastrulation: axis formation in the mouse embryo. *Curr Opin Genet Dev,* 11**,** 384-92.

LUO, K. 2004. Ski and SnoN: negative regulators of TGF-beta signaling. *Curr Opin Genet Dev,* 14**,** 65-70.

LUPIANEZ, D. G., KRAFT, K., HEINRICH, V., KRAWITZ, P., BRANCATI, F., KLOPOCKI, E., HORN, D., KAYSERILI, H., OPITZ, J. M., LAXOVA, R., SANTOS-SIMARRO, F., GILBERT-DUSSARDIER, B., WITTLER, L., BORSCHIWER, M., HAAS, S. A., OSTERWALDER, M., FRANKE, M., TIMMERMANN, B., HECHT, J., SPIELMANN, M., VISEL, A. & MUNDLOS, S. 2015. Disruptions of topological chromatin domains cause pathogenic rewiring of gene-enhancer interactions. *Cell,* 161**,** 1012-25.

LUPIANEZ, D. G., SPIELMANN, M. & MUNDLOS, S. 2016. Breaking TADs: How Alterations of Chromatin Domains Result in Disease. *Trends Genet,* 32**,** 225-37.

MACIAS, M. J., MARTIN-MALPARTIDA, P. & MASSAGUE, J. 2015. Structural determinants of Smad function in TGF-beta signaling. *Trends Biochem Sci,* 40**,** 296-308.

MARIKAWA, Y., TAMASHIRO, D. A., FUJITA, T. C. & ALARCON, V. B. 2011. Dual roles of Oct4 in the maintenance of mouse P19 embryonal carcinoma cells: as negative regulator of Wnt/beta-catenin signaling and competence provider for Brachyury induction. *Stem Cells Dev,* 20**,** 621-33.

MASSAGUE, J. 2008. TGFbeta in Cancer. *Cell,* 134**,** 215-30.

MASSAGUE, J. 2012. TGFbeta signalling in context. *Nat Rev Mol Cell Biol,* 13**,** 616-30.

MEINHART, A., KAMENSKI, T., HOEPPNER, S., BAUMLI, S. & CRAMER, P. 2005. A structural perspective of CTD function. *Genes Dev,* 19**,** 1401-15.

MILLER, D. S. H., C.S 2016. TGF-b superfamily signalling. *Encyclopedia of Cell Biology*.

MITCHELL, H., CHOUDHURY, A., PAGANO, R. E. & LEOF, E. B. 2004. Ligand-dependent and -independent transforming growth factor-beta receptor recycling regulated by clathrin-mediated endocytosis and Rab11. *Mol Biol Cell,* 15**,** 4166-78.

MIZUTANI, A., KOINUMA, D., TSUTSUMI, S., KAMIMURA, N., MORIKAWA, M., SUZUKI, H. I., IMAMURA, T., MIYAZONO, K. & ABURATANI, H. 2011. Cell type-specific target selection by combinatorial binding of Smad2/3 proteins and hepatocyte nuclear factor 4alpha in HepG2 cells. *J Biol Chem,* 286**,** 29848-60.

MORIKAWA, M., KOINUMA, D., TSUTSUMI, S., VASILAKI, E., KANKI, Y., HELDIN, C. H., ABURATANI, H. & MIYAZONO, K. 2011. ChIP-seq reveals cell type-specific binding patterns of BMP-specific Smads and a novel binding motif. *Nucleic Acids Res,* 39**,** 8712-27.

MOUSAVI, K., ZARE, H., DELL'ORSO, S., GRONTVED, L., GUTIERREZ-CRUZ, G., DERFOUL, A., HAGER, G. L. & SARTORELLI, V. 2013. eRNAs promote transcription by establishing chromatin accessibility at defined genomic loci. *Mol Cell,* 51**,** 606-17.

MOUSTAKAS, A. & HELDIN, C. H. 2009. The regulation of TGFbeta signal transduction. *Development,* 136**,** 3699-714.

MUELLER, T. D. & NICKEL, J. 2012. Promiscuity and specificity in BMP receptor activation. *FEBS Lett,* 586**,** 1846-59.

MULLEN, A. C., ORLANDO, D. A., NEWMAN, J. J., LOVEN, J., KUMAR, R. M., BILODEAU, S., REDDY, J., GUENTHER, M. G., DEKOTER, R. P. & YOUNG, R. A. 2011. Master transcription factors determine cell-type-specific responses to TGF-beta signaling. *Cell,* 147**,** 565-76.

NAHMAD, M. & LANDER, A. D. 2011. Spatiotemporal mechanisms of morphogen gradient interpretation. *Curr Opin Genet Dev,* 21**,** 726-31.

NAKAYA, K., MURAKAMI, M. & FUNABA, M. 2008. Regulatory expression of Brachyury and Goosecoid in P19 embryonal carcinoma cells. *J Cell Biochem,* 105**,** 801-13.

NECHAEV, S. & ADELMAN, K. 2011. Pol II waiting in the starting gates: Regulating the transition from transcription initiation into productive elongation. *Biochim Biophys Acta,* 1809**,** 34-45.

NECHAEV, S., FARGO, D. C., DOS SANTOS, G., LIU, L., GAO, Y. & ADELMAN, K. 2010. Global analysis of short RNAs reveals widespread promoter-proximal stalling and arrest of Pol II in Drosophila. *Science,* 327**,** 335-8.

NORA, E. P., LAJOIE, B. R., SCHULZ, E. G., GIORGETTI, L., OKAMOTO, I., SERVANT, N., PIOLOT, T., VAN BERKUM, N. L., MEISIG, J., SEDAT, J., GRIBNAU, J., BARILLOT, E., BLUTHGEN, N., DEKKER, J. & HEARD, E. 2012. Spatial partitioning of the regulatory landscape of the X-inactivation centre. *Nature,* 485**,** 381-5.

NORRIS, D. P. & ROBERTSON, E. J. 1999. Asymmetric and node-specific nodal expression patterns are controlled by two distinct cis-acting regulatory elements. *Genes Dev,* 13**,** 1575-88.

OHLER, U., LIAO, G. C., NIEMANN, H. & RUBIN, G. M. 2002. Computational analysis of core promoters in the Drosophila genome. *Genome Biol,* 3**,** RESEARCH0087.

OROM, U. A. & SHIEKHATTAR, R. 2011. Noncoding RNAs and enhancers: complications of a long-distance relationship. *Trends Genet,* 27**,** 433-9.

OSTUNI, R., PICCOLO, V., BAROZZI, I., POLLETTI, S., TERMANINI, A., BONIFACIO, S., CURINA, A., PROSPERINI, E., GHISLETTI, S. & NATOLI, G. 2013. Latent enhancers activated by stimulation in differentiated cells. *Cell,* 152**,** 157-71.

PAPAGEORGIOU, I., NICHOLLS, P. K., WANG, F., LACKMANN, M., MAKANJI, Y., SALAMONSEN, L. A., ROBERTSON, D. M. & HARRISON, C. A. 2009. Expression of nodal signalling components in cycling human endometrium and in endometrial cancer. *Reprod Biol Endocrinol,* 7**,** 122.

PAUKLIN, S. & VALLIER, L. 2015. Activin/Nodal signalling in stem cells. *Development,* 142**,** 607-19.

PENG, J., LI, Q., WIGGLESWORTH, K., RANGARAJAN, A., KATTAMURI, C., PETERSON, R. T., EPPIG, J. J., THOMPSON, T. B. & MATZUK, M. M. 2013. Growth differentiation factor 9:bone morphogenetic protein 15 heterodimers are potent regulators of ovarian functions. *Proc Natl Acad Sci U S A,* 110**,** E776-85.

PERRIMON, N., PITSOULI, C. & SHILO, B. Z. 2012. Signaling mechanisms controlling cell fate and embryonic patterning. *Cold Spring Harb Perspect Biol,* 4**,** a005975.

PHILLIPS, J. E. & CORCES, V. G. 2009. CTCF: master weaver of the genome. *Cell,* 137**,** 1194-211.

PIPER, J., ELZE, M. C., CAUCHY, P., COCKERILL, P. N., BONIFER, C. & OTT, S. 2013. Wellington: a novel method for the accurate identification of digital genomic footprints from DNase-seq data. *Nucleic Acids Res,* 41**,** e201.

POSTIGO, A. A. 2003. Opposing functions of ZEB proteins in the regulation of the TGFbeta/BMP signaling pathway. *EMBO J,* 22**,** 2443-52.

QUAIL, D. F., SIEGERS, G. M., JEWER, M. & POSTOVIT, L. M. 2013. Nodal signalling in embryogenesis and tumourigenesis. *Int J Biochem Cell Biol,* 45**,** 885-98.

QUINLAN, A. R. & HALL, I. M. 2010. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics,* 26**,** 841-2.

RADA-IGLESIAS, A., BAJPAI, R., SWIGUT, T., BRUGMANN, S. A., FLYNN, R. A. & WYSOCKA, J. 2011. A unique chromatin signature uncovers early developmental enhancers in humans. *Nature,* 470**,** 279-83.

RAFIEE, M. R., GIRARDOT, C., SIGISMONDO, G. & KRIJGSVELD, J. 2016. Expanding the Circuitry of Pluripotency by Selective Isolation of Chromatin-Associated Proteins. *Mol Cell,* 64**,** 624-635.

RAHL, P. B., LIN, C. Y., SEILA, A. C., FLYNN, R. A., MCCUINE, S., BURGE, C. B., SHARP, P. A. & YOUNG, R. A. 2010. c-Myc regulates transcriptional pause release. *Cell,* 141**,** 432-45.

RAJAGOPAL, N., ERNST, J., RAY, P., WU, J., ZHANG, M., KELLIS, M. & REN, B. 2014. Distinct and predictive histone lysine acetylation patterns at promoters, enhancers, and gene bodies. *G3 (Bethesda),* 4**,** 2051-63.

RAN, F. A., HSU, P. D., WRIGHT, J., AGARWALA, V., SCOTT, D. A. & ZHANG, F. 2013. Genome engineering using the CRISPR-Cas9 system. *Nat Protoc,* 8**,** 2281-308.

RANDALL, R. A., GERMAIN, S., INMAN, G. J., BATES, P. A. & HILL, C. S. 2002. Different Smad2 partners bind a common hydrophobic pocket in Smad2 via a defined proline-rich motif. *EMBO J,* 21**,** 145-56.

RANDALL, R. A., HOWELL, M., PAGE, C. S., DALY, A., BATES, P. A. & HILL, C. S. 2004. Recognition of phosphorylated-Smad2-containing complexes by a novel Smad interaction motif. *Mol Cell Biol,* 24**,** 1106-21.

REITER, F., WIENERROITHER, S. & STARK, A. 2017. Combinatorial function of transcription factors and cofactors. *Curr Opin Genet Dev,* 43**,** 73-81.

RENDEIRO, A. F., SCHMIDL, C., STREFFORD, J. C., WALEWSKA, R., DAVIS, Z., FARLIK, M., OSCIER, D. & BOCK, C. 2016. Chromatin accessibility maps of chronic lymphocytic leukaemia identify subtype-specific epigenome signatures and transcription regulatory networks. *Nat Commun,* 7**,** 11938.

ROBERTSON, I. B. & RIFKIN, D. B. 2016. Regulation of the Bioavailability of TGF-beta and TGF-beta-Related Proteins. *Cold Spring Harb Perspect Biol,* 8.

ROSS, S., CHEUNG, E., PETRAKIS, T. G., HOWELL, M., KRAUS, W. L. & HILL, C. S. 2006. Smads orchestrate specific histone modifications and chromatin remodeling to activate transcription. *EMBO J,* 25**,** 4490-502.

ROSS, S. & HILL, C. S. 2008. How the Smads regulate transcription. *Int J Biochem Cell Biol,* 40**,** 383-408.

ROSSANT, J. 2015. Mouse and human blastocyst-derived stem cells: vive les differences. *Development,* 142**,** 9-12.

SAKAKI-YUMOTO, M., LIU, J., RAMALHO-SANTOS, M., YOSHIDA, N. & DERYNCK, R. 2013. Smad2 is essential for maintenance of the human and mouse primed pluripotent stem cell state. *J Biol Chem,* 288**,** 18546-60.

SAPKOTA, G., ALARCON, C., SPAGNOLI, F. M., BRIVANLOU, A. H. & MASSAGUE, J. 2007. Balancing BMP signaling through integrated inputs into the Smad1 linker. *Mol Cell,* 25**,** 441-54.

SCHIER, A. F. 2009. Nodal morphogens. *Cold Spring Harb Perspect Biol,* 1**,** a003459.

SCHMIERER, B. & HILL, C. S. 2007. TGFbeta-SMAD signal transduction: molecular specificity and functional flexibility. *Nat Rev Mol Cell Biol,* 8**,** 970-82.

SCHMIERER, B., TOURNIER, A. L., BATES, P. A. & HILL, C. S. 2008. Mathematical modeling identifies Smad nucleocytoplasmic shuttling as a dynamic signal-interpreting system. *Proc Natl Acad Sci U S A,* 105**,** 6608-13.

SCHMITT, A. D., HU, M. & REN, B. 2016. Genome-wide mapping and analysis of chromosome architecture. *Nat Rev Mol Cell Biol,* 17**,** 743-755.

SCHNERCH, A., CERDAN, C. & BHATIA, M. 2010. Distinguishing between mouse and human pluripotent stem cell regulation: the best laid plans of mice and men. *Stem Cells,* 28**,** 419-30.

SHEN, L., SHAO, N. Y., LIU, X., MAZE, I., FENG, J. & NESTLER, E. J. 2013. diffReps: detecting differential chromatin modification sites from ChIP-seq data with biological replicates. *PLoS One,* 8**,** e65598.

SHEN, M. M. 2007. Nodal signaling: developmental roles and regulation. *Development,* 134**,** 1023-34.

SHI, Y. & MASSAGUE, J. 2003. Mechanisms of TGF-beta signaling from cell membrane to the nucleus. *Cell,* 113**,** 685-700.

SHI, Y., WANG, Y. F., JAYARAMAN, L., YANG, H., MASSAGUE, J. & PAVLETICH, N. P. 1998. Crystal structure of a Smad MH1 domain bound to DNA: insights on DNA binding in TGF-beta signaling. *Cell,* 94**,** 585-94.

SILVESTRI, C., NARIMATSU, M., VON BOTH, I., LIU, Y., TAN, N. B., IZZI, L., MCCAFFERY, P., WRANA, J. L. & ATTISANO, L. 2008. Genome-wide identification of Smad/Foxh1 targets reveals a role for Foxh1 in retinoic acid regulation and forebrain development. *Dev Cell,* 14**,** 411-23.

SIMON, C. S., DOWNES, D. J., GOSDEN, M. E., TELENIUS, J., HIGGS, D. R., HUGHES, J. R., COSTELLO, I., BIKOFF, E. K. & ROBERTSON, E. J. 2017. Functional characterisation of cis-regulatory elements governing dynamic Eomes expression in the early mouse embryo. *Development,* 144**,** 1249-1260.

SIMON, J. M., GIRESI, P. G., DAVIS, I. J. & LIEB, J. D. 2012. Using formaldehyde-assisted isolation of regulatory elements (FAIRE) to isolate active regulatory DNA. *Nat Protoc,* 7**,** 256-67.

SINGH, A. M., REYNOLDS, D., CLIFF, T., OHTSUKA, S., MATTHEYSES, A. L., SUN, Y., MENENDEZ, L., KULIK, M. & DALTON, S. 2012. Signaling network crosstalk in human pluripotent cells: a Smad2/3-regulated switch that controls the balance between self-renewal and differentiation. *Cell Stem Cell,* 10**,** 312-26.

SPITZ, F. & FURLONG, E. E. M. 2012. Transcription factors: from enhancer binding to developmental control. *Nat Rev Genet,* 13**,** 613-626.

STRIZZI, L., POSTOVIT, L. M., MARGARYAN, N. V., SEFTOR, E. A., ABBOTT, D. E., SEFTOR, R. E., SALOMON, D. S. & HENDRIX, M. J. 2008. Emerging roles of nodal and Cripto-1: from embryogenesis to breast cancer progression. *Breast Dis,* 29**,** 91-103.

SUI, L., BOUWENS, L. & MFOPOU, J. K. 2013. Signaling pathways during maintenance and definitive endoderm differentiation of embryonic stem cells. *Int J Dev Biol,* 57**,** 1-12.

SUNG, M. H., BAEK, S. & HAGER, G. L. 2016. Genome-wide footprinting: ready for prime time? *Nat Methods,* 13**,** 222-228.

SUNG, M. H., GUERTIN, M. J., BAEK, S. & HAGER, G. L. 2014. DNase footprint signatures are dictated by factor dynamics and DNA sequence. *Mol Cell,* 56**,** 275-85.

SWINSTEAD, E. E., MIRANDA, T. B., PAAKINAHO, V., BAEK, S., GOLDSTEIN, I., HAWKINS, M., KARPOVA, T. S., BALL, D., MAZZA, D., LAVIS, L. D., GRIMM, J. B., MORISAKI, T., GRONTVED, L., PRESMAN, D. M. & HAGER, G. L. 2016a. Steroid Receptors Reprogram FoxA1 Occupancy through Dynamic Chromatin Transitions. *Cell,* 165**,** 593-605.

SWINSTEAD, E. E., PAAKINAHO, V., PRESMAN, D. M. & HAGER, G. L. 2016b. Pioneer factors and ATP-dependent chromatin remodeling factors interact dynamically: A new perspective: Multiple transcription factors can effect chromatin pioneer functions through dynamic interactions with ATP-dependent chromatin remodeling factors. *Bioessays*.

TADA, H., KUBOKI, K., NOMURA, K. & INOKUCHI, T. 2001. High glucose levels enhance TGF-beta1-thrombospondin-1 pathway in cultured human mesangial cells via mechanisms dependent on glucose-induced PKC activation. *J Diabetes Complications,* 15**,** 193-7.

TAKAHASHI, H., PARMELY, T. J., SATO, S., TOMOMORI-SATO, C., BANKS, C. A., KONG, S. E., SZUTORISZ, H., SWANSON, S. K., MARTIN-BROWN, S., WASHBURN, M. P., FLORENS, L., SEIDEL, C. W., LIN, C., SMITH, E. R., SHILATIFARD, A., CONAWAY, R. C. & CONAWAY, J. W. 2011. Human mediator subunit MED26 functions as a docking site for transcription elongation factors. *Cell,* 146**,** 92-104.

TANAKA, C., SAKUMA, R., NAKAMURA, T., HAMADA, H. & SAIJOH, Y. 2007. Long-range action of Nodal requires interaction with GDF1. *Genes Dev,* 21**,** 3272-82.

TANG, L. Y., YAMASHITA, M., COUSSENS, N. P., TANG, Y., WANG, X., LI, C., DENG, C. X., CHENG, S. Y. & ZHANG, Y. E. 2011. Ablation of Smurf2 reveals an inhibition in TGF-beta signalling through multiple mono-ubiquitination of Smad3. *EMBO J,* 30**,** 4777-89.

TESAR, P. J., CHENOWETH, J. G., BROOK, F. A., DAVIES, T. J., EVANS, E. P., MACK, D. L., GARDNER, R. L. & MCKAY, R. D. 2007. New cell lines from mouse epiblast share defining features with human embryonic stem cells. *Nature,* 448**,** 196-9.

TILLO, D., KAPLAN, N., MOORE, I. K., FONDUFE-MITTENDORF, Y., GOSSETT, A. J., FIELD, Y., LIEB, J. D., WIDOM, J., SEGAL, E. & HUGHES, T. R. 2010. High nucleosome occupancy is encoded at human regulatory sequences. *PLoS One,* 5**,** e9129.

TOGASHI, Y., KOGITA, A., SAKAMOTO, H., HAYASHI, H., TERASHIMA, M., DE VELASCO, M. A., SAKAI, K., FUJITA, Y., TOMIDA, S., KITANO, M., OKUNO, K., KUDO, M. & NISHIO, K. 2015. Activin signal promotes cancer progression and is involved in cachexia in a subset of pancreatic cancer. *Cancer Lett,* 356**,** 819-27.

TOPCZEWSKA, J. M., POSTOVIT, L. M., MARGARYAN, N. V., SAM, A., HESS, A. R., WHEATON, W. W., NICKOLOFF, B. J., TOPCZEWSKI, J. & HENDRIX, M. J. 2006. Embryonic and tumorigenic pathways converge via Nodal signaling: role in melanoma aggressiveness. *Nat Med,* 12**,** 925-32.

TROMPOUKI, E., BOWMAN, T. V., LAWTON, L. N., FAN, Z. P., WU, D. C., DIBIASE, A., MARTIN, C. S., CECH, J. N., SESSA, A. K., LEBLANC, J. L., LI, P., DURAND, E. M., MOSIMANN, C., HEFFNER, G. C., DALEY, G. Q., PAULSON, R. F., YOUNG, R. A. & ZON, L. I. 2011. Lineage regulators direct BMP and Wnt pathways to cell-specific programs during differentiation and regeneration. *Cell,* 147**,** 577-89.

TSOMPANA, M. & BUCK, M. J. 2014. Chromatin accessibility: a window into the genome. *Epigenetics Chromatin,* 7**,** 33.

TSUKAZAKI, T., CHIANG, T. A., DAVISON, A. F., ATTISANO, L. & WRANA, J. L. 1998. SARA, a FYVE domain protein that recruits Smad2 to the TGFbeta receptor. *Cell,* 95**,** 779-91.

TSUNEYOSHI, N., TAN, E. K., SADASIVAM, A., POOBALAN, Y., SUMI, T., NAKATSUJI, N., SUEMORI, H. & DUNN, N. R. 2012. The SMAD2/3 corepressor SNON maintains pluripotency through selective repression of mesendodermal genes in human ES cells. *Genes Dev,* 26**,** 2471-6.

VALLIER, L., ALEXANDER, M. & PEDERSEN, R. A. 2005. Activin/Nodal and FGF pathways cooperate to maintain pluripotency of human embryonic stem cells. *J Cell Sci,* 118**,** 4495-509.

VALLIER, L., MENDJAN, S., BROWN, S., CHNG, Z., TEO, A., SMITHERS, L. E., TROTTER, M. W., CHO, C. H., MARTINEZ, A., RUGG-GUNN, P., BRONS, G. & PEDERSEN, R. A. 2009. Activin/Nodal signalling maintains pluripotency by controlling Nanog expression. *Development,* 136**,** 1339-49.

VALLIER, L., REYNOLDS, D. & PEDERSEN, R. A. 2004. Nodal inhibits differentiation of human embryonic stem cells along the neuroectodermal default pathway. *Dev Biol,* 275**,** 403-21.

VALTON, A. L. & DEKKER, J. 2016. TAD disruption as oncogenic driver. *Curr Opin Genet Dev,* 36**,** 34-40.

VAN BOXTEL, A. L., CHESEBRO, J. E., HELIOT, C., RAMEL, M. C., STONE, R. K. & HILL, C. S. 2015. A Temporal Window for Signal Activation Dictates the Dimensions of a Nodal Signaling Domain. *Dev Cell,* 35**,** 175-85.

VAN DER HEYDEN, M. A. & DEFIZE, L. H. 2003. Twenty one years of P19 cells: what an embryonal carcinoma cell line taught us about cardiomyocyte differentiation. *Cardiovasc Res,* 58**,** 292-302.

VAN DIJK, E. L., AUGER, H., JASZCZYSZYN, Y. & THERMES, C. 2014. Ten years of next-generation sequencing technology. *Trends Genet,* 30**,** 418-426.

VASTENHOUW, N. L. & SCHIER, A. F. 2012. Bivalent histone modifications in early embryogenesis. *Curr Opin Cell Biol,* 24**,** 374-86.

VIERSTRA, J. & STAMATOYANNOPOULOS, J. A. 2016. Genomic footprinting. *Nat Methods,* 13**,** 213-21.

VIZAN, P., MILLER, D. S., GORI, I., DAS, D., SCHMIERER, B. & HILL, C. S. 2013. Controlling long-term signaling: receptor dynamics determine attenuation and refractory behavior of the TGF-beta pathway. *Sci Signal,* 6**,** ra106.

VOIGT, P., LEROY, G., DRURY, W. J., 3RD, ZEE, B. M., SON, J., BECK, D. B., YOUNG, N. L., GARCIA, B. A. & REINBERG, D. 2012. Asymmetrically modified nucleosomes. *Cell,* 151**,** 181-93.

VOIGT, P., TEE, W. W. & REINBERG, D. 2013. A double take on bivalent promoters. *Genes Dev,* 27**,** 1318-38.

VOSS, T. C. & HAGER, G. L. 2014. Dynamic regulation of transcriptional states by chromatin and transcription factors. *Nat Rev Genet,* 15**,** 69-81.

WAKEFIELD, L. M. & HILL, C. S. 2013. Beyond TGFbeta: roles of other TGFbeta superfamily members in cancer. *Nat Rev Cancer,* 13**,** 328-41.

WARMFLASH, A., ZHANG, Q., SORRE, B., VONICA, A., SIGGIA, E. D. & BRIVANLOU, A. H. 2012. Dynamics of TGF-beta signaling reveal adaptive and pulsatile behaviors reflected in the nuclear localization of transcription factor Smad4. *Proc Natl Acad Sci U S A,* 109**,** E1947-56.

WOTTON, D., LO, R. S., LEE, S. & MASSAGUE, J. 1999. A Smad transcriptional corepressor. *Cell,* 97**,** 29-39.

WU, J., HUANG, B., CHEN, H., YIN, Q., LIU, Y., XIANG, Y., ZHANG, B., LIU, B., WANG, Q., XIA, W., LI, W., LI, Y., MA, J., PENG, X., ZHENG, H., MING, J., ZHANG, W., ZHANG, J., TIAN, G., XU, F., CHANG, Z., NA, J., YANG, X. & XIE, W. 2016. The landscape of accessible chromatin in mammalian preimplantation embryos. *Nature,* 534**,** 652-7.

WU, J. W., FAIRMAN, R., PENRY, J. & SHI, Y. 2001. Formation of a stable heterodimer between Smad2 and Smad4. *J Biol Chem,* 276**,** 20688-94.

WU, M. Y. & HILL, C. S. 2009. Tgf-beta superfamily signaling in embryonic development and homeostasis. *Dev Cell,* 16**,** 329-43.

XI, Q., HE, W., ZHANG, X. H., LE, H. V. & MASSAGUE, J. 2008. Genome-wide impact of the BRG1 SWI/SNF chromatin remodeler on the transforming growth factor beta transcriptional program. *J Biol Chem,* 283**,** 1146-55.

XI, Q., WANG, Z., ZAROMYTIDOU, A. I., ZHANG, X. H., CHOW-TSANG, L. F., LIU, J. X., KIM, H., BARLAS, A., MANOVA-TODOROVA, K., KAARTINEN, V., STUDER, L., MARK, W., PATEL, D. J. & MASSAGUE, J. 2011. A poised chromatin platform for TGF-beta access to master regulators. *Cell,* 147**,** 1511-24.

YAN, X., LIU, Z. & CHEN, Y. 2009. Regulation of TGF-beta signaling by Smad7. *Acta Biochim Biophys Sin (Shanghai),* 41**,** 263-72.

YANG, Z., YIK, J. H., CHEN, R., HE, N., JANG, M. K., OZATO, K. & ZHOU, Q. 2005. Recruitment of P-TEFb for stimulation of transcriptional elongation by the bromodomain protein Brd4. *Mol Cell,* 19**,** 535-45.

YAO, L. C., BLITZ, I. L., PEIFFER, D. A., PHIN, S., WANG, Y., OGATA, S., CHO, K. W., ARORA, K. & WARRIOR, R. 2006. Schnurri transcription factors from Drosophila and vertebrates can mediate Bmp signaling through a phylogenetically conserved mechanism. *Development,* 133**,** 4025-34.

YE, J., COULOURIS, G., ZARETSKAYA, I., CUTCUTACHE, I., ROZEN, S. & MADDEN, T. L. 2012. Primer-BLAST: a tool to design target-specific primers for polymerase chain reaction. *BMC Bioinformatics,* 13**,** 134.

YING, Q. L., NICHOLS, J., CHAMBERS, I. & SMITH, A. 2003. BMP induction of Id proteins suppresses differentiation and sustains embryonic stem cell self-renewal in collaboration with STAT3. *Cell,* 115**,** 281-92.

YING, Q. L. & SMITH, A. 2017. The Art of Capturing Pluripotency: Creating the Right Culture. *Stem Cell Reports,* 8**,** 1457-1464.

YING, Q. L., WRAY, J., NICHOLS, J., BATLLE-MORERA, L., DOBLE, B., WOODGETT, J., COHEN, P. & SMITH, A. 2008. The ground state of embryonic stem cell self-renewal. *Nature,* 453**,** 519-23.

YOUNG, R. A. 2011. Control of the embryonic stem cell state. *Cell,* 144**,** 940-54.

ZANCONATO, F., FORCATO, M., BATTILANA, G., AZZOLIN, L., QUARANTA, E., BODEGA, B., ROSATO, A., BICCIATO, S., CORDENONSI, M. & PICCOLO, S. 2015. Genome-wide association between YAP/TAZ/TEAD and AP-1 at enhancers drives oncogenic growth. *Nat Cell Biol,* 17**,** 1218-27.

ZARET, K. S. & CARROLL, J. S. 2011. Pioneer transcription factors: establishing competence for gene expression. *Genes Dev,* 25**,** 2227-41.

ZAWEL, L., DAI, J. L., BUCKHAULTS, P., ZHOU, S., KINZLER, K. W., VOGELSTEIN, B. & KERN, S. E. 1998. Human Smad3 and Smad4 are sequence-specific transcription activators. *Mol Cell,* 1**,** 611-7.

ZENTNER, G. E. & HENIKOFF, S. 2013. Regulation of nucleosome dynamics by histone modifications. *Nat Struct Mol Biol,* 20**,** 259-66.

ZENTNER, G. E. & SCACHERI, P. C. 2012. The chromatin fingerprint of gene enhancer elements. *J Biol Chem,* 287**,** 30888-96.

ZENTNER, G. E., TESAR, P. J. & SCACHERI, P. C. 2011. Epigenetic signatures distinguish multiple classes of enhancers with distinct cellular functions. *Genome Res,* 21**,** 1273-83.

ZHANG, L., HUANG, H., ZHOU, F., SCHIMMEL, J., PARDO, C. G., ZHANG, T., BARAKAT, T. S., SHEPPARD, K. A., MICKANIN, C., PORTER, J. A., VERTEGAAL, A. C., VAN DAM, H., GRIBNAU, J., LU, C. X. & TEN DIJKE, P. 2012. RNF12 controls embryonic stem cell fate and morphogenesis in zebrafish embryos by targeting Smad7 for degradation. *Mol Cell,* 46**,** 650-61.

ZHANG, Y., LIU, T., MEYER, C. A., EECKHOUTE, J., JOHNSON, D. S., BERNSTEIN, B. E., NUSBAUM, C., MYERS, R. M., BROWN, M., LI, W. & LIU, X. S. 2008. Model-based analysis of ChIP-Seq (MACS). *Genome Biol,* 9**,** R137.

ZHOU, V. W., GOREN, A. & BERNSTEIN, B. E. 2011. Charting histone modifications and the functional organization of mammalian genomes. *Nat Rev Genet,* 12**,** 7-18.

ZHU, L. J., GAZIN, C., LAWSON, N. D., PAGES, H., LIN, S. M., LAPOINTE, D. S. & GREEN, M. R. 2010. ChIPpeakAnno: a Bioconductor package to annotate ChIP-seq and ChIP-chip data. *BMC Bioinformatics,* 11**,** 237.